ST333/ST406
Applied Stochastic Processes
2023/24

Sam Olesker-Taylor
sam.olesker-taylor@warwick.ac.uk

14th August 2023

Contents

List of Figures

Introduction

Acknowledgements and Comparison with Previous Years

These notes build upon previous versions of the ST333/ST406 course lecture notes. The 2022/23 lecture notes are available on the Moodle page. Be warned, though: the structure of the early sections is a bit different to those; this change was made to try to emphasise and make clearer certain sections which the students found difficult.

The underlying material is based on older versions of the course, taught by W.S. Kendall, D. Hobson, V. Henderson, L. Alili, A. Papavasiliou and K. Habermann. Students Keegan Kang, Iain Carson and Carmen van-del'Isle have contributed by typesetting lecture notes of previous years.

I took over the course in 2022/23. The 2021/22 lecture notes, as taught by Habermann and Kendall, are also available on the Moodle page. However, I have changed the course quite a bit since then—primarily, removing technical calculations and statements, focussing more on intuition and qualitative understanding of continuous-time Markov chains—so those notes may not be so relevant.

The notes are likely to be updated during Term 1, as we work together through the material. Please email me with details if you find any typos or mistakes.

Extra Reading Material

The primary source for these lectures notes, which may be useful to fill in gaps in or supplement the lecture notes, is the book *Markov Chains* by Norris [Nor97] The advanced topics for the ST406 variant are based on Kelly and Yudovina [KY14]. It is important to note that the advanced topics *are not* the same as in 2021/22.

See the reading list for more details of these and other possibilities. These are accessible at third/fourth year undergraduate level.

Exercises are given throughout the lecture notes. These are optional. The primary learning resources are the lecture notes and the example sheets. If more material is desired, feel free to email me and I can provide.

Structure of the Lecture Notes

Chapter 0 is a refresher on some of the most important aspects of *discrete-time Markov chains*. We will not cover this material in lectures, but rather will start immediately on Chapter 1, which introduces continuous-time Markov chains.

Chapter 2 studies a particular case of continuous-time Markov chains: *birth-and-death processes*. These take values in $\mathbb{N} := \{0, 1, ...\}$ and only ever jump by ± 1.

Chapter 3 is devoted to *queueing theory*. This is where the majority of the interesting results and proofs lie. The first chapter is a little dry, being mostly about fundamentals of general continuous-time Markov chains, and the second is a bit specialised. This third chapter on queueing theory is really where the material shines.

Finally, Chapter 4 is on epidemic models. This is a very deep topic, which could be the basis of an entire lecture course, and we only really scratch the surface.

Additional Administrative Details

Further administrative details, including details on recording of lectures and the schedule, can be found on the Moodle page.

0 Preliminaries: Discrete-Time Markov Chains

We begin by reviewing the basics of probability and random processes. All this material should be covered in the union of the prerequisite modules:

- ST202 *Stochastic Processes*;
- *either* the ST202 first-year prerequisites ST111 *Probability Part A* and ST112 *Probability Part B* *or* ST115 *Introduction to Probability*.

For MSc students taking the ST406 version of this module, prerequisites are whatever the equivalent material from your undergraduate degree studies was.

If you encounter material which seems unfamiliar when working through this chapter, you should, as a matter of urgency, spend time revising preliminary material. Norris [Nor97, Chapter 1] is particularly recommended for this, as a lot of our later material is based on [Nor97, Chapter 2]. [GW14, Chapter 12] is also relevant. We do not give proofs here, but all statements and proofs can be found in these books.

We speak interchangeably of *random processes* and of *stochastic processes*. They are the same thing. Find out more about the adjective "stochastic" on Wikipedia.

0.1 Markov Property

Definition 0.1.1 (Markov Property)**.** Let I be a countable set.[1] Each $i \in I$ is called a *state* and I is the *state space*. Let $X = (X_n)_{n=0}^{\infty}$ be a stochastic process taking values in I. Then, X is a *Markov chain* if it satisfies the *Markov property*:

$$\mathbb{P}\{X_n = i_n \mid X_0 = i_0, \ldots, X_{n-1} = i_{n-1}\} = \mathbb{P}\{X_n = i_n \mid X_{n-1} = i_{n-1}\}$$

for all times $n \geq 0$ and all states $i_0, \ldots, i_n \in I$. △

[1] We suppose that I is a subset of the integers endowed with its natural, rather trivial, topology; any countable set can be represented in this manner

Definition 0.1.2 (Time-Homogeneity)**.** A Markov chain X is *time-homogeneous* if the conditional probabilities $\mathbb{P}\{X_n = i_n \mid X_{n-1} = i_{n-1}\}$ do not depend on n. Write

$$p_{i,j} \coloneqq \mathbb{P}_i\{X_1 = j\} \coloneqq \mathbb{P}\{X_1 = j \mid X_0 = i\}.$$

The (time-homogeneous) matrix of conditional probabilities $P \coloneqq (p_{i,j})_{i,j\in I}$ is called the *transition matrix* of the Markov chain. △

The Markov property says, informally,

"Conditional on the present, the past and future are independent."

It turns out that this holds when t is replaced with a *stopping time* T.

Theorem 0.1.3 (Strong Markov Property)**.** *Let X be a time-homogeneous Markov chain on I with transition matrix P. Let T be a stopping time for X. Then, for all $i \in I$, conditional on $T < \infty$ and $X_T = i$, the process $(X_{T+t})_{t\geq 0}$ is a time-homogeneous Markov chain on I with transition matrix P and is independent of $(X_s)_{s\leq T}$.*

Hint. Only the Markov property and basic manipulations are required. △

We can construct any finite-dimensional distribution from a transition matrix and the marginal distribution of X_0—the initial distribution—using the Markov property. Thus, these fully characterise the distribution of $(X_n)_{n\geq 0}$.

It can be checked that the transition matrix P is a *stochastic matrix*.

Definition 0.1.4 (Stochastic Matrix)**.** A matrix $P = (p_{i,j})_{i,j}$ is a *stochastic matrix* if it has non-negative entries and unit row-sums:

$$p_{i,j} \geq 0 \quad \text{for all} \quad i, j \in I \quad \text{and} \quad \textstyle\sum_{j\in I} p_{i,j} = 1 \quad \text{for all} \quad i \in I. \qquad \triangle$$

Example 0.1.5. Consider the four-state Markov chain transition matrix

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/6 & 1/3 & 1/2 & 0 \\ 0 & 1/6 & 1/3 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This is represented in the *state diagram* in Figure 0.1. In particular,

$$p_{i,j} \quad \text{is the probability of going} \quad \textit{to } j \quad \textit{from } i.$$

The *rows* are indexed by i and the *columns* by j. Eg, $p_{2,1} = 1/6$ and $p_{1,2} = 0$. △

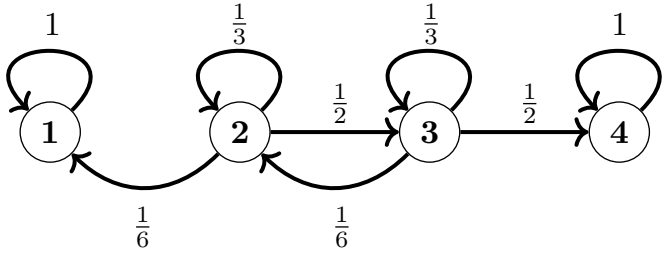

Figure 0.1. Typical state diagram for four-state Markov chain. Transitions are annotated with probabilities

0.2 Further Properties

Definition 0.2.1 (Communication). State $i \in I$ *leads to* state $j \in I$, written $i \to j$, if

$$\mathbb{P}_i\{X_n = j \text{ for some } n > 0\} > 0.$$

States $i, j \in I$ *communicate*, written $i \leftrightarrow j$, if $i \to j$ and $j \to i$.

The relation $\sim$ on I is defined by requiring $i \leftrightarrow j$ or $i = j$, for $i, j \in I$. The relation $\sim$ is an equivalence relation on I (exercise) and thus partitions I into equivalence classes, which are called *communicating classes*.

A chain is *irreducible* if there is only a single communicating class. △

Lemma 0.2.2 (Chapman–Kolmogorov Equations). *Let P be a stochastic matrix and let X be the associated Markov chain. Then, for all $i, j \in I$ and all $n \geq 0$,*

$$\mathbb{P}_i\{X_n = j\} = (P^n)_{i,j} =: p_{i,j}(n),$$

where P^n is the n-th matrix power of P.

Definition 0.2.3 (Aperiodicity). $i \in I$ is *aperiodic* if $\gcd\{n \geq 1 \mid p_{i,i}(n) > 0\} = 1$. △

Lemma 0.2.4. *If a chain is irreducible and has at least one aperiodic state, then all states are aperiodic.*

We now move onto *hitting times* and *recurrence/transience*.

Definition 0.2.5 (Hitting Times). The *hitting times* of a state $i \in I$ are given by

$$H_i := \inf\{n \geq 0 \mid X_n = i\} \quad \text{and} \quad H_i^+ := \inf\{n \geq 1 \mid X_n = i\}. \qquad \triangle$$

Definition 0.2.6 (Recurrence and Transience for States)**.** Let $i \in I$ be a state.

- It is *recurrent* if $\mathbb{P}_i\{H_i^+ < \infty\} = 1$; it is *positive recurrent* if $\mathbb{E}_i(H_i^+) < \infty$.
- It is *transient* if $\mathbb{P}_i\{H_i^+ < \infty\} < 1$. △

Lemma 0.2.7. *Let $i \in I$ be a state. Then, the following equivalences hold:*

$$\mathbb{P}_i\{H_i^+ < \infty\} = 1 \iff \mathbb{P}_i\{|\{n \geq 0 \mid X_n = i\}| = \infty\} = 1;$$
$$\mathbb{P}_i\{H_i^+ < \infty\} < 1 \iff \mathbb{P}_i\{|\{n \geq 0 \mid X_n = i\}| = \infty\} = 0.$$

In particular, $\mathbb{P}_i\{|\{n \geq 0 \mid X_n = i\}| = \infty\} \in \{0, 1\}$.

Hint. Consider the probability of returning to the starting state and apply the strong Markov property (Theorem 0.1.3) at the return time, if it is finite. △

The following equivalence can also be shown.

Theorem 0.2.8 (Recurrence–Transience Dichotomy)**.** *Let $i \in I$ be a state.*

$$\mathbb{P}_i\{H_i^+ < \infty\} = 1 \iff \textstyle\sum_{n\geq 1} p_{i,i}(n) = \infty;$$
$$\mathbb{P}_i\{H_i^+ < \infty\} < 1 \iff \textstyle\sum_{n\geq 1} p_{i,i}(n) < \infty.$$

In particular, every state is either transient or recurrent.

Definition 0.2.9 (Recurrence and Transience for Chains)**.** If a chain is irreducible—ie, has a unique communicating class—then it is *recurrent*, respectively *transient*, if all the states are recurrent, respectively transient. △

Corollary 0.2.10. *If a chain is on a finite state space with a unique communicating class, then the chain is recurrent.*

0.3 Invariant Distribution and Ergodic Theorem

Definition 0.3.1 (Invariant Distribution)**.** A row vector $\pi = (\pi_i)_{i\in I}$ is *invariant* wrt P if $\pi P = \pi$—ie, $\sum_{i\in I} \pi_i p_{i,j} = \pi_j$ for all $j \in I$. It is an *invariant distribution* if, in addition, it is non-negative and has unit sum—ie, $\pi_i \geq 0$ for all $i \in I$ and $\sum_{i\in I} \pi_i = 1$. The phrases *equilibrium* and *stationary distribution* are also used. △

Definition 0.3.2 (Ergodicity)**.** A chain is *ergodic* if it is irreducible and aperiodic. △

Theorem 0.3.3 (Ergodic Theorem). *Suppose that X is an ergodic Markov chain. Suppose it has an invariant distribution π. Let μ be any distribution on I. Then,*

$$\mathbb{P}_\mu\{X_n = j\} \to \pi_j \quad as \quad n \to \infty \quad for\ all \quad j \in I.$$

In particular, π is the unique invariant distribution and taking $\mu := \delta_i$[2] gives

$$p_{i,j}(n) \to \pi_j \quad as \quad n \to \infty \quad for\ all \quad i, j \in I.$$

Exercise 0.3.4. *1. Prove directly that an ergodic (ie, irreducible and aperiodic), positive-recurrent Markov chain has a unique invariant distribution.*

2. Prove that "aperiodic" is not necessary in the above statement.

0.4 Detailed Balance and Reversibility

The equation $\pi P = \pi$ for invariance of a distribution is sometimes referred to as *global balance*. When *detailed balance* holds, more can be said.

Definition 0.4.1 (Detailed Balance and Reversibility). Let P be a transition matrix of a Markov chain and π a measure, both on the state space I. If the chain satisfies the *detailed balance equations*—ie, $\pi_i p_{i,j} = \pi_j p_{j,i}$ for all $i, j \in I$—then the chain is said to be *reversible* (*in equilibrium*). △

Exercise 0.4.2. *If π satisfies detailed balance wrt P, then it is invariant wrt P.*

Exercise 0.4.3. *If π and P are in detailed balance, then*

$$\pi_{i_0} p_{i_0,i_1} \cdot \ldots \cdot p_{i_{n-1},i_n} = \pi_{i_n} p_{i_n,i_{n-1}} \cdot \ldots \cdot p_{i_1,i_0}$$

for all integers $n \geq 1$ and sequences $(i_0, \ldots, i_n)$ of states.

Theorem 0.4.4 (Time-Reversed Chain). *Suppose that π and P satisfy detailed balance. Let $N \in \mathbb{N}$. Suppose that $(X_n)_{n=0}^N$ is a Markov chain with transition matrix P and initial state distributed as π. Set $Y_n := X_{N-n}$. Then, $(Y_n)_{n=0}^N$ is also a Markov chain with transition matrix P and initial state distributed as π.*

The concept of reversibility can be extended beyond chains satisfying the detailed balance equations. Define $\widehat{P}$ by $\widehat{p}_{i,j} := \pi_i p_{i,j}/\pi_j$ for $i, j \in I$. Then, in the notation of the previous theorem, the Markov chain $(Y_n)_{n=0}^N$ has transition matrix $\widehat{P}$.

[2] δ_i is the point-mass at i: $\delta_i(i) = 1$ and $\delta_i(j) = 0$ for $j \neq i$

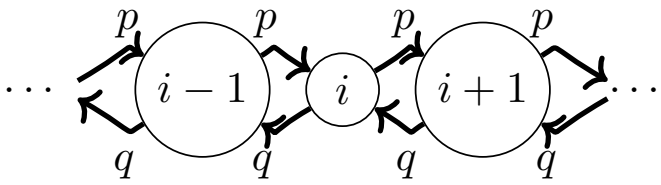

Figure 0.2. State diagram for simple, asymmetric RW on $\mathbb{Z}$

0.5 Examples

Exercise 0.5.1 (Simple Asymmetric Random Walk on $\mathbb{Z}$)**.** *Let $p \in (0,1)$ and $q := 1-p$. Let $I = \mathbb{Z}$ and let the non-zero entries of the transition matrix P be given by*

$$p_{i,i+1} = p \quad \text{and} \quad p_{i,i-1} = q \quad \text{for all} \quad i \in \mathbb{Z}.$$

This is represented in the state diagram in Figure 0.2. Suppose that $p \neq q$.

1. *Show that the chain is irreducible.*
2. *Show that the chain is periodic—ie, is not aperiodic.*
3. *Is the chain recurrent or transient?*

Hitting probabilities and times can be written as solutions to systems of linear equations, fundamentally due to the Markov property. See [Nor97, §1.3] for details.

4. *(Harder) Let $i, j, k \in \mathbb{Z}$ with $i \leq j \leq k$. What is the probability that the chain started from j hits i before k?* **Hint.** *Adjust the chain to be absorbed at k and ask for the probability that it ever hits i and use [Nor97, Theorem 1.3.2].*

Exercise 0.5.2 (Simple Symmetric Random Walk on $\mathbb{Z}$)**.** *Answer questions analogous to those of the previous exercise in the symmetric case $p = q$.*

Example 0.5.3 (Use of Strong/Markov Property for Hitting Probabilities)**.** Let X be a simple, asymmetric random walk on $\mathbb{Z}$ as above. By the Markov property,

$$\alpha := \mathbb{P}_1\{H_0 < \infty\} = p \cdot 1 + q \cdot \mathbb{P}_2\{H_0 < \infty\}.$$

In order to get from 2 to 0, the walk must go through 1. So, by the strong Markov property applied at the stopping time H_1 and translational invariance of the system,

$$\mathbb{P}_2\{H_0 < \infty\} = \mathbb{P}_2\{H_1 < \infty\}\mathbb{P}_2\{H_2 < \infty \mid H_1 < \infty\}$$

$$= \mathbb{P}_2\{H_1 < \infty\}\mathbb{P}_1\{H_0 < \infty\} = \alpha^2.$$

This means that we have reduced to solving the quadratic

$$\alpha = p + q\alpha^2, \quad \text{which has solutions} \quad \{1,\, p/q\}.$$

General theory from [Nor97, Theorem 1.3.2] tells us that α is the *minimal* solution:

$$\alpha = \begin{cases} 1 & \text{if} \quad p \geq q, \\ p/q & \text{if} \quad p < q. \end{cases}$$

△

1 Continuous-Time Markov Chains

This module is about Markov chains in continuous time. In this chapter, we introduce some general principles and some basic definitions. A fully rigorous survey of the theory would take us much too far afield. We therefore make four *Standing Assumptions* which allow us to cover many of the immediate applications of the theory.

Throughout, indices $k, \ell, m, n \in \mathbb{N} := \{0, 1, ...\}$ are discrete, whilst $s, t, u, v \in \mathbb{R}_+ := [0, \infty)$ are continuous. Indices $i, j, x, y \in I$ are states in the state space I.

1.1 Intuitive Constructions and Discussion

We start with the simplest construction of a continuous-time Markov chain $X = (X_t)_{t \geq 0}$ from its discrete-time *jump chain* $\hat{X} = (\hat{X}_n)_{n \geq 0}$, itself a Markov chain.

Definition 1.1.1 (Simplest Continuous-Time Construction)**.** Suppose that $\hat{X}$ is given.

1. Sample *holding times* $T_1, T_2, ...$:
$$T_1, T_2, ... \sim^{\text{iid}} \mathsf{Exp}(1).$$
2. From these, define *jump times* $S_0, S_1, ...$:
$$S_0 := 0 \quad \text{and} \quad S_k := S_{k-1} + T_k = ... = T_1 + ... + T_k \quad \text{for} \quad k \geq 0.$$
3. Define X via $\hat{X}$ from these:
$$X_t := \hat{X}_n \quad \text{for} \quad t \in [S_n, S_{n+1}).$$

In other words, $X_t = X_0$ until the first jump time $S_1 = T_1$, at which it 'jumps' to $X_{S_1} = \hat{X}_1$. It is 'held' at $\hat{X}_1$ for a further T_2 units of time (the holding time), 'jumping' to $X_{S_2} = \hat{X}_2$ at the second jump time $S_2 = T_1 + T_2$. The continues indefinitely. $\triangle$

Remark. We detail later why it is the Exponential distribution which is used. But, in short, it is because this is the only distribution with the *memoryless property*:

$$\mathbb{P}\{S > s+t \mid S > s\} = \mathbb{P}\{S > t\} \quad \text{if and only if} \quad S \sim \mathsf{Exp}(\lambda) \quad \text{for some} \quad \lambda \geq 0.$$

This is necessary for the *Markov property* to hold: "conditional on the present, the future is independent of the past". In particular, knowing how long it was since the chain last jumped must not affect the law of the time until the next jump. △

The parameter λ in $\mathsf{Exp}(\lambda)$ is called the *rate*. In essence, this is because

$$\mathbb{P}\{S < \delta\} = \lambda\delta + o(\delta) \quad \text{as} \quad \delta \downarrow 0.$$

Using the memoryless property, $S \sim \mathsf{Exp}(\lambda)$ can be approximated as follows.

- Discretise $[0,\infty) = [0,\delta) \dot\cup [\delta, 2\delta) \dot\cup [2\delta, 3\delta) \dot\cup \ldots$.
- Sample $B_1, B_2, \ldots \sim^{\mathsf{iid}} \mathsf{Bern}(\lambda\delta)$.
- Set $S \coloneqq \inf\{k \geq 1 \mid B_k = 1\}$.

The interpretation of B_k is "an attempt to 'fire' in the interval $[(k-1)\delta, k\delta)$". These attempts are independent across intervals, by the memoryless property.

This is the most important viewpoint of the Exponential holding time: in every infinitesimal interval of length δ, it tries to 'fire', succeeding with probability $\approx \delta$.

Jumping always at rate 1 is the simplest case. It is easy to change this (uniform) rate to any $\lambda > 0$. However, we can also let the rate depend on the current state $i \in I$—it cannot depend on the past, of course, by the Markov property.

- Sample *holding times* $T_1^{(i)}, T_2^{(i)}, \ldots \sim^{\mathsf{iid}} \mathsf{Exp}(q_i)$ independently for each state i.
- Use holding time $T_k^{(i)}$ for the k-th visit to state $i \in I$.

We have defined the continuous-time X from the discrete $\hat{X}$ along with the holding times $T = (T_k^{(i)})_{i \in I, k \geq 0}$. However, it is more natural to construct the simultaneously.

Definition 1.1.2 (Vertex-Rate Construction)**.** Let I be a finite set. Let $\Pi = (\pi_{i,j})_{i,j \in I}$ be a *transition matrix* on I. Let $q = (q_i)_{i \in I} \in (0,\infty)^I$—ie, $q_i \in (0,\infty)$ for all $i \in I$.

1. Suppose that the current time is $t \geq 0$ and state is $i \in I$.
   - Sample $T \sim \mathsf{Exp}(q_i)$.
   - Sample $J \sim \pi_{i,\cdot}$— ie, $\mathbb{P}\{J = j\} = \pi_{i,j}$ for all $j \in I$.
2. Set $X_s \coloneqq X_t$ for $s \in [t, t+T)$ and $X_{t+T} \coloneqq J$.

All random variables sampled are independent of all others, across different steps. △

The interpretation is that the chain is continually trying to jump, at rate q_i if it is currently at state i, and moves according to the transition matrix P when it jumps.

Remark. A consequence of this lack of uniformity in $(q_i)_{i\in I}$ is that it is no longer possible to pre-define the jump times $S_1, S_2, ...$ if where the chain will jump—ie, $\hat{X} = (\hat{X}_n)_{n\geq 0}$—is not known. This tends not to be so important in practice. △

The Exponential distribution possesses another important property.

Lemma 1.1.3 (Competition of Exponentials). *Let $\lambda_1, \ldots, \lambda_n > 0$ and let $E_i \sim \mathsf{Exp}(\lambda_i)$ independently for each i; let $E \coloneqq \min\{E_1, \ldots, E_n\}$ and $\lambda \coloneqq \lambda_1 + \ldots + \lambda_n$. Then,*

$$E \sim \mathsf{Exp}(\lambda) \quad \textit{and} \quad \mathbb{P}\{E = E_i\} = \lambda_i/\lambda \quad \textit{for all} \quad i.$$

This allows us to decompose an Exponential random variable T with rate $q_i = \sum_j q_i\pi_{i,j}$ into a minimum $\min_j T_j$ where $T_j \sim \mathsf{Exp}(q_{i,j} \coloneqq q_i\pi_{i,j})$ independently.

Definition 1.1.4 (Edge-Rate Construction). Let I be a finite set. Let $q = (q_{i,j})_{i,j\in I} \in \mathbb{R}_+^{I\times I}$—ie, $q_{i,j} \geq 0$ for all $i, j \in I$—with $q_i \coloneqq \sum_j q_{i,j} > 0$ for all i.

1. Suppose that the current time is $t \geq 0$ and state is $i \in I$.
   - Sample $T_j \sim \mathsf{Exp}(q_{i,j})$ independently for each $j \in I$.
   - Set $T \coloneqq \min_j T_j$ and $J \coloneqq \arg\min_j T_j$—ie, $T = T_J$.
2. Set $X_s \coloneqq X_t$ for $s \in [t, t + T)$ and $X_{t+T} \coloneqq J$.

All random variables sampled are independent of all others, across different steps. △

Competition of Exponentials implies that the jump rate is $q_i = \sum_{i,j} q_{i,j}$ from state $i \in I$, and that the resulting state J is distributed as $q_{i,\cdot}$—ie, $\mathbb{P}\{J = j\} = q_{i,j}/q_i$. Hence, the jump chain has transition matrix $\Pi = (\pi_{i,j} \coloneqq q_{i,j}/q_i)_{i,j\in I}$.

A consequence of this construction is the following infinitesimal description:

$$p_{i,j}(t + \delta) = p_{i,j}(t) + q_{i,j}\delta + o(\delta) \quad \text{as} \quad \delta \downarrow 0; \quad \text{equivalently,} \quad \tfrac{d}{dt}p_{i,j}(t) = q_{i,j}.$$

Some transitions may not be permitted: $q_{i,j} = 0$ is allowed, provided $q_i = \sum_j q_{i,j} > 0$.

- The edge-rate construction is the important of a continuous-time Markov chain:

  whilst at $i \in I$, it attempts to jump to j at (infinitesimal) rate $q_{i,j}$ simultaneously (and independently) for each $j \in I$.

- The vertex-rate construction is the second most important:

  whilst at $i \in I$, it attempts to jump at (infinitesimal) rate q_i; when it jumps, it moves according to the transition matrix Π.

1.2 Basic Definitions and Terminology

We proceed with the fundamental definition of a *Markov chain* in continuous time.

Definition 1.2.1 (Markov Property). A random process $X = (X_t)_{t\geq 0}$ with countable state-space I has the *Markov property* if

$$\mathbb{P}\{X_t = j \mid X_s = i, X_{u_1} = k_1, ..., X_{u_{n-1}} = k_{n-1}\} = \mathbb{P}\{X_t = j \mid X_s = i\}$$
$$\text{whenever} \quad 0 \leq u_1 \leq \ldots \leq u_n \leq s \leq t \quad \text{and} \quad k_1, \ldots, k_n, i, j \in I.$$

This can be expressed more succinctly using its *natural filtration*:

$$\mathbb{P}\{X_t = j \mid X_s = i, \sigma((X_s)_{s\leq t})\} = \mathbb{P}\{X_t = j\} \quad \text{whenever} \quad 0 \leq s \leq t \quad \text{and} \quad i, j \in I.$$

A stochastic process that has the Markov property is called a *Markov chain*, or *process*. The probabilities $\mathbb{P}\{X_t = j \mid X_s = i\}$ are called its *transition probabilities*. △

Definition 1.2.2 (Time Homogeneity). A Markov chain X is *time-homogeneous* if $\mathbb{P}\{X_t = j \mid X_s = i\} = \mathbb{P}\{X_{t-s} = j \mid X_0 = i\}$ depends only on (s, t) via $t - s$. Then,

$$p_{i,j}(t) \coloneqq \mathbb{P}\{X_t = j \mid X_0 = i\} \quad \text{for} \quad t \geq 0 \quad \text{and} \quad i, j \in I.$$

We often view $P(t) \coloneqq (p_{i,j}(t))_{i,j\in I}$ as a matrix or linear operator. △

The Markov property says $(X_{t+s} \mid X_t = i) \sim (X_s \mid X_0 = i)$ for deterministic $s, t \geq 0$ in the time-homogeneous set-up. The *strong* Markov property extends this to allow t to be a *stopping time* for discrete-time chains. The set-up is more delicate in continuous-time, but an analogous statement does hold under suitable regularity conditions—including *non-explosivity*, which means that it is not possible for infinitely many steps to be taken in finite time. We do not detail these technical conditions or give a proof; that would require a significant detour into measure theory.

Definition 1.2.3 (Stopping Time). A random time $T \in [0, \infty]$ is a *stopping time for* $X = (X_t)_{t\geq 0}$ if the event $\{T \leq t\}$ depends only on $(X_s)_{s\in[0,t]}$ for all $t \geq 0$. △

Theorem 1.2.4 (Strong Markov Property). *Let X be a time-homogeneous, continuous-time Markov chain I. Let T be a stopping time for X—ie, $\{T \leq t\}$ depends only on $(X_s)_{s\leq t}$. Then, under certain regularity conditions, for all $i \in I$, conditional on $T < \infty$ and $X_T = i$, the process $(X_{T+t})_{t\geq 0}$ is a Markov chain on I started from i with the same transition probabilities $(P(t))_{t\geq 0}$ as X and is independent of $(X_s)_{s\leq T}$.*

We impose a simplifying *Standing Assumption*—the first of four—which makes the theory much cleaner whilst covering many applications.

Standing Assumption I. *Markov chains in this course are continuous-time stochastic processes with countable state-space and are time-homogeneous.*

Here is a further simplifying *Standing Assumption*, which excludes the possibility of some annoying and pathological behaviour.

Standing Assumption II. *The functions $t \mapsto p_{i,j}(t)$ are continuous at 0 for all $i, j \in I$:*

$$\lim_{t \downarrow 0} p_{i,j}(t) = p_{i,j}(0) = \mathbf{1}\{i = j\} = \begin{cases} 1 & \text{if} \quad i = j, \\ 0 & \text{if} \quad i \neq j. \end{cases}$$

We can prove the *Chapman–Kolmogorov* (*CK*) *equations* as in discrete time.

Theorem 1.2.5 (Chapman–Kolmogorov Equations). *For all $s, t \geq 0$ and all $i, j \in I$,*

$$p_{i,j}(s+t) = \textstyle\sum_{k \in I} p_{i,j}(s) p_{k,j}(t).$$

This can be summarised efficiently as a matrix product: $P(s+t) = P(s)P(t)$.

Proof. The proof is as in discrete time. Let $s, t \geq 0$ and $i, j \in I$. We have

$$\begin{aligned} p_{i,j}(s+t) &= \mathbb{P}\{X_{s+t} = j \mid X_0 = i\} \\ \text{(law of total prob)} &= \textstyle\sum_{k \in I} \mathbb{P}\{X_{s+t} = j,\ X_s = k \mid\ X_0 = i\} \\ \text{(Bayes's theorem)} &= \textstyle\sum_{k \in I} \mathbb{P}\{X_{s+t} = j \mid X_s = k,\ X_0 = i\} \mathbb{P}\{X_s = k \mid X_0 = i\} \\ \text{(Markov property)} &= \textstyle\sum_{k \in I} \mathbb{P}\{X_{s+t} = j \mid X_s = k\} \mathbb{P}\{X_s = k \mid X_0 = i\} \\ \text{(time-homogeneity)} &= \textstyle\sum_{k \in I} p_{k,j}(t) p_{i,k}(s). \end{aligned}$$

The same formula holds for products of matrices, hence the final statement. □

Corollary 1.2.6. *Suppose that Standing Assumption II holds. For a given initial state $i \in I$, the transition probabilities $p_{i,j}(t)$ are uniformly continuous in the time $t \geq 0$ and the final state $j \in I$.*

It follows that the family $(P(t))_{t \geq 0}$ of transition probability matrices forms a *semigroup of stochastic matrices*.

1. $(P(t))_{t \geq 0}$ forms a *semigroup*:
   - $P(0)$ is the identity matrix/operator on I;
   - $P(t+s) = P(t)P(s)$ for all $s, t \geq 0$.
2. $P(t)$ is *stochastic* for each $t \geq 0$:

- $p_{i,j}(t) \geq 0$ for all $i, j \in I$;
- $\sum_{j \in I} p_{i,j}(t) = 1$ for all $i \in I$.

The converse also holds: a semigroup of stochastic matrices defines a family of transition probability matrices.

Theorem 1.2.7. *Suppose that Standing Assumption I holds. Suppose also that the Markov chain $X = (X_t)_{t \geq 0}$ is right-continuous: for all $\omega \in \Omega$ and $t \geq 0$, there exists an $\varepsilon > 0$ such that*

$$X_t(\omega) = X_s(\omega) \quad \text{for all} \quad s \in [t, t + \varepsilon].$$

Then, the distribution of the Markov chain is fully characterised by its family of transition probabilities $(P(t))_{t \geq 0}$ and its initial distribution of X_0.

Proof (Sketch). It is a standard result in measure theory that the probability of any event depending on a right-continuous process can be determined from its finite-dimensional distributions: the probabilities

$$\mathbb{P}\{X_{t_0} = i_0, \ldots, X_{t_n} = i_n\}$$

for any $n \geq 0$, $i_0, \ldots, i_n \in I$ and $0 \leq t_0 \leq \ldots \leq t_n$. We take this as given; see [Nor97, §6.6] for details. Now, the finite-dimensional distributions can be handled using Bayes's theorem, the Markov property and time-homogeneity:

$$\begin{aligned}
&\mathbb{P}\{X_{t_0} = i_0, \ldots, X_{t_n} = i_n\} \\
&\quad = \mathbb{P}\{X_0 = i_0\} \textstyle\prod_{k=1}^{n} \mathbb{P}\{X_{t_k} = i_k \mid X_{t_{k-1}} = i_{k-1}, \ldots, X_0 = i_0\} \\
&\quad = \mathbb{P}\{X_0 = i_0\} \textstyle\prod_{k=1}^{n} \mathbb{P}\{X_{t_k} = i_k \mid X_{t_{k-1}} = i_{k-1}\} \\
&\quad = \mathbb{P}\{X_0 = i_0\} \textstyle\prod_{k=1}^{n} p_{i_{k-1}, i_k}(t_k - t_{k-1}).
\end{aligned}$$

This depends only on the law of X_0 and on the transition probabilities $(P(t))_{t \geq 0}$. □

Exercise 1.2.8. *Consider a two-state Markov chains X, say on $I = \{0, 1\}$. The two states may represent "broken" or "working" for a machine. Let $\lambda, \mu \in (0, \infty)$. We propose the following model for the transition probabilities, for $t \geq 0$:*

$$\begin{aligned}
\mathbb{P}\{X_t = 1 \mid X_0 = 1\} &= e^{-\lambda t}; \\
\mathbb{P}\{X_t = 0 \mid X_0 = 0\} &= e^{-\mu t}.
\end{aligned}$$

Construct the family $(P(t))_{t \geq 0}$ of transition matrices. Do they form a semigroup?

Exercise 1.2.9. *Consider the set-up of the previous exercise. Suppose instead that the transition probabilities are, for $t \geq 0$, proposed by the following:*

$$\mathbb{P}\{X_t = 1 \mid X_0 = 1\} = \tfrac{1}{2}(1 + e^{-2\lambda t});$$

$$\mathbb{P}\{X_0 = 0 \mid X_0 = 0\} = \tfrac{1}{2}(1 + e^{-2\lambda t}).$$

Show that these form a semigroup.

1.3 From Discrete- to Continuous-Time Markov Chains

In this section, we study which of the concepts and properties of Markov chains from transfer from discrete-time to continuous-time, and which do not.

1.3.1 The Jump Chain

First, show how to view a (suitably regular) continuous-time Markov chain contains an embedded discrete-time Markov chain: essentially, sample the continuous-time chain at the times at which it jumps.

Definition 1.3.1 (Holding and Jump Times)**.** Let $X = (X_t)_{t\geq 0}$ be a continuous-time Markov chain with right-continuous paths. The *jump times* $S = (S_n)_{n\geq 0}$ are

$$S_0 \coloneqq 0 \quad \text{and} \quad S_n \coloneqq \inf\{t \geq S_{n-1} \mid X_t \neq X_{S_{n-1}}\} \quad \text{for} \quad n \geq 1.$$

The *inter-arrival*, or *holding*, *times* $T = (T_n)_{n\geq 1}$ are

$$T_n \coloneqq \begin{cases} S_n - S_{n-1} & \text{if} \quad S_{n-1} < \infty, \\ \infty & \text{if} \quad S_{n-1} = \infty. \end{cases}$$

It follows from right-continuity of paths that $S_n, T_n > 0$ for all $n \geq 1$. △

Definition 1.3.2 (Jump Chain)**.** Given a Markov chain $X = (X_t)_{t\geq 0}$ with jump times $S = (S_n)_{n\geq 0}$, define the *jump chain* to be the discrete-time process $\hat{X} = (\hat{X}_n)_{n\geq 0}$ by

$$\hat{X}_n \coloneqq X_{S_n} \quad \text{for} \quad n \geq 0.$$

This is a discrete-time Markov chain. Let $\Pi = (\pi_{i,j})_{i,j}$ denote its transition matrix[1]:

$$\pi_{i,j} \coloneqq \mathbb{P}\{X_{S_1} = j \mid X_0 = i\} \quad \text{for} \quad i \in I.$$

This is the probability that it jumps to state j the first time it leaves state i. △

We can study many of the properties of the continuous-time Markov chain by studying the corresponding jump chain. In order to do so, we need to compute the transition matrix Π—the *jump matrix*—with entries given by $\pi_{i,j} = \mathbb{P}_i\{X_{S_1} = j\}$.

[1] recall from Standing Assumption I that the chain in time-homogeneous: its transitions do not depend on the starting time

Definition 1.3.3 (Transition Derivatives and Matrix)**.** Define

$$q_{i,j} \coloneqq \lim_{\delta \downarrow 0} \tfrac{1}{\delta}\big(p_{i,j}(\delta) - p_{i,j}(0)\big) \quad \text{for} \quad i,j \in I.$$

Set $q_i \coloneqq -q_{i,i}$ for $i \in I$. Define the *transition-rates matrix* $Q \coloneqq (q_{i,j})_{i,j\in I}$. △

This leads us to the *infinitesimal definition* of a continuous-time Markov chain.

Definition 1.3.4 (Infinitesimal Definition)**.** Suppose that $P = (P(t))_{t\geq 0}$, with $P(t) = (p_{i,j}(t))_{i,j\in I}$ for all $t \geq 0$, is a bona fied family of time-homogeneous transition probabilities. Define the transition-rates matrix $Q = (q_{i,j})_{i,j\in I}$ as in Definition 1.3.3. The *infinitesimal definition* of the associated continuous-time Markov chain is

$$p_{i,j}(t+\delta) = p_{i,j}(t) + q_{i,j}\delta + o(\delta) \quad \text{as} \quad \delta \to 0. \qquad \triangle$$

We want to understand what it means for jumps to happen at certain *rates*.

Remark (Understanding Transition Rates)**.** By Taylor's theorem,

$$p_{i,j}(\delta) = p_{i,j}(0) + q_{i,j}\delta + o(\delta) \quad \text{as} \quad \delta \to 0.$$

In a tiny interval of length δ, the probability of jumping $i \to j$ is approximately $q_{i,j}\delta$. Equivalently, we can view the process as jumping *somewhere* from i at rate q_i and, upon jumping, choosing the location with probabilities proportional to $(q_{i,j})_{j\in I}$. △

Lemma 1.3.5 (Jump Matrix)**.** *The jump matrix* $\Pi = (\pi_{i,j})_{i,j\in I}$ *satisfies*

$$\pi_{i,j} = \begin{cases} q_{i,j}/q_i & \text{if} \quad i \neq j \quad \text{and} \quad q_i \neq 0, \\ 0 & \text{if} \quad i = j \quad \text{and} \quad q_i \neq 0, \\ 0 & \text{if} \quad i \neq j \quad \text{and} \quad q_i = 0, \\ 1 & \text{if} \quad i = j \quad \text{and} \quad q_i = 0, \end{cases} \quad \text{for all} \quad i,j \in I.$$

Proof. If $q_i = 0$, then the chain never leaves i, so the claim is trivial. Assume $q_i \neq 0$.

Let $\delta > 0$ be small. Define $Y_n \coloneqq X_{n\delta}$. Then, $Y = (Y_n)_{n\geq 0}$ is a *discrete-time* Markov chain with transition matrix $P = I + \delta Q + o(\delta)$ as $\mathbb{P}_i\{Y_1 = j\} = p_{i,j}(\delta)$. Now,

$$\mathbb{P}_i\{Y_1 \neq i\} = \textstyle\sum_{j:j\neq i} \mathbb{P}_i\{Y_1 = j\} = \delta \sum_{j:j\neq i} q_{i,j} + o(\delta) = \delta q_i + o(\delta).$$

Define $N \coloneqq \inf\{n \geq 0 \mid Y_n \neq Y_0\}$; then, $N \sim \mathsf{Geom}_1(\delta q_i + o(\delta))$. Then,

$$\mathbb{P}_i\{Y_N = j\} = q_{i,j}/q_i + o(1) \quad \text{where} \quad o(1) \to 0 \text{ as } \delta \to 0 \quad \text{when} \quad i \neq j,$$

by the Markov property for Y. Also by the Markov property for Y, the probability that X jumps twice by time δ is $O(\delta^2)$. This means that

$$\mathbb{P}\{\delta(N-1) < T < \delta N\} = 1 - o(1) \quad \text{where} \quad T \coloneqq \inf\{t \geq 0 \mid X_t \neq X_0\}.$$

In words, this is says that, with probability tending to 1 as $\delta \to 0$, the first jump for X happens in the same interval as for Y. Hence,

$$\pi_{i,j} = \mathbb{P}_i\{X_T = j\} \approx \mathbb{P}_i\{Y_N = j\} \approx q_{i,j}/q_i \quad \text{when} \quad i \neq j,$$

where the "$\approx$" sign hides additive $o(1)$ errors. Taking $\delta \to 0$ proves the lemma. □

1.3.2 Instantaneous Transition Rates

The first question we ask of a continuous-time Markov chain is,

"How long does it stay in its current state?"

We claimed in §1.1.1 that this is Exponentially distributed[2], as a consequence of the Markov property. We prove this now.

Definition 1.3.6 (Memoryless Property). T has the *memoryless property* if

$$\mathbb{P}\{T \geq t + s \mid X \geq t\} = \mathbb{P}\{T \geq s\} \quad \text{for all} \quad s, t \geq 0. \qquad \triangle$$

We are looking at all the holding times. But, by the Markov property and time homogeneity, we may re-index to assume that it is currently time 0.

Lemma 1.3.7 (Memoryless Property of Holding Time). *The first jump time T_1 has the memoryless property given the initial state X_0.*

Proof. This is a consequence of the Markov property:

$$\begin{aligned}\mathbb{P}_i\{T_1 \geq t + s \mid T_1 \geq s\} &= \mathbb{P}\{X_u = i \,\forall\, u \in [0, t+s) \mid X_u = i \,\forall\, u \in [0, s)\} \\ &= \mathbb{P}\{X_u = i \,\forall\, u \in [s, t+s) \mid X_u = i \,\forall\, u \in [0, s)\} \\ &= \mathbb{P}\{X_u = i \,\forall\, u \in [0, t) \mid X_0 = i\} = \mathbb{P}_i\{T_1 \geq t\},\end{aligned}$$

using the Markov property and time homogeneity in the penultimate equality. □

The *only* continuous, non-negative random variable with the memoryless property is the Exponential: $T \sim \mathsf{Exp}(\lambda)$ for some λ.

[2] a random variable T has the *Exponential distribution with rate* λ if $\mathbb{P}\{T > t\} = e^{-\lambda t}$

Lemma 1.3.8. *Suppose that T is a continuous random variable taking values in $[0,\infty)$. If T has the memoryless property, then it is exponentially distributed:*

$$\mathbb{P}\{T \geq t\} = e^{-\lambda t} \quad \textit{for some} \quad \lambda \geq 0.$$

Proof. Write $\bar{F}(t) \coloneqq \mathbb{P}\{T \geq t\}$ for the complement of the cdf. Then, the memoryless property says precisely that

$$\bar{F}(t+s) = \bar{F}(t)\bar{F}(s) \quad \text{for all} \quad s,t \geq 0.$$

It follows that

$$\big(\bar{F}(t+s) - \bar{F}(t)\big)/s = \bar{F}(t)\big(\bar{F}(s) - 1\big)/s.$$

Taking $s \downarrow 0$ gives

$$\bar{F}'(t) = \bar{F}(t) \cdot \bar{F}'(0) = -\lambda \bar{F}(t)$$

where $\lambda \coloneqq -\bar{F}'(0) > 0$. This DE is easy to solve:

$$\bar{F}(t) = Ae^{-\lambda t} \quad \text{for some constant } A.$$

But, $\bar{F}(0) = 1$, so $A = 1$. Hence, $\bar{F}(t) = e^{-\lambda t}$ as required. □

Hence, $(T_1 \mid X_0 = i) \sim \mathsf{Exp}(\lambda)$ for some λ. But, what is λ?

Lemma 1.3.9 (Exponential Rate). *Let $i \in I$. Then, $(T_1 \mid X_0 = i) \sim \mathsf{Exp}(q_i)$.*

Proof. The previous two lemmas give $(T_1 \mid X_0 = i) \sim \mathsf{Exp}(\lambda)$ with $\lambda = -\bar{F}'(0)$, where $\bar{F}$ is the complementary cdf of T_1. It remains to calculate the derivative

$$\bar{F}'(0) = \lim_{\delta \downarrow 0}\big(\bar{F}(0) - 1\big)/\delta = \lim_{\delta \downarrow 0}\big(\mathbb{P}_i\{T_1 > \delta\} - 1\big)/\delta.$$

Now, if δ is very small, then $\{T_1 > \delta\} \approx \{X_\delta = i\}$, given $X_0 = i$. The error comes from having at least two jumps before δ, which has probability order $\delta^2 = o(\delta)$. Thus,

$$\begin{aligned}\big(\mathbb{P}_i\{T_1 > \delta\} - 1\big) &= \big(\mathbb{P}_i\{X_\delta = i\} + o(\delta) - 1\big)/\delta \\ &= \big(p_{i,i}(\delta) - p_{i,i}(0)\big)/\delta + o(1) = p'_{i,i}(0) + o(1) = -q_i + o(1).\end{aligned}$$

Taking $\delta \downarrow 0$ gives $\lambda = q_i$. Hence, $(T_1 \mid X_0 = i) \sim \mathsf{Exp}(q_i)$. □

We now know that the continuous-time chain X waits a time $E \sim \mathsf{Exp}(q_i)$, if it is initially at i, then jumps to j with probability $\pi_{i,j} = q_{i,j}/q_i$. The following lemma shows that E can be written as $\min_j E_j$ where $E_j \sim \mathsf{Exp}(q_{i,j} = q_i \pi_{i,j})$ independently.

Lemma 1.3.10 (Competition of Exponentials). *Let $\lambda_1, \ldots, \lambda_n > 0$ and $E_i \sim \mathsf{Exp}(\lambda_i)$ independently for each i; let $E \coloneqq \min\{E_1, \ldots, E_n\}$ and $\lambda \coloneqq \lambda_1 + \ldots + \lambda_n$. Then,*

$$E \sim \mathsf{Exp}(\lambda) \quad \textit{and} \quad \mathbb{P}\{E = E_i\} = \lambda_i/\lambda \quad \textit{for all} \quad i.$$

Exercise 1.3.11. *Prove this lemma. See [Nor97, Theorem 2.3.3] for details. There, the countably-infinite case is handled too, assuming that* $\lambda := \lambda_1 + \lambda_2 + \ldots < \infty$.

We apply this with $T_j \sim \mathsf{Exp}(q_{i,j} = q_i \pi_{i,j})$ independently for $j \in I$:

$$\textstyle\sum_j q_{i,j} = q_i \quad \text{and} \quad q_{i,j}/q_i = \pi_{i,j} \quad \text{for all} \quad i, j \in I.$$

The interpretation of this is that, in a short interval of length δ, all T_j try to 'fire' independently; T_j has probability $\approx q_{i,j}\delta$ of firing.

- If none fire, then the chain stays put.
- If precisely one fires—say, T_J—then the chain jumps to J.
- The probability that two or more fire is $\mathcal{O}(\delta^2)$, which is ignored when $\delta \to 0$.

This process is then repeated in the next (infinitesimal) interval of length δ.

This representation leads to the following naming convention.

Definition 1.3.12 (Instantaneous Transition Rates)**.** We call $q_{i,j}$ the (*instantaneous*) *transition rate* from i to j. The matrix $Q = (q_{i,j})_{i,j \in I}$ is the (*instantaneous*) *transition-rate matrix*, or the *generator of the transition semigroup* $P = (P(t))_{t \geq 0}$. △

We want to rule out the possibility that the chain can leave a state infinitely fast. This can lead to very pathological behaviour from the application point of view.

Standing Assumption III. *We have* $q_i < \infty$ *for all* $i \in I$. *This implies that* $\sup_{i \in I} q_i < \infty$ *if* I *is finite, but it* does not *require* $\sup_{i \in I} q_i < \infty$ *if* I *is countably infinite.*

We also want to rule out the possibility that the chain leaves a state at a rate different to the sum of all the rates at which it jumps to other states. This may seem obviously true, rather than requiring a *Standing Assumption* of its own. But, again, there are pathological examples where this does not hold.

Standing Assumption IV. *We have*

$$q_i = \textstyle\sum_{j \in I \setminus \{i\}} q_{i,j} \quad \text{for all} \quad i \in I.$$

This holds whenever the state space I *is finite.*

The Standing Assumptions imply that the matrix Q is a *Q-matrix*.

Definition 1.3.13 (Q-Matrix)**.** A matrix Q is a *Q-matrix* if it satisfies the following.

- Finite, non-positive diagonal entries: $0 \leq -q_{i,i} < \infty$ for all $i \in I$.
- Finite, non-negative off-diagonal entries: $q_{i,j} \geq 0$ for all $i, j \in I$ with $i \neq j$.
- Zero row-sums: $\sum_{j \in I} q_{i,j} = 0$ for all $i \in I$. △

1.3.3 Class Structure, Recurrence and Transience

The Markov chain dynamics can be used to classify the state space.

Definition 1.3.14 (Communication). State i *leads to* j, written $i \to j$, for $i, j \in I$ if

$$\mathbb{P}\{X_t = j \text{ for some } t \geq 0 \mid X_0 = i\} > 0.$$

If $i \to j$ and $j \to i$, then i and j *communicate*, written $i \leftrightarrow j$. This is an equivalence relation. It partitions the space into equivalences classes called *communicating classes*. A Markov chain is *irreducible* if it has a unique communicating class. △

Theorem 1.3.15. *Let X be a continuous-time Markov chain and let $\hat{X}$ be its jump chain. For all $i, j \in I$ with $i \neq j$, the following are equivalent:*

1. *$i \to j$ for X;*
2. *$i \to j$ for $\hat{X}$;*
3. *there exist $m \geq 1$ and $k_1, ..., k_m \in I$ with $\pi_{i,k_1}, \pi_{k_1,k_2}, ..., \pi_{k_{m-1},k_m}, \pi_{k_m,j} > 0$;*
4. *$p_{i,j}(t) > 0$ for all $t > 0$.*

Exercise 1.3.16. *Prove this theorem. See [Nor97, Theorem 3.2.1] for the details.*

We want to distinguish between states that are visited at arbitrarily large times and those which are eventually left unvisited forever. We have to adjust the definition compared with discrete time: the continuous nature of time means that a vertex is either never visited, or is visited at infinitely many distinct times; we replace "infinitely many" with "an unbounded set". As in discrete-time, this is equivalent returning almost surely—again, though, "return" means "return after stepping away", not "there exists a future time at which the chain is in the same state".

Definition 1.3.17 (Hitting Times). Let X be a continuous-time Markov chain, and T_1 its first jump time. The *hitting times* of a state $i \in I$ are given by

$$H_i \coloneqq \inf\{t \geq 0 \mid X_t = i\} \quad \text{and} \quad H_i^+ \coloneqq \inf\{t \geq T_1 \mid X_t = i\}. \qquad \triangle$$

Definition 1.3.18 (Recurrence and Transience for States). Let $i \in I$ be a state.

- It is *recurrent* if $\mathbb{P}_i\{H_i^+ < \infty\} = 1$; it is *positive recurrent* if $\mathbb{E}_i(H_i^+) < \infty$.
- It is *transient* if $\mathbb{P}_i\{H_i^+ < \infty\} < 1$. △

Lemma 1.3.19. *Let $i \in I$ be a state. Then, the following equivalences hold:*

$$\mathbb{P}_i\{H_i^+ < \infty\} = 1 \iff \mathbb{P}_i\{\{t \geq 0 \mid X_t = i\} \text{ is unbounded}\} = 1;$$

$$\mathbb{P}_i\{H_i^+ < \infty\} < 1 \iff \mathbb{P}_i\{\{t \geq 0 \mid X_t = i\} \text{ is unbounded}\} = 0.$$

In particular, $\mathbb{P}_i\{\{t \geq 0 \mid X_t = i\} \text{ is unbounded}\} \in \{0, 1\}$.

Exercise 1.3.20. *Prove this lemma.* **Hint.** *Consider the probability of returning state and apply the strong Markov property at the return time, if it is finite.*

Hint. Consider the probability of returning to the starting state and apply the strong Markov property (Theorem 1.2.4) at the return time, if it is finite. △

Theorem 1.3.21. *A state is recurrent for the continuous-time Markov chain X if and only if it is recurrent for the discrete-time jump chain $\hat{X}$.*

The following dichotomy holds, analogous to discrete-time case: cf Theorem 0.2.8.

Theorem 1.3.22. *Let $i \in I$. If $q_i = 0$, then state i is recurrent. If $q_i > 0$, then the following recurrence–transience dichotomy holds:*

- *if $\mathbb{P}_i\{H_i^+ < \infty\} = 1$, then state i is recurrent and $\int_0^\infty p_{i,i}(t)dt = \infty$;*
- *if $\mathbb{P}_i\{H_i^+ < \infty\} < 1$, then state i is transient and $\int_0^\infty p_{i,i}(t)dt < \infty$.*

Exercise 1.3.23. *Prove this theorem.* **Hint.** *Compare the integral $\int_0^\infty p_{i,i}(t)dt$ with the sum $\sum_{n=0}^\infty \pi_{i,i}(n)$, where $\pi_{i,i}(n) = \mathbb{P}_i\{\hat{X}_n = i\}$, and use Theorem 0.2.8.*

A full proof is given in [Nor97, Theorem 3.4.2]. "Fubini's theorem" there is a theorem justifying swapping the order of an integral and an expectation.

1.3.4 Small Exercises

Exercise 1.3.24. *Let X be the three-state Markov chain with transition rates*

$$Q = \begin{pmatrix} -\lambda & \frac{1}{3}\lambda & \frac{2}{3}\lambda \\ \frac{2}{3}\lambda & -\lambda & \frac{1}{3}\lambda \\ \frac{2}{3}\lambda & \frac{1}{3}\lambda & -\lambda \end{pmatrix},$$

for some $\lambda > 0$. This is represented by the state diagram in Figure 1.1.

1. *Explain why this is a Q-matrix.*
2. *Construct the jump matrix Π.*
3. *Find the class structure of the Markov chain.*
4. *Decide which states are recurrent and which are transient.*

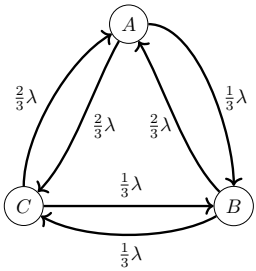

Figure 1.1. State diagram for three-state Markov chain

Exercise 1.3.25. *Consider the Markov chain with the following transition rates:*

$$Q = \begin{pmatrix} -2 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & -4 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Construct its jump matrix. What is its class structure?

1.4 Invariant Distribution and Reversibility

1.4.1 Invariant Distribution and Stochastic Equilibrium

One highly important aspect of a stochastic process is whether it settles down into a stochastic equilibrium. First, we introduce *invariant*, or *equilibrium*, *distributions*.

Definition 1.4.1 (Invariant Measure)**.** A measure λ on I is *invariant for a Q-matrix* Q if $\lambda Q = 0$, where $(\lambda Q)_j := \sum_{i \in I} \lambda_i q_{i,j}$ for all $j \in I$. A measure μ is *invariant for a stochastic matrix* P if $\mu P = \mu$. If the total mass of the measure is 1, then it defines a probability measure which we call an *invariant distribution*. △

Theorem 1.4.2. *Let Q be a Q-matrix with jump matrix Π. Let λ be a measure on I. Define the measure μ on I by $\mu_i := \lambda_i q_i$ for $i \in I$. Then, λ is an invariant measure for Q if and only if μ is an invariant measure for Π; ie $\lambda Q = 0$ if and only if $\mu \Pi = \mu$.*

Proof. From Lemma 1.3.5, the jump matrix $\Pi = (\pi_{i,j})_{i,j \in I}$ satisfies

$$q_{i,j} = q_i(\pi_{i,j} - \delta_{i,j}) \quad \text{where} \quad \delta_{i,j} := \mathbf{1}\{i = j\} \quad \text{for all} \quad i, j \in I.$$

Then,

$$\big(\mu(\Pi - I)\big)_j = \sum_i \mu_i(\pi_{i,j} - \delta_{i,j}) = \sum_i \lambda_i q_i(\pi_{i,j} - \delta_{i,j}) = \sum_i \lambda_i q_{i,j} = (\lambda Q)_j.$$

Hence, $\lambda Q = 0$ if and only if $\mu(\Pi - I) = 0$, ie $\mu\Pi = \mu$. □

Theorem 1.4.3. *An irreducible, continuous-time Markov chain on a finite state space has a unique invariant distribution.*

Proof. By Theorem 1.4.2, it suffices to prove uniqueness of the invariant distribution for the jump chain $\hat{X}$. First, irreducibility of X implies that $\hat{X}$ is also irreducible; finiteness of I then implies positive recurrence.

A measure μ satisfies $\mu\Pi = \mu$ if and only if $\mu(I+\Pi)/2 = \mu$. Hence, Π has a unique invariant distribution if and only if its *lazy* version $\Pi' := \frac{1}{2}(I+\Pi)$ does. Irreducibility and positive recurrence is preserved under lazification. Hence, the lazy version is ergodic. The ergodic theorem for discrete-time Markov chains (Theorem 0.3.3) then implies that it has a unique invariant distribution, completing the proof. □

Theorem 1.4.4. *Suppose that Q is an irreducible and recurrent Q-matrix and let $(P(t))_{t\geq 0}$ be the transition-probabilities semigroup that it generates. Then,*

$$\lambda Q = 0 \quad \textit{if and only if} \quad \lambda P(t) = \lambda \quad \textit{for all} \quad t \geq 0.$$

Proof. By definition of Q and time-homogeneity,

$$p_{i,j}(t+\delta) = p_{i,j}(t) + q_{i,j}\delta + o(\delta) \quad \text{for all} \quad i,j \in I \quad \text{and} \quad t \geq 0.$$

For $t \geq 0$, let $f(t) := \lambda P(t) \in \mathbb{R}^I$—ie, $f_j(t) = \sum_i \lambda_i p_{i,j}(t)$ for $j \in I$. Then,

$$\lambda P(t) = \lambda \quad \text{for all} \quad t \geq 0 \quad \text{if and only if} \quad f'(t) = 0 \in \mathbb{R}^I.$$

We now show that $f_j'(t) = (\lambda Q)_j$, from which the result immediately follows:

$$\begin{aligned}\tfrac{1}{\delta}\big(f_j(t+\delta) - f_j(t)\big) &= \tfrac{1}{\delta}\sum_i \lambda_i\big(p_{i,j}(t+\delta) - p_{i,j}(t)\big)\\ &= \sum_i \lambda_i\big(q_{i,j} + o(1)\big) = (\lambda Q)_j + o(1).\end{aligned}$$

Technically, we need uniformity of the first $o(1)$ term to deduce this if $|I| = \infty$, and to consider $\delta < 0$ too; we ignore these subtleties. This completes the proof. □

Theorem 1.4.5 (Ergodic Theorem)**.** *Suppose that Q is irreducible and recurrent on a finite state space. Suppose that λ is an invariant distribution for Q. Then,*

$$p_{i,j}(t) \to \lambda_j \quad \textit{as} \quad t \to \infty \quad \textit{for all} \quad i,j \in I.$$

The ergodic theorem says that the invariant distribution λ gives a description of the equilibrium behaviour, if it exists. The probability λ_j is the long-run probability of ending up in state j, regardless of the initial state.

Remark. We often stated and proved results for *finite*-state Markov chains. Typically, analogous results hold for countable state spaces, under the assumption of positive recurrence. The results are often harder to prove, requiring the error terms to be handled carefully, sometimes using the precision of measure theory. △

Example 1.4.6 (Switcher Model Continued)**.** Recall that the transition rates of the switcher chain are given by

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

Solving $\alpha Q = 0$ for α such that $\sum_i \alpha_i = 1$ gives

$$\alpha_1 = \frac{\mu}{\mu + \lambda} \quad \text{and} \quad \alpha_2 = \frac{\lambda}{\mu + \lambda}. \qquad \triangle$$

1.4.2 Reversibility and Detailed-Balanced Equations

Informally, a Markov chain is *reversible* if, when started from equilibrium, it is statistically impossible to tell whether it is being run forward or backward in time.

Solving the equilibrium, or *global balance*, equations $\sum_{j \in I} \pi_j q_{j,i} = 0$ for all $i \in I$ is often far from a pleasant experience. Many Markov chains satisfy an often-easier-to-solve set of equations, called the *detailed-balance* equations:

$$\pi_j q_{j,i} = \pi_i q_{i,j} \quad \text{for all} \quad i, j \in I.$$

The *detailed* balance equations require the 'probability flux' to be balanced across each pair of states; the *global* ones only require the total flux be balanced at each state.

We see later how detailed balance relates to a notion of *reversibility*. Before this, we get acquainted with some fundamental consequences of detailed balance. First, detailed balance implies global balance.

Definition 1.4.7 (Detailed Balance)**.** A matrix Q and a measure λ are in *detailed balance* if $\lambda_i q_{i,j} = \lambda_j q_{j,i}$ for all $i, j \in I$. △

Lemma 1.4.8 (Detailed Balance Implies Global Balance)**.** *If Q and λ are in detailed balance, then λ is an invariant measure for Q.*

Proof. This is a simple calculation, using detailed balance in the second equality:

$$(\lambda Q)_i = \sum_{j \in I} \lambda_j q_{j,i} = \sum_{j \in I} \lambda_i q_{i,j} = \lambda_i \sum_{j \in I} q_{i,j} = 0 \quad \text{for all} \quad i \in I. \qquad \square$$

This means that it is often worth looking for a solution to the detailed balance equations directly when trying to find the equilibrium distribution. After all, they are significantly simpler. Be warned, though: chains with a unique invariant distribution may not satisfy detailed balance! Eg, if $q_{i,j} \neq 0 = q_{j,i}$ for some $i, j \in I$, then detailed balance *can never* be satisfied.

Example 1.4.9 (Equivalence Between Jump- and Continuous-Time Chain)**.** Consider a continuous-time Markov chain with transition-rates matrix Q. Assume that $q_i \neq 0$ for all $i \in I$. Recall that the transition matrix Π of the jump chain satisfies

$$\pi_{i,j} = q_{i,j}/q_i \quad \text{for all} \quad i, j \in I \quad \text{with} \quad i \neq j.$$

Let λ and μ be measures on I satisfying $\mu_i = \lambda_i q_i$ for all $i \in I$. Then,

$$\lambda_i q_{i,j} = \lambda_j q_{j,i} \iff \mu_i \pi_{i,j} = \mu_j \pi_{j,i} \quad \text{for all} \quad i, j \in I.$$

The latter are the detailed balance equations for discrete-time Markov chains. △

Example 1.4.10 (Birth-and-Death Chains)**.** A continuous-time Markov chain X on $\mathbb{N}$ is called a *birth-and-death chain* if it can only increment by ± 1—ie, $q_{i,j} = 0$ if $|i-j| > 1$. Such chains *always* have an invariant measure satisfying detailed-balance, but may not be normalisable. Indeed, the detailed-balanced equations reduce to

$$\pi_n q_{n,n-1} = \pi_{n-1} q_{n-1,n} \quad \text{for all} \quad n \geq 1.$$

Rearranging this and iterating defines π:

$$\pi_n = \pi_{n-1} \frac{q_{n-1,n}}{q_{n,n-1}} = \ldots = \pi_0 \prod_{m=1}^{n} \frac{q_{m-1,m}}{q_{m,m-1}}.$$

This defines an invariant *distribution* if it is normalisable: ie, if

$$\sum_{n=0}^{\infty} \prod_{m=1}^{n} \frac{q_{m-1,m}}{q_{m,m-1}} < \infty, \quad \text{in which case} \quad \pi_0 = 1 \Big/ \sum_{n=0}^{\infty} \prod_{m=1}^{n} \frac{q_{m-1,m}}{q_{m,m-1}}.$$

There are two typical examples of this:

$$\left\{ \begin{matrix} q_{m,m-1} = \mu \\ q_{m-1,m} = \lambda \end{matrix} \right\} \quad \text{and} \quad \left\{ \begin{matrix} q_{m,m-1} = m\mu \\ q_{m-1,m} = \lambda. \end{matrix} \right\}.$$

These are $M/M/1$ and $M/M/\infty$ queues, respectively, to be encountered in Chapter 3.

$M/M/1$ Queue. Here, $q_{m-1,m}/q_{m,m-1} = \lambda/\mu$. So,

$$\pi_0^{-1} = \textstyle\sum_{n=0}^{\infty} (\lambda/\mu)^n < \infty \quad \text{if and only if} \quad \lambda < \mu.$$

$M/M/\infty$ Queue. Here, $q_{m-1,m}/q_{m,m-1} = \lambda/(m\mu)$. So,

$$\pi_0^{-1} = \sum_{n=0}^{\infty} \tfrac{1}{n!}(\lambda/\mu)^n = e^{\lambda/\mu} < \infty \quad \text{for all} \quad \lambda, \mu > 0. \qquad \triangle$$

We now discuss the concept of reversibility. It is analogous to the discrete-time set-up. There is one minor technicality which we address upfront. Our continuous-time Markov chains $X = (X_t)_{t\geq 0}$ are defined to be right-continuous: $X_t = \lim_{s\downarrow t} X_s$. However, if we fix $T > 0$ and set $\hat{X}_t := X_{T-t}$, then the process $(X_t)_{0\leq t\leq T}$ is *left*-continuous. Instead, officially, we need to set $X_t := \lim_{s\uparrow t} X_{T-s}$ to obtain the *right*-continuous version of the time-reversal of $(X_t)_{t\geq 0}$. We ignore this technicality.

Theorem 1.4.11. *Fix $T \in (0, \infty)$. Let $X = (X_t)_{t\geq 0}$ be a Markov chain with instantaneous transition-rates matrix Q which is irreducible and non-explosive. Suppose that the initial distribution π of X is invariant for Q. For $i, j \in I$ and $t \geq 0$, let*

$$\hat{q}_{j,i} := \pi_i q_{i,j}/\pi_j \quad \text{and} \quad \hat{p}_{j,i}(t) := \pi_i p_{i,j}(t)/\pi_j.$$

Let $(\hat{X}_t := X_{T-t})_{t\in[0,T]}$ be the right-continuous version of the time-reversal of $(X_t)_{t\in[0,T]}$. Then, $(\hat{X}_t)_{t\in[0,T]}$ is a Markov chain with initial distribution π, instantaneous transition-rates matrix $\hat{Q}$ and transition probabilities $\hat{P} = ((\hat{p}_{i,j}(t))_{i,j\in I})_{t\geq 0}$. Moreover, $(\hat{X}_t)_{t\in[0,T]}$ is also irreducible, non-explosive and has initial distribution π.

We delay this proof until we can use *Kolmogorov differential equations* in §1.5.

A chain is *reversible* if $\hat{Q} = Q$—ie, $(X_{T-t})_{0\leq t\leq T}$ has transition-rates matrix Q.

Definition 1.4.12 (Reversibility)**.** Let $X = (X_t)_{t\geq 0}$ be a Markov chain with instantaneous transition-rates matrix Q which is irreducible and non-explosive. It is *reversible* if its invariant distribution π is in detailed balance with Q:

$$\pi_i q_{i,j} = \pi_j q_{j,i} \quad \text{for all} \quad i, j \in I. \qquad \triangle$$

Remark (Why "Reversible"?)**.** The 'probability flux' is equal in both directions.

- $q_{i,j}$ is the rate at which the chain jumps to j if it is at i.
- π_i is the proportion of time that the chain is at i in equilibrium.
- So, $\pi_i q_{i,j}$ is the i-to-j 'rate', or 'flux', in equilibrium.
- "Reversibility" simply means that the i-to-j flux equals the j-to-i flux. $\triangle$

Remark (Preservation of Reversibility Under Restriction)**.** The following remark is of major importance in applications—eg, when calculating conditional probabilities for Bayesian posterior distributions using Markov chain Monte Carlo.

The property of reversibility easily carries over to Markov chains X' which are *restrictions* of the original chain X. If $I' \subseteq I$ and we forbid the chain to leave I', by setting $q'_{i,j} := q_{i,j}\mathbf{1}\{i,j \in I'\}$ for $i,j \in I$ and $q_{i,i}$ appropriately for $i \in I$, then the new Markov chain X' is still reversible wrt the invariant distribution of X. △

1.5 The Kolmogorov Differential Equations

We have constructed the transition-rates matrix Q from the transition probabilities semigroup P by taking derivatives. We will see that the opposite also holds: under favourable regularity conditions, typically satisfied in common applications, it is possible to construct P from a Q-matrix. The law of the Markov chain is thus characterised by just Q and the initial distribution. This is achieved by solving the

Kolmogorov Differential Equations (KDEs).

1.5.1 The KDEs and Examples

The following two theorems do appear in [Nor97, §2.8], but in a rather different form and set-up to ours. See, instead, [Bré20, §13] for a more similar proof.

Theorem 1.5.1 (KBDE). *Suppose that X is a continuous-time Markov chain with transition probabilities P and corresponding transition rates Q. Then, P and Q satisfy the Kolmogorov Backward Differential Equation (KBDE):*

$$\tfrac{d}{dt}P(t) = QP(t), \quad P(0) = I.$$

This is equivalent to the following by-entry version:

$$\begin{aligned} \tfrac{d}{dt}p_{i,j}(t) &= \textstyle\sum_{k\in I} q_{i,k}p_{k,j}(t) \quad \text{for all} \quad i,j \in I; \\ p_{i,j}(0) &= \mathbf{1}\{i=j\} \qquad\qquad\quad \text{for all} \quad i,j \in I. \end{aligned}$$

Proof (Sketch). Consider the time interval $(0, t+\delta] = (0,\delta] \dot\cup (\delta, t+\delta]$ and suppose that δ is small. By definition, $q_{i,k} = p'_{i,k}(t)$. Thus, by Taylor's theorem,

$$p_{i,k}(\delta) = p_{i,k}(0) + q_{i,k}\delta + \varepsilon_{i,k}(\delta)\delta \quad \text{for} \quad \delta > 0$$

for some function $\varepsilon_{i,k}(\cdot)$ with $\lim_{\delta\to 0}\varepsilon_{i,k}(\delta) = 0$. This can be written succinctly as

$$p_{i,k}(\delta) = p_{i,k}(0) + q_{i,k}\delta + o(\delta),$$

but this does hide dependence on the states. If $|I| < \infty$, though, then the $o(\delta)$ error is uniform over the states, as there are only finitely many to take a maximum over.

Suppose that $X_0 = i$ and average over $X_\delta = k$, using CK (Theorem 1.2.5):

$$\begin{aligned} p_{i,j}(t+\delta) &= \textstyle\sum_{k\in I} p_{i,k}(\delta)p_{k,j}(t) \\ &= \textstyle\sum_{k\in I} \big(p_{i,k}(0) + q_{i,k}\delta + \varepsilon_{i,k}(\delta)\delta\big)p_{k,j}(t) \\ &= p_{i,j}(t) + \delta \textstyle\sum_{k\in I} q_{i,k}p_{k,j}(t) + \delta \sum_{k\in I} \varepsilon_{i,k}(\delta)p_{k,j}(t), \end{aligned}$$

using $p_{i,k}(0) = \mathbf{1}\{i = k\}$. Rearranging and taking the limit $\delta \to 0$ gives

$$p'_{i,j}(t) = \textstyle\sum_{k\in I} q_{i,k}p_{k,j}(t) + \lim_{\delta\to 0} \sum_{k\in I} \varepsilon_{i,k}(\delta)p_{k,j}(t).$$

Technically, we also need to look at $\delta < 0$. We ignore this subtlety here.

If we can swap the limit and the sum, then we would be done as

$$\lim_{\delta\to 0} \varepsilon_{i,k}(\delta) = 0, \quad \text{so then} \quad \tfrac{d}{dt}p_{i,j}(t) = \textstyle\sum_{k\in I} q_{i,k}p_{k,j}(t).$$

If $|I| < \infty$, then we can always swap the limit and the now-*finite* sum. However, for countably infinite I, it is not obvious. We do not explore this further. We simply assume that the swapping is legitimate. It is in all the examples we care about. □

A *forward* version also holds under typically-satisfied regularity conditions.

Theorem 1.5.2 (KFDE). *Suppose that X is a continuous-time Markov chain with transition probabilities P and corresponding transition rates Q. Assume that $\sum_{k\in I} p_{i,k}(t)q_k < \infty$ for all $i \in I$ and $t \geq 0$. Then, P and Q satisfy the Kolmogorov Forward Differential Equation (KFDE):*

$$\tfrac{d}{dt}P(t) = P(t)Q, \quad P(0) = I.$$

This is equivalent to the following by-entry version:

$$\begin{aligned} \tfrac{d}{dt}p_{i,j}(t) &= \textstyle\sum_{k\in I} p_{i,k}(t)q_{k,j} \quad \text{for all} \quad i,j \in I; \\ p_{i,j}(0) &= \mathbf{1}\{i = j\} \qquad\qquad\quad \text{for all} \quad i,j \in I. \end{aligned}$$

Proof (Sketch). The proof is similar to that of Theorem 1.5.1 except that more care is required to justify an exchange of limit and sum. The main difference is that we average over $X_t = k$, rather than $X_\delta = k$. The Chapman–Kolmogorov equations give

$$p_{i,j}(t+\delta) = \textstyle\sum_{k\in I} p_{i,k}(t)p_{k,j}(\delta).$$

The rest of the proof is analogous, but even more care is required in swapping the limit and the sum in the case of infinite I; it is always fine for finite I, though. □

Remark (Existence and Uniqueness of KDE Solutions). When choosing whether to solve the *forward* or *backward* equations, bear in mind the following points.

- The KBDEs have *at least* one solution: a solution always exists, but it need not be unique. We are interested in the minimal solution if there are multiple.

  We have uniqueness for finite state spaces and in general there is always a *minimal* solution. Uniqueness can fail if the chain explodes.
- The KFDEs have *at most* one solution: a solution is unique *if* it exists.

We do not worry about lack of existence or uniqueness. In most cases we consider, both KDEs have unique solutions, so the rates specify the Markov chain uniquely. △

Now that we have introduced the KDEs, we can give the deferred proof of Theorem 1.4.11 on time reversal. We also give an alternative proof of Theorem 1.4.4.

Proof of Theorem 1.4.11 (Assuming I Is Finite). By the KDEs and finiteness of I, the semigroup $P = (P(t))_{t\geq 0}$ of Q is the unique solution of the forward equation

$$P'(t) = P(t)Q \quad \text{with} \quad P(0) = I.$$

Inputting $\hat{p}_{j,i}(t) = \pi_i p_{i,j}(t)/\pi_j$ and using the forward equation,

$$\hat{P}'(t) = \hat{Q}\hat{P}(t) \quad \text{with} \quad \hat{P}(0) = I.$$

But, this is precisely the backward equation for $\hat{Q}$, which is itself a Q-matrix. Thus, $\hat{P} = (\hat{P}(t))_{t\geq 0}$ is the semigroup of $\hat{Q}$. Also, π is invariant for $\hat{P}(t)$ for all $t \geq 0$.

Finally, whenever $0 = t_0 < \ldots < t_n = T$, letting $s_k := t_k - t_{k-1}$, we have

$$\begin{aligned}&\mathbb{P}\{\hat{X}_{t_0} = i_0, \ldots, \hat{X}_{t_n} = i_n\} = \mathbb{P}\{X_{T-t_0} = i_0, \ldots, X_{T-t_n} = i_n\}\\ &= \pi_{i_n} p_{i_n,i_{n-1}}(s_n) \cdot \ldots \cdot p_{i_1,i_0}(s_1) = \pi_{i_0}\hat{p}_{i_0,i_1}(s_1) \cdot \ldots \cdot \hat{p}_{i_{n-1},i_n}(s_n).\end{aligned}$$

Hence, $(\hat{X}_t)_{0\leq t\leq T}$ is a Markov chain with transition probabilities $\hat{P} = (\hat{P}(t))_{t\geq 0}$. □

Alternative Proof of Theorem 1.4.4: "$\lambda Q = 0 \Leftrightarrow \lambda P(t) = \lambda$". From the KBDE,

$$\tfrac{d}{dt}(\lambda P(t)) = \lambda \tfrac{d}{dt} P(t) = \lambda(QP(t)) = (\lambda Q)P(t).$$

If $\lambda Q = 0$, then $\frac{d}{dt}(\lambda P(t)) = 0$, so $\lambda P(t) = \lambda P(0) = \lambda$ for all $t \geq 0$. Conversely, if $\lambda P(t) = \lambda$ for all $t \geq 0$, then $\frac{d}{dt}(\lambda P(t)) = 0$, so $\lambda Q = 0$ by taking $t = 0$ above. □

The KDEs allow us to compute P from the Q-matrix. This is easier said than done, though. The equations are often note straightforward to solve, even when there is a unique solution. The method of integrating factors is a useful tool for solving such equations. To get some intuition for how this works, we consider a few examples.

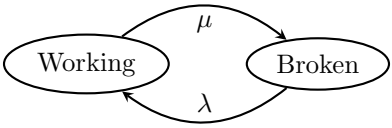

Figure 1.2. State diagram for the switcher model

Example 1.5.3 (Switcher Model)**.** A two-state stochastic process is commonly referred to as the *switcher model*: it is used to model systems which switch between two states—eg, "working" and "broken". Its state diagram is shown in Figure 1.2. The instantaneous transition-rate matrix for this model is

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

Since $P(t)$ is a stochastic matrix for all t, it can be parametrised by just two unknowns:

$$P(t) = \begin{pmatrix} \phi(t) & 1-\phi(t) \\ 1-\psi(t) & \psi(t) \end{pmatrix} \quad \text{for all} \quad t \geq 0.$$

Method 1: *Using integrated factors to solve the KFDEs.* The KFDE system is a matrix product which we can write out in full:

$$\overbrace{\begin{pmatrix} \phi'(t) & -\phi'(t) \\ -\psi'(t) & \psi'(t) \end{pmatrix}}^{P'(t)} = \overbrace{\begin{pmatrix} \phi(t) & 1-\phi(t) \\ 1-\psi(t) & \psi(t) \end{pmatrix}}^{P(t)} \cdot \overbrace{\begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}}^{Q}$$
$$= \begin{pmatrix} -\lambda\phi + \mu(1-\phi) & \lambda\phi - \mu(1-\phi) \\ -\lambda(1-\psi) + \psi\mu & \lambda(1-\psi) - \mu\psi \end{pmatrix}.$$

We thus get the following *autonomous* DEs[3]:

$$\phi'(t) = -\lambda\phi(t) + \mu\big(1-\phi(t)\big), \quad \phi(0) = 1;$$
$$\psi'(t) = -\mu\psi(t) + \lambda\big(1-\psi(t)\big), \quad \psi(0) = 1.$$

The two equations are equivalent, up to swapping λ and μ, so it suffices to solve only the first. Rearranging it gives

$$\phi'(t) + (\lambda+\mu)\phi(t) = \mu \quad \text{with} \quad \phi(0) = 1.$$

[3] an *autonomous* DE is one which involves only one function of time; compare this with *coupled* DEs which have multiple functions of time, as is the case in **Method 2**

In the *method of integrating factors*, one searches for some function f such that

$$\tfrac{d}{dt}\phi(t) + (\lambda+\mu)\phi(t) = f(t)^{-1}\tfrac{d}{dt}\big(f(t)\phi(t)\big).$$

This only defines f up to a constant factor. Expanding the derivative and simplifying,

$$f'(t)/f(t) = \lambda+\mu, \quad \text{so we take} \quad f(t) \coloneqq e^{(\lambda+\mu)t}.$$

Using these equations, we obtain

$$\begin{aligned}\mu &= \tfrac{d}{dt}\phi(t) + (\lambda+\mu)\phi(t)\\ &= f(t)^{-1}\tfrac{d}{dt}\big(f(t)\phi(t)\big) = e^{-(\lambda+\mu)t}\tfrac{d}{dt}\big(e^{(\lambda+\mu)t}\phi(t)\big);\end{aligned}$$

ie,

$$\tfrac{d}{dt}\big(e^{(\lambda+\mu)t}\phi(t)\big) = e^{(\lambda+\mu)t}\mu.$$

Integrating both sides and using the initial conditions, we solve the equations:

$$\begin{aligned}\phi(t) &= \tfrac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} + \tfrac{\mu}{\lambda+\mu};\\ \psi(t) &= \tfrac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t} + \tfrac{\lambda}{\lambda+\mu}.\end{aligned}$$

Method 2: *Solving the KBDE directly.* The KBDE system is a matrix product which we can write out in full:

$$\begin{aligned}\overbrace{\begin{pmatrix}\phi'(t) & -\phi'(t)\\ -\psi'(t) & \psi'(t)\end{pmatrix}}^{P'(t)} &= \overbrace{\begin{pmatrix}-\lambda & \lambda\\ \mu & -\mu\end{pmatrix}}^{Q}\cdot\overbrace{\begin{pmatrix}\phi(t) & 1-\phi(t)\\ 1-\psi(t) & \psi(t)\end{pmatrix}}^{P(t)}\\ &= \begin{pmatrix}-\lambda\phi+\lambda(1-\psi) & -\lambda(1-\phi)+\lambda\psi\\ \mu\phi-\mu(1-\psi) & \mu(1-\phi)-\mu\psi\end{pmatrix}.\end{aligned}$$

This is equivalent to the following *coupled*, or *linked*, DEs:

$$\begin{aligned}\phi'(t) &= \lambda\big(1-\phi(t)-\psi(t)\big)\\ \psi'(t) &= \mu\big(1-\phi(t)-\psi(t)\big).\end{aligned}$$

There are multiple ways of solving this. One way is to differentiate again:

$$\begin{aligned}\phi''(t) &= -\lambda\phi'(t) - \lambda\psi'(t)\\ &= -\lambda\phi'(t) - \lambda\cdot\mu\big(1-\phi(t)-\psi(t)\big)\\ &= -\lambda\phi'(t) - \mu\cdot\lambda\big(1-\phi(t)-\psi(t)\big)\\ &= -\lambda\phi'(t) - \mu\phi'(t)\\ &= -(\lambda+\mu)\phi'(t).\end{aligned}$$

This can then be solved in the usual way, noting that

$$\phi'(0) = -\lambda\big(1 - \phi(0) - \psi(0)\big) = -\lambda.$$

The same solution as before is obtained.

Special Case: $\lambda = \mu$. Without loss of generality, $\lambda = \mu = 1$. A trick for solving coupled DEs which comes up surprisingly often is to look for linear combinations of (ϕ, ψ) which decouple the DEs.[4] Let $(\alpha, \beta) \coloneqq (\phi + \psi, \phi - \psi)$:

$$\begin{aligned} \alpha'(t) &= \phi'(t) + \psi'(t) = 2 - 2\alpha(t), & \alpha(0) &= \phi(0) + \psi(0) = 2; \\ \beta'(t) &= \phi'(t) - \psi'(t) = 0, & \beta(0) &= \phi(0) - \psi(0) = 0. \end{aligned}$$

This is a set of decoupled DEs which can be solved easily for (α, β), then converted into a solution for (ϕ, ψ). △

Example 1.5.4. Suppose X is a continuous-time Markov chain with transition rates

$$Q = \begin{pmatrix} -\lambda & \lambda/2 & \lambda/2 \\ \lambda/2 & -\lambda & \lambda/2 \\ \lambda/2 & \lambda/2 & -\lambda \end{pmatrix}.$$

Our goal is to find the transition probabilities P.

Fix a time $t \geq 0$. There are 9 unknowns: $p_{i,j}(t)$ for $i, j \in \{1, 2, 3\}$. The rates depend only on whether we stay put or jump to another state. So, $p_{i,i}(t)$ does not depend on i and $p_{i,j}(t)$ does not depend on either i or j if $i \neq j$. With this in mind, let

$$\left\{ \begin{matrix} \phi(t) \coloneqq p_{1,1}(t) \\ \psi(t) \coloneqq p_{1,2}(t) \end{matrix} \right\} \quad \text{so that} \quad P(t) = \begin{pmatrix} \phi(t) & \psi(t) & \psi(t) \\ \psi(t) & \phi(t) & \phi(t) \\ \psi(t) & \psi(t) & \phi(t) \end{pmatrix}.$$

Additionally, $P(t)$ has unit row-sums, so

$$\phi(t) + 2\psi(t) = 1.$$

The forward equations give

$$\begin{aligned} \phi'(t) &= -\lambda\phi(t) + \tfrac{1}{2}\lambda\psi(t) + \tfrac{1}{2}\lambda\psi(t) \\ &= -\lambda\phi(t) + \lambda\psi(t) \\ &= -\lambda\phi(t) + \tfrac{1}{2}\lambda\big(1 - \phi(t)\big) \end{aligned}$$

[4] Formally, this is equivalent to finding a basis in which a certain matrix is diagonal, which corresponds to decoupled/autonomous differential equations: if $\mathbf{x}'(t) = M\mathbf{x}(t) + \mathbf{b}$ and $M = UDU^{-1}$ with D diagonal, then $\mathbf{y}'(t) = D\mathbf{y}(t) + \mathbf{c}$, ie $y_i'(t) = D_{i,i}y_i(t) + c_i$ for all i, where $\mathbf{y}(t) \coloneqq U^{-1}\mathbf{x}(t)$ and $\mathbf{c} \coloneqq U^{-1}\mathbf{b}$

$$= \tfrac{1}{2}\lambda - \tfrac{3}{2}\lambda\phi(t).$$

We again look for an integrating factor:

$$f(t)^{-1}\tfrac{d}{dt}\big(f(t)\phi(t)\big) = \phi'(t) + \tfrac{3}{2}\lambda\phi(t)$$

implies

$$f'(t)/f(t) = \tfrac{3}{2}\lambda, \quad \text{so we take} \quad f(t) = \exp\big(\tfrac{3}{2}\lambda t\big).$$

Substituting this back in and integrating gives

$$\phi(t) = \tfrac{1}{3} + \tfrac{2}{3}\exp\big(-\tfrac{3}{2}\lambda t\big).$$

Finally, using $\phi(t) + 2\psi(t) = 1$, we get

$$\psi(t) = \tfrac{1}{3} - \tfrac{1}{3}\exp\big(-\tfrac{3}{2}\lambda t\big). \qquad \triangle$$

Exercise 1.5.5. *Suppose that*

$$Q = \begin{pmatrix} -\lambda & p\lambda & (1-p)\lambda \\ (1-q)\mu & -\mu & q\mu \\ r\gamma & (1-r)\gamma & -\gamma \end{pmatrix}.$$

Find $P = (P(t))_{t\geq 0}$. This example is expanded upon in Example Sheet 1.

1.5.2 Solving KDEs Abstractly via Matrix Exponentials

Another way of constructing the transition probabilities semigroup P from a Q-matrix is by computing the *exponential of the matrix* tQ, denoted by e^{tQ}.

Definition 1.5.6 (Matrix Exponential)**.** For a square matrix A, the exponential e^A is

$$e^A := \sum_{k=0}^{\infty} \frac{A^k}{k!}, \quad \text{where } A^k \text{ is the } k\text{-th power of the matrix.}$$

This always converges for finite-dimensional matrices. So, for finite I, we may set

$$e^{tQ} := \sum_{k=0}^{\infty} \frac{t^k Q^k}{k!} \quad \text{for all} \quad t \geq 0. \qquad \triangle$$

We assume that I is finite for the rest of the section. The methods also apply for the case where I is countable, provided that the series converges.

The following two theorems imply that $(e^{tQ})_{t\geq 0}$ is a semigroup of stochastic matrices solving the KDEs. In other words, Q determines the transition-probabilities semigroup $P = (P(t))_{t\geq 0}$ of the Markov chain with transition rates Q.

Theorem 1.5.7. *Let Q be a Q-matrix[5] on a finite set I. Set $P(t) \coloneqq e^{tQ}$ for $t \geq 0$. Then, $P = (P(t))_{t\geq 0}$ has the following properties.*

1. *P is a semigroup: $P(t+s) = P(t)P(s)$ for all $s, t \geq 0$ and $P(0) = I$, the identity.*
2. *P is the unique solution to the KBDE:*
$$P'(t) = QP(t), \quad P(0) = I.$$
3. *P is the unique solution to the KFDE:*
$$P'(t) = P(t)Q, \quad P(0) = I.$$
4. *For $k \in \{1, 2, ...\}$,*
$$\tfrac{d^k}{dt^k} P(t)\Big|_{t=0} = Q^k.$$

Proof. We use the following two properties of matrix exponentials without proof.

- If A and B are square matrices that commute, ie $AB = BA$, then
$$e^{A+B} = e^A e^B.$$
- The matrix-valued power series given by
$$t \mapsto \textstyle\sum_{k\geq 0} (tQ)^k/k!$$
has infinite radius of convergence.

Clearly, $P(0) = e^{0Q} = I$, the identity. The matrices sQ and tQ commute, so
$$P(t)P(s) = e^{tQ}e^{sQ} = e^{(t+s)Q} = P(t+s).$$

Proving that P is a solution to the KDEs is straightforward:
$$\begin{aligned}
\tfrac{d}{dt} P(t) &= \tfrac{d}{dt}\big(\textstyle\sum_{k\geq 0} t^k Q^k/k!\big) \\
&= \textstyle\sum_{k\geq 0} \big(\tfrac{d}{dt} t^k\big) Q^k/k! \\
&= \textstyle\sum_{k\geq 0} t^{k-1} Q^k/(k-1)! \\
&= \textstyle\sum_{\ell\geq 0} t^\ell Q^{\ell+1}/\ell!
\end{aligned}$$

But, $Q^\ell \cdot Q = Q^{\ell+1} = Q \cdot Q^\ell$. From this, it follows that
$$P(t)Q = \big(\textstyle\sum_{\ell\geq 0} t^\ell Q^\ell/\ell!\big)Q = \tfrac{d}{dt} P(t) = Q\big(\textstyle\sum_{\ell\geq 0} t^\ell Q^\ell/\ell!\big) = QP(t).$$

[5] see Definition 1.3.13 to recall the conditions for a matrix to be a *Q-matrix*

We now need to show uniqueness. Suppose that $R = (R(t))_{t\geq 0}$ is another family of matrices satisfying the KBDE. Then,

$$\tfrac{d}{dt}(P(t)^{-1}R(t)) = \tfrac{d}{dt}(e^{-tQ}R(t)) = (-e^{-tQ}Q)R(t) + e^{-tQ}(QR(t)) = 0.$$

It follows that $t \mapsto P(t)^{-1}R(t)$ is constant—independent of t. But, $P(0) = I = R_0$. Thus, $P(t) = R(t)$ for all $t \geq 0$. Uniqueness for KFDE is proved similarly.

Iterating the KBDEs, using linearity and finiteness of the state space, gives

$$\tfrac{d^k}{dt^k}P(t) = Q\tfrac{d^{k-1}}{dt^{k-1}}P(t) = \ldots = Q^kP(t).$$

The final claim is established by taking $t \coloneqq 0$ as $P(0) = I$. □

Theorem 1.5.8. *A matrix Q on a finite set I is a Q-matrix if and only if $P(t) \coloneqq e^{tQ}$ is a stochastic matrix for all $t \geq 0$.*

Proof. We have seen that $P(t) \coloneqq e^{tQ}$ satisfies $Q = P'(0)$. So,

$$P(t) = I + tQ + \mathcal{O}(t^2) \quad \text{as} \quad t \downarrow 0.$$

Thus, $q_{i,j} \geq 0$ if and only if $p_{i,j}(t) \geq 0$ for all $t \geq 0$ sufficiently small, for $i \neq j$. But $P(t) = P(t/n)^n$ for all n, so $q_{i,j} \geq 0$ if and only if $p_{i,j}(t) \geq 0$ for all $t \geq 0$, for $i \neq j$.

If Q has null row-sums, then $\mathbf{1}$—the vector consisting only of 1s—is an eigenvector of Q with eigenvalue 0: $Q\mathbf{1} = \mathbf{0} = 0\mathbf{1}$. Hence,

$$Q^n\mathbf{1} = Q^{n-1}Q\mathbf{1} = \mathbf{0}.$$

Null row-sums for Q^n for all $n \geq 0$ implies that $P(t)$ has unit row-sums for every $t \geq 0$:

$$P(t)\mathbf{1} = (\textstyle\sum_{n\geq 0} t^nQ^n/n!)\mathbf{1} = I\mathbf{1} + \sum_{n\geq 1} t^n(Q^n\mathbf{1})/n! = \mathbf{1}.$$ □

1.5.3 Constructing Matrix Exponentials via Eigenstatistics

We have seen that the family of matrices $(P(t) \coloneqq e^{tQ})_{t\geq 0}$ solves the KDEs. An efficient way of computing all powers of Q is to diagonalise the matrix: try to write

$$Q = U\Lambda U^{-1} \quad \text{where } \Lambda \text{ is diagonal and } U \text{ is invertible.}$$

When this is possible, we can choose $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ to be the diagonal matrix of eigenvalues and $U = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ to be the matrix of (right-)eigenvectors; that is, $Q\mathbf{v}_i = \lambda_i\mathbf{v}_i$ for all i.[6] Then, calculating Q^k is easy:

$$Q^k = (U\Lambda U^{-1}) \cdot (U\Lambda U^{-1}) \cdots (U\lambda U^{-1}) = U\Lambda^k U^{-1} \quad \text{for} \quad k \in \mathbb{N}.$$

[6] Typically, we take $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ to be an orthonormal set of eigenvectors so that U is a unitary matrix—ie, $U^\dagger = U^{-1}$—when diagonalising *abstractly*, but this is not necessary *practically*

Plugging this into the power series,

$$\textstyle\sum_{k\geq 0} t^k Q^k/k! = \sum_{k\geq 0} t^k U\Lambda^k U^{-1} = U\big(\sum_{k\geq 0}(t\Lambda)^k/k!\big)U^{-1} = Ue^{t\Lambda}U^{-1}$$

and $\Lambda^k = \operatorname{diag}(\lambda_1^k, \ldots, \lambda_n^k)$ for $k \geq 0$, so

$$e^{t\Lambda} = \operatorname{diag}(e^{t\lambda_1}, \ldots, e^{t\lambda_n}).$$

So, the (candidate) solution for $P = (P(t))_{t\geq 0}$ is given by

$$P(t) \coloneqq Ue^{t\Lambda}U^{-1} \quad \text{for} \quad t \geq 0.$$

Finding all the eigenvalues and eigenvectors is often computationally intensive. It is manageable for 2×2 matrices, though, as in the *Switcher Model* (Example 1.5.3).

Example 1.5.9 (Switcher Model Continued)**.** Recall that the transition matrix is

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

We need to compute the exponential e^{tQ}.

First, we compute the eigenvalues of Q, which involves solving

$$\det(Q - \theta I) = 0 \quad \text{for} \quad \theta.$$

This leads to the characteristic equation

$$\theta(\theta + \lambda + \mu) = 0.$$

Hence, the eigenvalues are $\theta_1 = 0$ and $\theta_2 = -(\lambda + \mu)$. Solving for the eigenvectors,

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_2 = \begin{pmatrix} \lambda \\ -\mu \end{pmatrix}.$$

Hence, our desired matrices are as follows:

$$\Lambda = \begin{pmatrix} 0 & 0 \\ 0 & -(\lambda+\mu) \end{pmatrix} \quad \text{and} \quad e^{t\Lambda} = \begin{pmatrix} 1 & 0 \\ 0 & e^{-t(\lambda+\mu)} \end{pmatrix};$$
$$U = \begin{pmatrix} 1 & \lambda \\ 1 & -\mu \end{pmatrix} \quad \text{and} \quad U^{-1} = \frac{1}{\lambda+\mu}\begin{pmatrix} \mu & \lambda \\ 1 & -1 \end{pmatrix}.$$

It follows that

$$P(t) = Ue^{t\Lambda}U^{-1} = \frac{1}{\lambda+\mu}\begin{pmatrix} \mu + \lambda e^{-t(\lambda+\mu)} & \lambda - \lambda e^{-t(\lambda+\mu)} \\ \mu - \mu e^{-t(\lambda+\mu)} & \lambda + \mu e^{-t(\lambda+\mu)} \end{pmatrix}. \qquad \triangle$$

Remark. Usually, finding the eigenvalues for 2×2 matrices is easy but hard for 3×3 matrices, since there is no nice "cubic formula". However, $\mathbf{1}$ is *always* an eigenvector of a Q-matrix with eigenvalue 0, since it corresponds to taking row-sums. So, finding the eigenvalues for 3×3 matrices really only corresponds to solving a *quadratic*. $\triangle$

The simplification in the previous remark applies primarily to eigen*values*. Knowing that $\mathbf{1}$ is an eigenvector with eigenvalue 0 does not particularly help in finding the other eigen*vectors*. Finding the remaining eigen*vectors*, possibly by inverting a 3×3 matrix or using Gaussian elimination, is still tedious.

If one only desires some elements of $P(t) = (p_{i,j}(t))_{i,j \in I}$, then an alternative, simultaneous-equations approach can be used. We describe this now via an example.

Example 1.5.10. Consider the three-state Markov chain with transition rates

$$Q = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -1 & 0 \\ 2 & 1 & -3 \end{pmatrix}.$$

Suppose that we only wish to know $p_{1,1}(t)$ for each $t \geq 0$.

First, we compute the eigenvalues. The characteristic equation of Q is

$$-\lambda(\lambda + 4)(\lambda + 2) = 0.$$

Thus, the eigenvalues are $\lambda_1 = 0$ (always an eigenvalue), $\lambda_2 = -2$ and $\lambda_3 = -4$. Now,

$$P(t) = Ue^{t\Lambda}U^{-1} \quad \text{where} \quad e^{t\Lambda} = \operatorname{diag}(1, e^{-2t}, e^{-4t})$$

for some matrix U. Rather than find and invert U, we just use the fact that $p_{1,1}(t)$ is *some* linear combination, independent of t, of $(1, e^{-2t}, e^{-4t})$:

$$p_{1,1}(t) = a + be^{-2t} + ce^{-4t} \quad \text{for some constants} \quad a, b, c \in \mathbb{R}.$$

We need to determine the constants. Initial conditions come from Theorem 1.5.7(4):

$$p_{1,1}(0) = 1, \quad p'_{1,1}(0) = q_{1,1} = -2, \quad p''_{1,1}(0) = (Q^2)_{1,1} = 7.$$

This leads to the following system of equations:

$$\begin{array}{rcrcrcr} a & + & b & + & c & = & 1, \\ & - & 2b & - & 4c & = & -2, \\ & & 4b & + & 16c & = & 7. \end{array}$$

Solving it gives $(a, b, c) = (\frac{3}{8}, \frac{1}{4}, \frac{3}{8})$ and hence

$$p_{1,1}(t) = \tfrac{3}{8} + \tfrac{1}{4}e^{-2t} + \tfrac{3}{8}e^{-4t}. \qquad \triangle$$

Remark. If the *invariant distribution* at j is known, say to be π_j, then, under mild assumptions, $p_{i,j}(t) \to \pi_1$ as $t \to \infty$. This often saves calculating the square $(Q^2)_{1,1}$. It is tractable since $p_{i,j}(t) \to a$ as $t \to \infty$, so this immediately gives $a = \pi_j$. $\triangle$

Remark. The final system of equations above is 3×3, which is often tedious to solve. Notice, however, that the constant term is removed as soon as even one derivative is taken. So, fundamentally, it is only 2×2, with the added definition $a := 1 - b - c$. $\triangle$

The matrix-exponential approach is attractively powerful for small state-space examples. However, it requires us to solve the characteristic equation, which is n-th order if there are n states, and then do further work with $n \times n$ matrices. Mathematical software can handle modestly-sized n, though we would need to explore numerical analysis methods to have practical ways of determining eigenvalues. But, n can be very large indeed. For example, in image analysis, a small 128×128 black–white pixel image leads to a state-space with $2^{128 \times 128} = 2^{16384}$ states, which is somewhere near 10^{5000}. Realistic agent-based modelling for the COVID-19 pandemic involved systems with a couple of million entities; again, in the simplest case, one expects k entities to lead to 2^k states.

We need to be able to understand properties of Markov chains *without* making explicit computations. This is where *theory* trumps simple-minded *computation*.

From a theoretical point of view, it is possible to vastly to generalize the above approaches using the theory of semigroups of operators on infinite-dimensional Banach spaces; see, in particular, the Hille–Yosida theorem. This offers the opportunity to deal with applications where our *Standing Assumptions* may not all apply. An exploration of these results would take this module far too far off track.

2 Birth–Death Processes

A *birth-and-death*, or *birth–death*, process is a continuous-time Markov chain with state space $\mathbb{N} \cup \{\infty\}$ where the only non-zero transition rates $q_{i,j}$ have $|i - j| \leq 1$:

- a "birth" corresponds to increasing the state by 1;
- a "death" corresponds to decreasing the state by 1.

The state space includes $\infty \notin \mathbb{N}$ as we need to allow for the possibility of populations' growing to infinite size in finite time: "explosion".

A *linear birth–death process* is a birth–death process with rates $q_{i,i\pm1}$ depending linearly on the current state, with a possible added constant:

$$\begin{aligned} q_{i,i-1} &= a_- + b_- i, \qquad \text{for all} \quad i \in \mathbb{N}, \\ q_{i,i+1} &= a_+ + b_+ i, \qquad \text{for some} \quad a_\pm, b_\pm \in \mathbb{R}. \end{aligned}$$

We will study the linear birth–death process in detail. We start with the simplest example: the Poisson process, which has $q_{i,j} = \mathbf{1}\{j = i + 1\}$;

2.1 The Poisson Process

The Poisson process models the growth in a number of occurrences of an event occurring by time t, as t increases. Events are suitable for modelling as a Poisson process if

- the probability of an event's occurring in an interval of length δ is $\propto \delta$ as $\delta \downarrow 0$,
- the occurrence of events is independent between intervals.

There are many, many applications: modelling times of child births, industrial accidents, imperfections in cotton, alpha-particle decay, extreme weather incidents, etc. Figure 2.1 shows an actual dataset which has been modelled by a Poisson process: the cumulative totals of cyclones—storms with wind speeds exceeding 88km/h—in the Bay of Bengal, recorded over a 100-year period. Additionally, the Poisson process is highly significant as a building block for all kinds of other random processes, as we will see later in the course—eg, that it is fundamental for *queueing theory*.

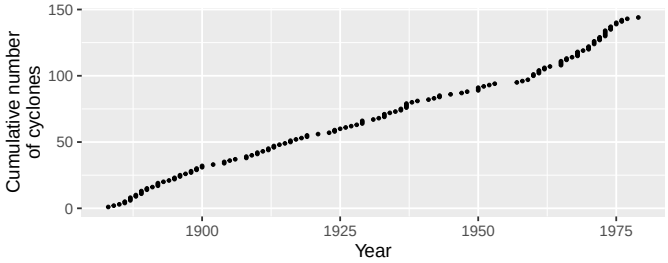

Figure 2.1. Plot of cumulative numbers of cyclones in Bay of Bengal as time varies from 1880 until 1979

2.1.1 Definition and Main Properties

Definition 2.1.1 (Poisson Counting Process)**.** A continuous-time Markov chain taking values in $\mathbb{N}$ is a *Poisson (counting) process* of rate λ, abbreviated $\mathsf{PP}(\lambda)$, if it satisfies the Standing Assumptions[1] and its only non-zero transition rates are

$$q_{i,i+1} = \lambda \quad \text{for all} \quad i \in \mathbb{N}.$$

Its instantaneous transition-rates matrix is given by

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

The following theorem gives an alternative interpretation of a PPs, which is very useful for simulating its paths. It is really the jump-chain construction of §1.3.1.

Theorem 2.1.2. *Let* $T_1, T_2, \ldots \sim^{\text{iid}} \mathsf{Exp}(\lambda)$. *Define* $N = (N_t)_{t \geq 0}$ *by*

$$N_t := \begin{cases} 0 & \text{if} \quad t < T_1, \\ n & \text{if} \quad T_1 + \ldots + T_n \leq t < T_1 + \ldots + T_{n+1}. \end{cases}$$

[1] recall that the Standing Assumptions are the following:

I it has countable state-space and is time-homogeneous;

II transition probabilities are continuous at 0—ie,

$$p_t(i,j) \to \mathbf{1}\{i = j\} \quad \text{as} \quad t \to \infty \quad \text{for all} \quad i, j \in I;$$

III $q_i < \infty$ for all i;

IV $q_i = \sum_{j \neq i} q_{i,j}$ for all i.

Then, $N \sim \mathsf{PP}(\lambda)$. Moreover, N has independent and stationary increments: $(N_{s+t} - N_s)_{t\geq 0} \sim \mathsf{PP}(\lambda)$ and is independent of $(N_u)_{u\leq s}$ for all $s \geq 0$.

Proof. We prove the independence of increments first, as this immediately implies that N is a Markov process—in fact, not only does it say that N_{s+t} given N_s is independent of the history $(N_u)_{u\leq s}$, but also that $N_{s+t} - N_s$ is independent of N_s.

To this end, let $M_t \coloneqq N_{s+t} - N_s$ for $t \geq 0$. Let S and T denote the jump and holding times for N, respectively; define S' and T' similarly. Let $i \geq 0$ and condition on $\{N_s = i\}$. Then, the following relations hold:

$$\begin{aligned} S'_k &= S_{i+k} - s \quad \text{for all} \quad k \geq 1; \\ T'_1 &= T_{i+1} - (s - S_i) \quad \text{and} \quad T'_k = T_{i+k} \quad \text{for all} \quad k \geq 2. \end{aligned}$$

Now, by the memoryless property and the fact that

$$\{N_s = i\} = \{S_i \leq s\} \cap \{S_{i+1} > s\} = \{S_i \leq s\} \cap \{T_{i+1} > s - S_i\},$$

the law $(T'_1 \mid X_s = i) \sim \mathsf{Exp}(\lambda)$. Moreover, the times $(T'_k)_{k\geq 2}$ are independent of $(T'_k)_{k\leq i}$, and hence of $(N_u)_{u\leq s}$. This proves the independence of increments.

Stationarity of increments also follows from the above argument, along with the fact that the times $(T'_k)_{k\geq 2}$ are $\mathsf{Exp}(\lambda)$-s, by definition. So, $(M_t)_{t\geq 0} \sim (N_t)_{t\geq 0}$.

We now prove that N is indeed a $\mathsf{PP}(\lambda)$. We must show that

$$p_{i,j}(t + \delta) = p_{i,j}(t) + \delta\lambda \mathbf{1}\{j = i + 1\} + o(\delta) \quad \text{as} \quad \delta \to 0. \tag{$\star$}$$

Using stationarity of increments and the cdf of $T_1, T_2 \sim^{\mathsf{iid}} \mathsf{Exp}(\lambda)$, the following hold:

$$\begin{aligned} \mathbb{P}\{N_{t+\delta} = N_t \mid N_t = i\} &= \mathbb{P}\{N_\delta = 0\} = \mathbb{P}\{T_1 > \delta\} = e^{-\lambda\delta} = 1 - \lambda\delta + o(\delta); \\ \mathbb{P}\{N_{t+\delta} \geq N_t + 1 \mid N_t = i\} &= \mathbb{P}\{N_\delta \geq 1\} = \mathbb{P}\{T_1 \leq \delta\} = 1 - e^{-\lambda\delta} = \lambda\delta + o(\delta); \\ \mathbb{P}\{N_{t+\delta} \geq N_t + 2 \mid N_t = i\} &= \mathbb{P}\{T_1 + T_2 \leq \delta\} \\ &\leq \mathbb{P}\{T_1 \leq \delta\}\mathbb{P}\{T_2 \leq \delta\} = (1 - e^{-\lambda\delta})^2 = o(\delta). \end{aligned}$$

This verifies ($\star$), since $\{N_{t+\delta} - N_t = 1\} = \{N_{t+\delta} - N_t \geq 1\} \setminus \{N_{t+\delta} - N_t \geq 2\}$. □

We now show that $N_t \sim \mathrm{Pois}(\lambda t)$, which is what gives the name "Poisson". This gives us an opportunity to practice using the KDEs, and solving the resulting DE.

Proposition 2.1.3. *The transition probabilities $(P(t))_{t\geq 0}$ of $(N_t)_{t\geq 0} \sim \mathsf{PP}(\lambda)$ are*

$$p_{i,j}(t) = (\lambda t)^{j-i} e^{-\lambda t}/(j - i)! \cdot \mathbf{1}\{j \geq i\} \quad \text{for all} \quad i, j \in \mathbb{N} \quad \text{and} \quad t \geq 0.$$

In particular, $N_t - N_0 \sim \mathrm{Pois}(\lambda t)$ for all $t \geq 0$.

Proof. The lower triangle of the matrix $P(t)$ vanishes because $i \to j$ if and only if $j > i$. Thus, $p_{i,j}(t) = 0$ whenever $j < i$. On the other hand, the transition rates are translation invariant, so $p_{i,j}(t)$ depends on (i, j) through $j - i$ only. Hence,

$$P(t) = \begin{pmatrix} \Phi_0(t) & \Phi_1(t) & \Phi_2(t) & \cdots \\ 0 & \Phi_0(t) & \Phi_1(t) & \cdots \\ 0 & 0 & \Phi_0(t) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad \text{where} \quad \Phi_k(t) \coloneqq p_{0,k}(t).$$

The KFDE—ie, $P'(t) = QP(t)$ with $P_0 = I$—becomes

$$\begin{cases} \Phi_0'(t) = -\lambda\Phi_0(t) & \text{with} \quad \Phi_0(0) = 1, \\ \Phi_k'(t) = -\lambda\Phi_k(t) + \lambda\Phi_{k-1}(t) & \text{with} \quad \Phi_k(0) = 0 \quad \text{for} \quad k \geq 1. \end{cases}$$

We can solve this inductively. First, clearly, $\Phi_0(t) = e^{-\lambda t}$. Next, using the method of integrating factors, we write the equation for $k \geq 1$ as

$$\tfrac{d}{dt}\big(e^{\lambda t}\Phi_k(t)\big) = e^{\lambda t}\lambda\Phi_{k-1}(t) \quad \text{with} \quad \Phi_k(0) = 0.$$

Integrating both sides, we obtain the recurrence relation

$$\Phi_k(t) = \lambda e^{-\lambda t}\textstyle\int_0^t e^{\lambda s}\Phi_{k-1}(s)ds \quad \text{for} \quad k \geq 1 \quad \text{with} \quad \Phi_0(t) = e^{-\lambda t}.$$

Eg,

$$\Phi_1(t) = \lambda e^{-\lambda t}\textstyle\int_0^t e^{\lambda s}\Phi_0(s)ds = \lambda e^{-\lambda t}\int_0^t 1ds = \lambda t e^{-\lambda t}.$$

The claimed formula can then be verified by induction on $k \geq 0$. □

Exercise 2.1.4. *Verify that the KBDE is satisfied by this solution.*

The above proposition says that the number of jumps made by a $\mathsf{PP}(\lambda)$ in time t is $\mathrm{Pois}(\lambda t)$. The 'inverse problem' is to determine at what time the n-th jump is made. The next proposition shows that this time is Γ-distributed.

Proposition 2.1.5. *Let $T_1, T_2, \ldots \sim^{\mathsf{iid}} \mathsf{Exp}(\lambda)$. Then, for all $n \geq 1$, the sum $T_1 + \ldots + T_n \sim \Gamma(\lambda, n)$—ie, it is a non-negative, continuous random variable with pdf*

$$t \mapsto \lambda(\lambda t)^{n-1}e^{-\lambda t}/(n-1)!.$$

Proof (Exercise). This can be proved by induction using the convolution formula for the pdf of the sum of independent random variables. Alternatively, it can be proved by computing the mgf of the independent sum $T_1 + \ldots + T_n$ and of the given pdf. □

The independence of increments and Poisson distribution just proved will be of fundamental importance. We restate them in their own theorem for ease of reference.

Theorem 2.1.6 (Independent, Poisson-Distributed Increments of a Poisson Process)**.** *If $N \sim \mathsf{PP}(\lambda)$, then the increment $N_{t+s} - N_t \sim \mathrm{Pois}(\lambda s)$ and is independent of $(N_u)_{u \le t}$.*

2.1.2 Further Properties

The Markov property says that $((X_{s+t})_{t \ge 0} \mid X_s = i) \sim ((X_t)_{t \ge 0} \mid X_0 = i)$ for deterministic $s, t \ge 0$, in the time-homogeneous set-up. The *strong* Markov property extended this to allow t to be a *stopping time* under certain regularity conditions.

Recall from Definition 1.2.3 that a random time T is a *stopping time* for X if $\{T \le t\}$ depends only on $(X_s)_{s \le t}$ for all $t \ge 0$. Informally, a random time T is a stopping time if its occurrence (or not) by time t can be determined only by information available up to time t. Typically, it will be the "time that an event happens". Eg, it may be the "time of the first jump" or the "time that a process first crosses a threshold". The "time of the last jump before s" is not a stopping time, though: this cannot be determined by the information up until time t if $t < s$; it is not known whether another jump happens in the interval (t, s).

The Poisson process is a Markov chain which satisfies the appropriate regularity conditions for the strong Markov property to hold.

Theorem 2.1.7 (Strong Poisson Increments)**.** *Let $N \sim \mathsf{PP}(\lambda)$. Let T be a stopping time for N. Then, conditional on $T < \infty$, the process $(N_{T+t} - N_T)_{t \ge 0} \sim \mathsf{PP}(\lambda)$ and is independent of $(N_s)_{s \le T}$.*

We now give an example of data which can be modelled via a Poisson process.

Example 2.1.8 (Coal-Mining Disasters in the UK)**.** Consider the specific decade 1933–1942 in the plot of coal-mining disasters illustrated by Figure 2.2, taken from Jarrett [Jar79]. Do disasters cluster in time?

Reference to the generating dataset shows that there are 14 disasters in the decade 1933–1942 (outlined in red), which suggests that we estimate the rate λ for that decade by $\lambda := 14/10 = 1.4$. Let us define, somewhat arbitrarily, a *clustered incident* as a disaster which is followed by another disaster less than $\delta := 0.25$ years later. Then, we expect about $\lambda t(1 - e^{-\lambda\delta}) \approx 4.1$ clustered incidents in this period of length $t = 10$ years, on the basis of Poisson variation; see below for justification of this expression. Data inspection shows that there are 5 clustered incidents, which is near enough to the expected value to mean that it is not strong evidence of any clustering.

Considering the entire range, from 1851 to 1962, it is clear that it is unreasonable to consider the intensity to be constant in time. A more sophisticated analysis, not

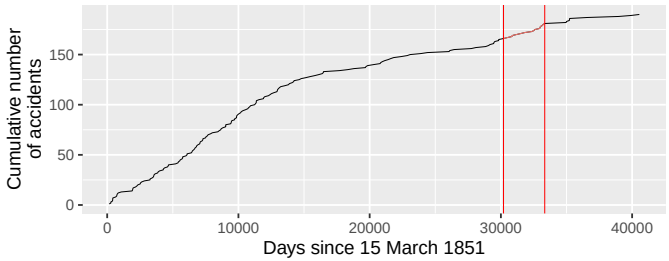

Figure 2.2. Cumulative numbers each year of coal-mining disasters in the UK over the period 15 March 1851 to 22 March 1962, taken from [Jar79]. The portion bracketed by red vertical lines highlights the decade 1933–42

taken in this course, is required.

The calculation of the expectation claimed above follows from Theorem 2.1.7. If $N \sim \mathsf{PP}(\lambda)$ describing the incidents and S_n is the time of the n-th incident, then $(N_{S_n+s} - N_{S_n})_{s\geq 0} \sim \mathsf{PP}(\lambda)$. Then, the probability that the n-th incident is clustered is

$$\mathbb{P}\{n\text{-th incident is clustered}\} = \mathbb{P}\{T_1 < \delta\} = \mathbb{P}\{\mathsf{Exp}(\lambda) < \delta\} = 1 - e^{-\lambda\delta}.$$

Thus, the expected number $\mathbb{E}(C_t)$ of clustered incidents before time t is given by

$$\textstyle\sum_n \mathbb{P}\{S_n \leq t,\ S_{n+1} - S_n \leq \delta\} = \sum_n \mathbb{P}\{S_n \leq t\}(1 - e^{-\lambda\delta}).$$

Now, N_t is the number of incidents which happen by time t. So,

$$N_t = \textstyle\sum_n \mathbf{1}\{S_n \leq t\} \quad \text{and so} \quad \mathbb{E}(N_t) = \sum_n \mathbb{P}\{S_n \leq t\}.$$

But, $N_t \sim \mathrm{Pois}(\lambda t)$, so $\mathbb{E}(N_t) = \lambda t$. Hence,

$$\mathbb{E}(C_t) = \lambda t(1 - e^{-\lambda\delta}). \qquad \triangle$$

The following theorem extends the previous increment-theorem to disjoint sets that are deterministic, but not necessarily intervals.

Theorem 2.1.9 (Poisson Counts). *Let $N \sim \mathsf{PP}(\lambda)$. Let A and B be disjoint (measurable) subsets of $\mathbb{R}_+$, of Lebesgue measure ('size') a and b respectively. Then, the number $N(A)$ of incidents counted by the process N in A has law $\mathrm{Pois}(\lambda a)$ and, moreover, $N(A)$ and $N(B)$ are independent.*

Proof. We first show that $N(A) \sim \mathrm{Pois}(\lambda a)$. We prove this for sets of the form

$$A = \dot{\cup}_{k=1}^{n}(s_k, t_k) \quad \text{where} \quad s_1 < t_1 < s_2 < t_2 < \ldots < s_n < t_n.$$

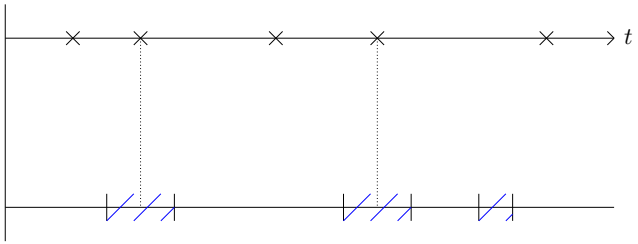

Figure 2.3. Illustration of Poisson counts

More general sets require measure theory that we do not cover here. We have

$$N(A) = \sum_{k=1}^{n} N(s_k, t_k) = \sum_{k=1}^{n} (N_{t_k} - N_{s_k}).$$

Now, Theorem 2.1.6 says that $N_{t_k} - N_{s_k} \sim \text{Pois}\big(\lambda(t_k - s_k)\big)$ independently of k. Sums of independent Poisson random variables are Poisson, so $N(A)$ is also Poisson-distributed. Its mean is $\sum_{k=1}^{n} \lambda(t_k - s_k) = \lambda a$. Hence, $N(A) \sim \text{Pois}(\lambda a)$.

Independence also follows from this proof, for sets which are unions of intervals. Indeed, $N_{t_k} - N_{s_k}$ is independent of $(N_u)_{u \le s_k}$, by Theorem 2.1.6, and of $(N_u)_{u > t_k}$, trivially. So, the number of arrivals in disjoint intervals is independent. □

Example 2.1.10. Consider the hatched regions in Figure 2.3, making up A. Cross-marks $\times$ indicate individual Poisson counts. Here, $N(A) = 2$. △

The following two theorems describe *superposition* of ('adding together') Poisson processes and *thinning* of ('removing arrivals from') Poisson process.

Theorem 2.1.11 (Superposition of Poisson Processes)**.** *If $N \sim \mathsf{PP}(\lambda)$ and $M \sim \mathsf{PP}(\mu)$ independently, then $N + M := (N_t + M_t)_{t \ge 0} \sim \mathsf{PP}(\lambda + \mu)$.*

Proof. The main part of the work lies in showing the Markov property. Let A and B be arbitrary events depending only on $(N_u)_{u \le s}$ and $(M_u)_{u \le s}$, respectively. The Markov property follows if $(N_{s+t} + M_{s+t}) - (N_s + M_s)$ is independent of A and B.

By independence of N and M first, then independence of increments second,

$$\begin{aligned}&\mathbb{P}\{N_{s+t} - N_s = i,\ M_{s+t} - M_s = j \mid A,\ B\} \\ &\quad = \mathbb{P}\{N_{s+t} - N_s = i \mid A\}\mathbb{P}\{M_{s+t} - M_s = j \mid B\} = \mathbb{P}\{N_t = i\}\mathbb{P}\{M_t = j\}.\end{aligned}$$

Now, summing over (i, j) such that $i + j = k$, we obtain

$$\mathbb{P}\{(N_{t+s} + M_{t+s}) - (N_t + M_t) = k \mid A \cap B\} = \sum_{(i,j):i+j=k} \mathbb{P}\{N_t = i\}\mathbb{P}\{M_t = j\}.$$

This does not depend on (A, B), which establishes the Markov property—in fact, it establishes stationary and independent increments.

We now use the infinitesimal definition of a PP to to determine the rates:

$$\begin{aligned}
&\mathbb{P}\{(N_{t+\delta} + M_{t+\delta}) - (N_t + M_t) = 0\} \\
&\quad = \mathbb{P}\{N_{t+\delta} - N_t = 0,\ M_{t+\delta} - M_t = 0\} = \mathbb{P}\{N_\delta = 0\}\mathbb{P}\{M_\delta = 0\} \\
&\quad = \big(1 - \lambda\delta + o(\delta)\big)\big(1 - \mu\delta + o(\delta)\big) = 1 - (\lambda + \mu)\delta + o(\delta); \\
&\mathbb{P}\{(N_{t+\delta} + M_{t+\delta}) - (N_t + M_t) = 1\} \\
&\quad = \mathbb{P}\{N_{t+\delta} - N_t = 1,\ M_{t+\delta} - M_t = 0\} + \mathbb{P}\{N_{t+\delta} - N_t = 0,\ M_{t+\delta} - M_t = 1\} \\
&\quad = \big(\lambda\delta + o(\delta)\big)\big(1 - \mu\delta + o(\delta)\big) + \big(\mu\delta + o(\delta)\big)\big(1 - \lambda\delta + o(\delta)\big) = (\lambda + \mu)\delta + o(\delta).
\end{aligned}$$

Hence, the rates are $q_{i,j} = \mathbf{1}\{j = i + 1\}$, as required. □

Theorem 2.1.12 (Poisson Thinning). *If $N \sim \mathsf{PP}(\lambda)$ and it is thinned by independent removal of events with probability $1-p$, for some $p \in (0, 1)$, then the thinned process $M \sim \mathsf{PP}(p\lambda)$, counting the remaining incidents. Also, $N - M \sim \mathsf{PP}((1 - p)\lambda)$. Moreover, M and $N - M$ are independent.*

Proof. Let T be the holding times of N. Let $B_1, B_2, \ldots \sim^{\mathsf{iid}} \mathsf{Bern}(p)$. Let $K_1, K_2, \ldots$ be the indices of successive 1-s in the sequence of 0-s and 1-s generated by the B-s:

$$K_0 \coloneqq 0 \quad \text{and} \quad K_m \coloneqq \min\{m > K_{n-1} \mid M_m = 1\} \quad \text{for} \quad m \geq 1.$$

The $K_1, K_2, \ldots$ are themselves random variables. Then, the holding times T' of M are

$$T'_n \coloneqq \textstyle\sum_{m=K_{n-1}+1}^{K_n} T_m.$$

The theorem then follows from the jump-chain construction in Theorem 2.1.2 if we can show that the new holding times T' are in fact independent $\mathsf{Exp}(p\lambda)$ random variables.

Their mutual independence follows easily: T'_n and $T'_{n'}$ depend on disjoint sets of T_m-s if $n \neq n'$; hence, they are independent.

We now show that $T'_n \sim \mathsf{Exp}(p\lambda)$. T'_n is a sum over $Z_n \coloneqq K_n - K_{n-1}$ random variables T_m $(m \in \{K_{n-1} + 1, \ldots, K_n\})$. Now, Z_n is defined in terms of M, so is independent of S. Moreover, $Z_n \sim \mathsf{Geom}_1(p)$:

$$\mathbb{P}\{Z_n = k\} = (1 - p)^{k-1}p \quad \text{for} \quad k \in \{1, 2, \ldots\}.$$

It is left as an exercise to deduce that $T'_n \sim \mathsf{Exp}(p\lambda)$. The easiest way is via mgfs:

$$\mathbb{E}\big(e^{xT'_n}\big) = \mathbb{E}\big(\mathbb{E}\big(e^{x\sum_{K_{n-1}<m\leq K_n} T_m} \mid Z_n\big)\big) = \mathbb{E}\big(\mathbb{E}\big(e^{xT_1}\big)^{Z_n}\big) \quad \text{for all} \quad x \in \mathbb{R}.$$

One can also directly compute the distribution using Proposition 2.1.5 or use the memoryless property for the T-s to deduce that T'_n has the memoryless property, which gives the Exponential distribution, then find the mean.

Now that we have shown $M \sim \mathsf{PP}(p\lambda)$, it is immediate, analogously, that $N-M \sim \mathsf{PP}((1-p)\lambda)$ since $N-M$ is a thinning on N with probability $1-p$.

It remains to prove that M and $N-M$ are independent. Both processes are right-continuous and increasing, so it suffices to check that the finite-dimensional marginals are independent: ie, that

$$\begin{aligned}&\mathbb{P}\{M_{t_1}=m_1,...,M_{t_k}=m_k,N_{t_1}-M_{t_1}=n_1,...,N_{t_k}-M_{t_k}=n_k\}\\&\quad=\mathbb{P}\{M_{t_1}=m_1,...,M_{t_k}=m_k\}\mathbb{P}\{N_{t_1}-M_{t_1}=n_1,...,N_{t_k}-M_{t_k}=n_k\}\\&\qquad\text{whenever}\quad t_1\le...\le t_k,\ m_1\le...\le m_k\ \text{ and } n_1\le...\le n_k.\end{aligned}$$

We only show this for $k=1$, but the general case follows similarly. First, observe that

$$\{M_t=m,N_t-M_t=m\}=\{N_t=n+m,M_t=m\}.$$

This event happens if and only if there are $n+m$ arrivals in the full process and precisely m of them are retained—ie, m of the $B_1,...,B_{n+m}\in\{0,1\}$ are equal to 1, or, equivalently, $B_1+...+B_{n+m}=m$. Now, $N_t\sim\text{Pois}(\lambda t)$ and $B_1+...+B_{n+m}\sim\mathsf{Bin}(n+m,p)$. Hence, using the pdf of the Poisson and Binomial distributions and the independence of the B_ℓ-s from N_t, we have

$$\begin{aligned}\mathbb{P}\{M_t=m,N_t-M_t=m\}&=\mathbb{P}\{N_t=n+m,M_t=m\}\\&=e^{-\lambda t}\frac{(\lambda t)^{n+m}}{(n+m)!}\cdot\binom{n+m}{m}p^m(1-p)^n\\&=\left(e^{-p\lambda t}(p\lambda t)^m/m!\right)\cdot\left(e^{-(1-p)\lambda t}\big((1-p)\lambda t\big)^n/n!\right)\\&=\mathbb{P}\{M_t=m\}\mathbb{P}\{N_t-M_t=m\},\end{aligned}$$

since $M_t\sim\text{Pois}(p\lambda t)$ and $N_t-M_t\sim\text{Pois}((1-p)\lambda t)$. This shows that M_t and N_t-M_t are independent for all t. □

Exercise 2.1.13. *Show that the thinned process M from the previous theorem is a $\mathsf{PP}(\lambda p)$ via the infinitesimal definition of a Poisson process.*

Example 2.1.14. Customers arrive at a refreshment stall according to a Poisson process of rate 2 per minute. Each individual customer, independently of everything else, wants a cup of tea with probability $\frac{1}{4}$, wanting a cup of coffee otherwise. Thus, the system of arrivals wanting tea is obtained by independent thinning of a Poisson process of rate $\lambda=2$ using retention probability $p=\frac{1}{4}$.

Theorem 2.1.12 shows that these arrivals form a Poisson process of rate $p\lambda=\frac{1}{2}$ per minute. Similarly, the system of arrivals wanting coffee forms a Poisson process of rate $\frac{3}{2}$. Further, the tea- and coffee-arrival processes are actually independent. △

2.1.3 A Brief Return to Instantaneous Transition Rates

Poisson thinning and superposition was briefly hinted at as far back as §1.3, albeit in disguise. There, we discussed ways of sampling continuous-time Markov chains.

- The most basic approach was to sample $T \sim \mathsf{Exp}(q_i)$ and $J \sim (q_{i,j})_{j \in I \setminus \{i\}}$.
- The more refined approach was to sample $T_j \sim \mathsf{Exp}(q_{i,j})$ for each $j \in I \setminus \{i\}$, then set $T \coloneqq \min_j T_j$ and $J \coloneqq \arg\min_j T_j$, so $T = T_J$.
- In either case, $X_t \coloneqq i$ for $t < T$ and $X_T \coloneqq J$.

Competition of exponentials (Lemma 1.3.10) established the equivalence of these.

We now frame this in the manner of (thinned/superimposed) Poisson processes. A discrete-time Markov chain can be seen as following a set of (random) instructions:

- every state has an infinite stack of (random) cards with instructions on them;
- upon arriving at a state, the next card is looked at to see where to jump next.

A continuous-time chain could be viewed similarly:

- when at state i, it waits $\mathsf{Exp}(q_i)$ to turn over the next instruction;
- alternatively, instructions arrive as $\mathsf{PP}(q_i)$ and are read upon arrival.

This alternative is better than the first, but there is an even better way:

- cards with instruction "jump to j" arrive as independent $\mathsf{PP}(q_{i,j})$-s for each j.

This means that the instruction "jump to j" is arriving to i at rate $q_{i,j}$. Hence, the continuous-time Markov chain at i "attempts to jump (from i) to j at rate $q_{i,j}$".

All this fundamentally boils down to the fact that if $T_1, T_2, \ldots \sim^{\mathsf{iid}} \mathsf{Exp}(\lambda)$ and $N \sim \mathsf{Geom}_1(p)$ independently, then $\sum_{n=1}^{N} T_n \sim \mathsf{Exp}(p\lambda)$.

2.2 Birth Processes

Birth processes are processes on $\mathbb{N} \cup \{\infty\}$ where only positive jumps are permitted—ie, they are *counting processes*. These simplest of these is the Poisson process, just discussed, where the birth rate is constant. Next, we study *pure-birth processes*, where the jump rate is allowed to vary from state to state.

2.2.1 Pure-Birth Processes

We start with the precise definition of a pure-birth process.

Definition 2.2.1 (Pure-Birth Process)**.** A continuous-time Markov chain taking values in $\mathbb{N} \cup \{\infty\}$ is a *pure-birth process* with rates $\lambda = (\lambda_i)_{i \geq 0}$, abbreviated $\mathsf{PBP}(\lambda)$, if its only non-zero transition rates are

$$q_{i,i+1} = \lambda_i \quad \text{for all} \quad i \in \mathbb{N}.$$

Its instantaneous transition-rates matrix is given by

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots & \cdots \\ 0 & -\lambda_1 & \lambda_1 & 0 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots \end{pmatrix}. \qquad \triangle$$

This definition is identical to that for Poisson processes, except that the jump-rate is allowed to depend on the location and we allow $\infty \in I$.

Remark. We take the state space to be $I = \mathbb{N} \cup \{\infty\}$. However, it is sometimes natural to exclude 0 and start from $X_0 = 1$—ie, one individual alive. In particular, when $\lambda_0 = 0$, starting from $X_0 > 0$ is necessary to get a non-trivial process—ie, one which does not remain at 0 forever. Eg, a population cannot grow if it has 0 members.

We include $\infty \in I$ as it is possible for the chain to reach ∞ in a finite time. Eg, if $q_i = 2^i$, then the expected time to transition $i \to i+1$ is 2^{-i}. Thus, the expected time to transition $0 \to n$ is $\sum_{i=0}^{n-1} 2^{-i} = 2 - 2^{-n+1} < 2$. So, the chain reaches ∞ in 2 units of time, in expectation. This is a phenomenon called *explosion*. We discuss it more later, but was already mentioned when discussing the strong Markov property. $\triangle$

Exercise 2.2.2. *Cosmic rays arrive at atmosphere top at height y km and with vertical velocity v km/s. They collide with atmosphere molecules occasionally as they fall and give birth to other particles; this happens at rate α Hz (per second). These, in turn, collide and give birth. All particles travel with constant vertical velocity v km/s.*

Assume that X_0 particles arrive simultaneously at atmosphere top at time 0 and set the number of particles at t seconds to be X_t. Show that $X = (X_t)_{t \geq 0}$ forms a pure-birth process and find its rates.

Pure-birth processes can be constructed via exponentials, as for Poisson processes.

Theorem 2.2.3. *Let $T_n \sim \mathsf{Exp}(\lambda_{n-1})$ independently for all $n \geq 1$. Let $m \in \mathbb{N}$. Set*

$$X_t \coloneqq \begin{cases} m & \text{if} \quad t < T_{m+1}, \\ n & \text{if} \quad T_{m+1} + \ldots + T_n \leq t < T_{m+1} + \ldots + T_{n+1}. \end{cases}$$

Figure 2.4. Cosmic rays' arriving at atmosphere top, colliding occasionally with molecules to give rise to further cosmic rays

Then, X is a pure-birth process with rates $\lambda = (\lambda_{n-1})_{n\geq 1}$ started from $X_0 = m$.

Proof. The proof of this theorem is very similar to that for the jump-chain construction of a Poisson process, Theorem 2.1.2. We just need to check the rates. The rates which were 0 before are still 0: the process only jumps $+1$, so cannot go down, and the chance of having two jumps in time δ is $o(\delta)$. For $n \geq 1$,

$$\mathbb{P}\{X_{t+\delta} - X_t \geq 1 \mid X_t = n\} = \mathbb{P}\{X_\delta \geq 1 \mid X_0 = n\} = \mathbb{P}\{T_n \leq \delta\} = \lambda\delta + o(\delta). \quad \square$$

Corollary 2.2.4. *The holding times $(T_n)_{n\geq 1}$ of a pure-birth process with rates $\lambda = (\lambda_{n-1})_{n\geq 1}$ are independent and have law $T_n \sim \mathsf{Exp}(\lambda_{n-1})$ for all $n \geq 1$.*

Several of the properties of Poisson processes also hold for general pure-birth processes. The following theorem generalises Theorem 2.1.6 for Poisson processes and it can be proved in exactly the same way.

Theorem 2.2.5 (Strong/Markov Property). *Let $X \sim \mathsf{PBP}(\lambda)$. Then, conditional on $X_t = i$, the process $(X_{t+s})_{s\geq 0} \sim \mathsf{PBP}(\lambda)$ starting from i and independent of $(X_s)_{s\leq t}$. Moreover, the same holds if the deterministic time t is replaced by a stopping time T.*

Exercise 2.2.6 (Hard!). *Let X be a pure-birth process with rates $\lambda = (\lambda_i)_{i\geq 0}$ such that $\lambda_i \neq \lambda_j$ whenever $n \neq m$. Show that*

$$\mathbb{P}_0\{X_t \leq j\} = \sum_{k\in\{0,\dots,j\}} e^{-\lambda_k t} \prod_{\ell\in\{0,\dots,j\}\setminus\{k\}} \frac{\lambda_\ell}{\lambda_\ell - \lambda_k}.$$

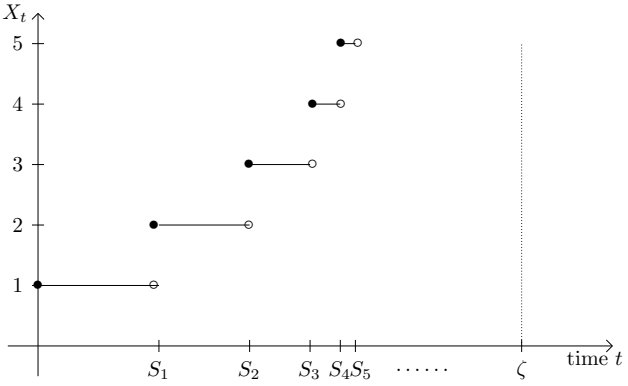

Figure 2.5. Illustration of explosion for a pure-birth process

2.2.2 Explosion

An interesting phenomenon in continuous-time Markov chains is the possibility for the process to *explode*. We describe this in the set-up of pure-birth processes.

Definition 2.2.7 (Explosion)**.** Let X be a pure-birth process with rates $(\lambda_{n-1})_{n\geq 1}$ and holding times $(T_n)_{n\geq 1}$. Let $S_n := \sum_{m=1}^{n} T_m$, the time of the n-th jump. Define

$$\zeta := \lim_{n\uparrow\infty} S_n = \sum_{n\geq 1} T_n.$$

If $\zeta < \infty$, then we say that the pure-birth process has *exploded*; ζ is the random *explosion time*—ie, the time at which the chain reaches ∞. △

Figure 2.5 illustrates explosion. We are interested in the probability of explosion—ie, $\mathbb{P}\{\zeta < \infty\}$. It turns out that $\{\zeta < \infty\}$ is a 0–1 event.

Theorem 2.2.8 (Explosion Dichotomy; [Nor97, Theorem 2.3.2])**.** *Let $(T_n)_{n\geq 1}$ be a sequence if independent random variables with $T_n \sim \mathsf{Exp}(\lambda_{n-1})$ where $\lambda_{n-1} \in (0,\infty)$ for all n. Recall that $\zeta = \sum_{n\geq 1} T_n$. Then, the following explosion dichotomy holds:*

$$\sum_{n\geq 1} 1/\lambda_{n-1} < \infty \implies \mathbb{P}\{\zeta < \infty\} = 1;$$
$$\sum_{n\geq 1} 1/\lambda_{n-1} = \infty \implies \mathbb{P}\{\zeta = \infty\} = 1.$$

We now rephrase this in terms of explosion for pure-birth processes.

Corollary 2.2.9. *Let $X \sim \mathsf{PBP}\big((\lambda_{n-1})_{n\geq 1}\big)$. Then, with probability 1, the process*

explodes if $\sum_{n\geq 1} 1/\lambda_n < \infty$*; otherwise, with probability* 1*, it does not explode.*

Example 2.2.10. If $\lambda_{n-1} = n\lambda$ for some $\lambda > 0$—this is the *simple birth process* and is discussed in the next section—then, with probability 1, the process does not explode.

If $\lambda_{n-1} = n^2\lambda$ for some $\lambda > 0$, then, with probability 1, the process explodes. △

Remark. The possibility of explosion is the reason why we include ∞ in the state space. It was not included for Poisson processes, but there $\lambda_{n-1} = \lambda$ for all n, so $\sum_{n\geq 1} 1/\lambda = \infty$, and hence there is no explosion. It is possible to avoid including ∞, but then the transition-probability matrix $P(t)$ becomes *sub-stochastic*: its rows sum to at most 1, the difference to 1 being the probability of "escaping to ∞". △

2.2.3 Simple Birth Processes

We now focus on linear birth processes: those whose transition rates depend linearly on the current state. The simplest of them is the following.

Definition 2.2.11 (Simple Birth Process, aka Yule Process)**.** A pure-birth process with rates $\lambda_n = n\lambda$ for some $\lambda > 0$ is called a *simple birth process*, abbreviated $\mathsf{SBP}(\lambda)$, or a *Yule process*, with parameter λ, if its only non-zero transition rates are

$$q_{i,i+1} = \lambda i \quad \text{for all} \quad i \in I.$$

Its instantaneous transition-rates matrix Q is given by

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & \cdots \\ 0 & -\lambda & \lambda & 0 & \cdots & \cdots \\ 0 & 0 & -2\lambda & 2\lambda & \cdots & \cdots \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots \end{pmatrix}. \qquad \triangle$$

We showed previously that the minimum of Exponential random variables is Exponentially distributed. This allows us to construct a SBP in an alternative manner.

Lemma 2.2.12 (Alternative Construction of SBP)**.** *Consider bacteria in a colony. Assign to each bacterium an independent* $\mathsf{Exp}(\lambda)$ *splitting timer:*

- *when this timer 'rings', the bacterium splits into two new bacteria;*
- *each is assigned a new, independent* $\mathsf{Exp}(\lambda)$ *timer.*

Let X_t *count the number of bacteria in the colony at time* t*. Then,* $X \sim \mathsf{SBP}(\lambda)$*.*

Proof. We can take $\lambda := 1$ for simplicity without loss of generality.

Suppose that a jump happens at time t, and $X_t = n$. Then, there are n bacteria, each with an independent $\mathsf{Exp}(1)$ timer. Some of the timers have already been running for an unknown length of time. However, the memoryless property of the Exponential distribution implies that the time until they ring is an independent $\mathsf{Exp}(1)$. Hence, the time until the first rings is the minimum of these n iid $\mathsf{Exp}(1)$-s, which is distributed as $\mathsf{Exp}(n)$. Moreover, the jump times are independent. So, $X \sim \mathsf{SBP}(1)$.

Analogously, if we suppose that $X_t = n$, and that the previous jump-time was at $s < t$, then we can again apply the memoryless property at time t to deduce that the time until the next jump is the minimum of iid Exponentials. □

A *superposition* result, analogous to that for PPs (Theorem 2.1.11), holds for SBPs.

Theorem 2.2.13 (Superposition of Simple Birth Processes). *If* $Y, Z \sim^{\mathsf{iid}} \mathsf{SBP}(\lambda)$, *then* $X = (X_t := Y_t + Z_t)_{t \geq 0} \sim \mathsf{SBP}(\lambda)$ *with* $X_0 = Y_0 + Z_0$.

The independent-evolution construction gives an intuitive proof.

Proof: Intuitive. We prove this for $Y_0 = 1 = Z_0$; the proof generalises easily (exercise). Lemma 2.2.12 implies that a $\mathsf{SBP}(\lambda)$ can be viewed as independent evolution of bacteria. This way, if we start with two bacteria, then we can evolve their processes—namely, the time at which they split, then the time at which their children split, etc—independently. Each is a $\mathsf{SBP}(\lambda)$ started with one individual. Looking at the processes combined takes us back to a $\mathsf{SBP}(\lambda)$ with two initial bacteria. □

The proof for PPs (Theorem 2.1.11) can also be followed, but it is less informative.

Proof: Technical. The proof of the Markov property follows that for PPs (Theorem 2.1.11). We just need to check the rates. Remember that a birth process is increasing, so $X_t = X_0$ implies $Y_t = Y_0$ and $Z_t = Z_0$. Thus,

$$\begin{aligned}
\mathbb{P}_r\{X_t = X_0\} &= \mathbb{P}\{Y_t = Y_0, Z_t = Z_0 \mid X_0 = r\} \\
&= \textstyle\sum_{k=0}^{r} \mathbb{P}\{Y_t = Y_0, Z_t = Z_0 \mid Y_0 = k, Z_0 = r-k\}\mathbb{P}\{Y_0 = k \mid X_0 = k\} \\
&= \textstyle\sum_{k=0}^{r} \mathbb{P}_k\{Y_t = Y_0\}\mathbb{P}_{r-k}\{Z_t = Z_0\}\mathbb{P}\{Y_0 = k \mid X_0 = r\} \\
&= \textstyle\sum_{k=0}^{r} e^{-\lambda t k} e^{-\lambda t (r-k)} \mathbb{P}\{Y_0 = k \mid X_0 = r\} = e^{-\lambda r t}.
\end{aligned}$$

Hence, $q_{r,r} = \frac{d}{dt} e^{-\lambda r t}\big|_{t=0} = -\lambda r$. But, X makes jumps of unit size: two jumps in time δ has order δ^2 probability. Thus, $q_{r,s} = 0$ if $s \notin \{r, r+1\}$ and $q_{r,r+1} = -q_{r,r} = \lambda r$. □

The next couple of results are not needed later, but rather are examples of how to perform calculations with the SBP. First, we look at the mean and variance.

Lemma 2.2.14 (Mean and Variance of Simple Birth Process). *Let $X \sim \mathsf{SBP}(\lambda)$ for some $\lambda > 0$. Find a differential equation satisfied by $\mu_k(t) \coloneqq \mathbb{E}_1(X_t^k)$ for $t \geq 0$ and $k \in \{1, 2\}$. Solve these and deduce expressions for the mean and variance of X_t:*

$$\mathbb{E}_1(X_t) = e^{\lambda t} \quad \text{and} \quad \mathbb{V}\mathrm{ar}_1(X_t) = e^{2\lambda t} - e^{\lambda t}.$$

Proof. We take $\lambda \coloneqq 1$ without loss of generality: replacing t by $t\lambda$ at the end retrieves the general-λ statement. Also, we abbreviate $p_n(t) \coloneqq p_{1,n}(t)$.

We start with $\mu_1(t) = \mathbb{E}_1(X_t) = \sum_{n\geq 1} np_n(t)$. Using the KFDEs,

$$\begin{aligned}
\mu_1'(t) &= \textstyle\sum_{n\geq 1} np_n'(t) \\
&= \textstyle\sum_{n\geq 1} n\big((n-1)p_{n-1}(t) - np_n(t)\big) \\
&= \textstyle\sum_{n\geq 1} \big((n-1)^2 p_{n-1}(t) - n^2 p_n(t) + (n-1)p_{n-1}(t)\big) \\
&= \textstyle\sum_{n\geq 1} (n-1)p_{n-1}(t) = \sum_{n\geq 1} np_n(t) = \mu_1(t),
\end{aligned}$$

where we wrote $n = (n-1) + 1$ and used an index shift, and $p_{1,0}(t) = 0$, to cancel

$$\textstyle\sum_{n\geq 1} \big((n-1)^2 p_{n-1}(t) - n^2 p_n(t)\big) = 0.$$

Solving this simple differential equation, using $\mu_1(0) = 1$, gives the mean:

$$\mathbb{E}_1(X_t) = \mu_1(t) = e^t.$$

This exponential increase with rate 1 is natural when viewed in the independent-reproductive manner of Lemma 2.2.12. Each member of the population is reproducing at constant rate λ. So, the rate of increase of the population is equal to the current population: $\mu_1'(t) = \lambda\mu(t)$; hence, $\mu_1(t) = e^{\lambda t}$.

We now turn to $\mu_2(t) = \mathbb{E}_1(X_t^2) = \sum_{n\geq 1} n^2 p_n(t)$. The same ideas give

$$\begin{aligned}
\mu_2'(t) &= \textstyle\sum_{n\geq 1} n^2 p_{1,n}'(t) \\
&= \textstyle\sum_{n\geq 1} n^2\big((n-1)p_{n-1}(t) - np_n(t)\big) \\
&= \textstyle\sum_{n\geq 1} \big((n-1)^3 p_{n-1}(t) - n^3 p_n(t) + 2(n-1)^2 p_{n-1}(t) + (n-1)p_{n-1}(t)\big) \\
&= 2\mu_2(t) + \mu_1(t).
\end{aligned}$$

Solving this, after inputting $\mu_1(t) = e^t$, gives

$$\mu_2(t) = 2e^{2t} - e^t.$$

Finally, we combine μ_2 and μ_1 to find the variance:

$$\mathbb{V}\mathrm{ar}_1(X_t) = \mu_2(t) - \mu_1(t)^2 = (2e^{2t} - e^t) - (e^t)^2 = e^{2t} - e^t. \qquad \square$$

We can find the transition probabilities of a simple birth process using the KDEs.

Proposition 2.2.15. *The transition probabilities of* SBP(λ) *are given by*

$$p_{m,n}(t) = \begin{cases} \binom{n-1}{m-1} e^{-\lambda m t}(1-e^{-\lambda t})^{n-m} & \text{if} \quad 0 \le m < n, \\ 1 & \text{if} \quad m = n = 0, \\ 0 & \text{if} \quad \text{otherwise.} \end{cases}$$

Proof: Technical. First, clearly, 0 is an absorbing state and $m \not\to n$ if $m > n$. Thus, the transition-probabilities matrix $P(t)$ can be written as

$$P(t) = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ 0 & p_{1,1}(t) & p_{1,2}(t) & p_{1,3}(t) & \cdots \\ 0 & 0 & p_{2,2}(t) & p_{2,3}(t) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Then, the KFDE lead to the following equations for $p_{m,n}$:

$$\begin{aligned} &\tfrac{d}{dt} p_{m,m}(t) = -\lambda m p_{m,m}(t) && \text{with} \quad p_{m,m}(0) = 1; \\ &\tfrac{d}{dt} p_{m,n}(t) = -\lambda n p_{m,n}(t) + \lambda(n-1) p_{m,n-1}(t) && \text{with} \quad p_{m,n}(0) = 0 \quad \text{for} \quad n > m. \end{aligned}$$

The first line is straight-forward to solve:

$$p_{m,m}(t) = e^{-\lambda m t} \quad \text{using} \quad p_{m,m}(0) = 1.$$

The method of integrating factors gives a recurrence relation for the second line:

$$\tfrac{d}{dt}\big(e^{\lambda n t} p_{m,n}(t)\big) = e^{\lambda n t}\big(\tfrac{d}{dt} p_{m,n}(t) + \lambda n p_{m,n}(t)\big) = \lambda(n-1) e^{\lambda n t} p_{m,n-1}(t),$$

so,

$$e^{\lambda n t} p_{m,n}(t) = \lambda(n-1) \textstyle\int_0^t e^{\lambda n s} p_{m,n-1}(s) ds.$$

We now use induction to establish the claimed form of $p_{m,n}(t)$. Indeed, by hypothesis,

$$\begin{aligned} p_{m,n}(t) &= \lambda(n-1) e^{-\lambda n t} \textstyle\int_0^t e^{\lambda n s} p_{m,n-1}(s) ds \\ &= \lambda(n-1) e^{-\lambda n t} \textstyle\int_0^t e^{\lambda n s} \binom{n-2}{m-1} e^{-\lambda m s}(1-e^{-\lambda s})^{n-1-m} ds \\ &= \lambda(n-1) e^{-\lambda n t} \binom{n-2}{m-1} \textstyle\int_0^t e^{\lambda s}(e^{\lambda s}-1)^{n-1-m} ds \\ &= \lambda(n-1) e^{-\lambda n t} \binom{n-2}{m-1} \frac{(e^{\lambda t}-1)^{n-m}}{\lambda(n-m)} \\ &= \frac{n-1}{n-m} \binom{n-2}{m-1} \cdot e^{-\lambda n t}(e^{\lambda t}-1)^{n-m} \\ &= \binom{n-1}{m-1} e^{-\lambda m t}(1-e^{\lambda t})^{n-m}. \end{aligned}$$

□

The above proof used induction: this requires knowing, or guessing, the final answer already. This is somewhat unsatisfactory. We now give a direct proof of Proposition 2.2.15, based on intuitive probabilistic arguments, rather than technical details.

Proof: Intuitive. First, we analyse case $m = 1$—ie, starting with 1 individual. We will then use the superposition theorem (Theorem 2.2.13) to deduce Proposition 2.2.15. As always, by homogeneity, there is no loss of generality if we assume that $\lambda = 1$.

Let $X \sim \mathsf{SBP}(1)$ with $X_0 = 1$. Let $H_{n-1} := \inf\{t \geq 0 \mid X_t = n\}$. Then,

$$H_{n-1} = T_1 + ... + T_{n-1} \quad \text{where} \quad T_m \sim \mathsf{Exp}(m) \quad \text{independently.}$$

Now, we must determine the law of H_{n-1}. We claim that

$$H_{n-1} \sim \max\{S_1, ..., S_{n-1}\} \quad \text{where} \quad S_1, ..., S_{n-1} \sim^{\mathsf{iid}} \mathsf{Exp}(1).$$

This has an intuitive proof. Consider $n-1$ independent $\mathsf{Exp}(1)$ timers, $S_1, ..., S_{n-1} \sim^{\mathsf{iid}} \mathsf{Exp}(1)$. Then, the time until the *first* timer rings is $\min\{S_1, ..., S_{n-1}\} \sim \mathsf{Exp}(n-1)$. Once this has rung, the time until the *next* ring is an independent $\mathsf{Exp}(n-2)$, by the memoryless property. In general, the time between the k-th and $(k+1)$-th is an independent $\mathsf{Exp}(n-1-k)$. Hence, the time of the *last* ring is an independent sum

$$\max\{S_1, ..., S_{n-1}\} \sim \mathsf{Exp}(n-1) + ... + \mathsf{Exp}(1) \sim T_{n-1} + ... + T_1 = H_{n-1}.$$

Lemma 2.2.18 below extends this result. Using this representation of H_{n-1},

$$\begin{aligned}\mathbb{P}_1\{X_t \geq n\} = \mathbb{P}\{H_{n-1} \leq t\} &= \mathbb{P}\{\max\{T'_1, ..., T'_{n-1}\} \leq t\} \\ &= \mathbb{P}\{\mathsf{Exp}(\lambda) \leq t\}^{n-1} = (1 - e^{-\lambda t})^{n-1}.\end{aligned}$$

We can now find $p_{1,n}(t)$:

$$p_{1,n}(t) = \mathbb{P}_1\{X_t = n\} = \mathbb{P}_1\{X_t \geq n\} - \mathbb{P}_1\{X_t \geq n+1\} = e^{-t}(1 - e^{-t})^{n-1}.$$

In particular, $(X_t \mid X_0 = 1) \sim \mathsf{Geom}_1(e^{-t})$. Another method for showing this, going via mgfs, is given in Example Sheet 3; there, the mean and variance are also found.

The superposition theorem (Theorem 2.2.13) says that

$$X_t \sim \textstyle\sum_{i=1}^m X_t^{(i)} \quad \text{where} \quad X \sim \mathsf{SBP}_m(1) \quad \text{and} \quad X^{(1)}, ..., X^{(m)} \sim^{\mathsf{iid}} \mathsf{SBP}_1(1),$$

where the subscript indicates the initial value, as usual. We have just proved that $X_t^{(i)} \sim \mathsf{Geom}_1(e^{-t})$. It is well known that a sum of independent Geometric random variables with the same parameter follows the *Negative Binomial* distribution, which has pmf as in Proposition 2.2.15. This completes the proof. □

Exercise 2.2.16. *Show that a sum of iid Geometrics has the Negative Binomial distribution. That is, suppose that* $X_1, ..., X_m \sim^{\text{iid}} \mathsf{Geom}_1(p)$*, for some* $p \in (0,1)$*. Show that*

$$\mathbb{P}\{\textstyle\sum_{i=1}^m X_i = n\} = \binom{n-1}{m-1} p^m (1-p)^{n-m}.$$

This exercise is typically proved by induction or by calculating the mgf of the sum and of the given pdf. Either proof requires knowing, or guessing, the desired formula, which is not particularly satisfying. However, this is a classical result regarding independent sums of standard random variables.

The formula can also be derived directly using a convolution, starting with a few base cases, then guessing the (relatively simple) formula.

Corollary 2.2.17 (Memoryless Property of the Simple Birth Process)**.** $X \sim \mathsf{SBP}(\lambda)$ *with* $X_0 = 1$ *satisfies the discrete memoryless property: for all* $m, n \geq 0$ *and all* $t \geq 0$*,*

$$\mathbb{P}_1\{X_t \geq 1 + m + n \mid X_t \geq 1 + m\} = \mathbb{P}_1\{X_t \geq 1 + n\}.$$

Proof. This is an immediate consequence of the fact that $(X_t \mid X_0 = 1)$ is Geometrically distributed. Indeed, if $X \sim \mathsf{Geom}_1(p)$—eg, with $p = e^{-\lambda t}$—then

$$\begin{aligned}
&\mathbb{P}\{X \geq 1 + m + n \mid X \geq 1 + m\} \\
&\quad = \frac{\mathbb{P}\{X \geq 1 + m + n\}}{\mathbb{P}\{X \geq 1 + m\}} = \frac{(1-p)^{m+n}}{(1-p)^m} = (1-p)^n = \mathbb{P}\{X \geq 1 + n\}.
\end{aligned}$$ □

During the proof that $(X_t \mid X_0 = 1) \sim \mathsf{Geom}_1(e^{-\lambda t})$ we found the law of the independent sum $\mathsf{Exp}(1) + ... + \mathsf{Exp}(n-1)$. We now extend this, as promised there.

Lemma 2.2.18. *Let* $T_1, ..., T_n$ *be independent with* $T_m \sim \mathsf{Exp}(m)$ *for each* m*. Let* $S_1, ..., S_n \sim^{\text{iid}} \mathsf{Exp}(1)$*. Then,*

$$(T_n,\ T_n + T_{n-1},\ ...,\ T_n + ... + T_1) \quad \textit{are the order statistics}[2] \textit{of} \quad (S_1,\ ...,\ S_n).$$

Proof. We are going to use the memoryless property repeatedly.

By competition of Exponentials (Lemma 1.3.10), $\min\{S_1, ..., S_n\} \sim \mathsf{Exp}(n)$. Give n objects an independent $\mathsf{Exp}(1)$ timer. Then, $T_n \sim \mathsf{Exp}(n) \sim \min\{S_1, ..., S_n\}$.

Suppose that $I_1 := \arg\min_i S_i$ is the first timer to ring—ie, $S_{I_1} = \min\{S_1, ..., S_n\}$. We wait for the second to ring: $I_2 := \arg\min_{i \neq I_1} S_i$. Then, $S_{I_2} = \min_{i \neq I_1} S_i$ is the minimum over $n-1$ $\mathsf{Exp}(1)$ random variables each conditioned to be larger than S_{I_1}. But, by the memoryless property, we can remove this conditioning by shifting by S_{I_1}: $S_{I_2} \sim S_{I_1} + E_{n-1}$ where $E_{n-1} \sim \mathsf{Exp}(n-1)$ is an independent random variable.

[2]The *order statistics* of a sequence $(s_1, ..., s_n)$ is simply the sequence written in increasing order

Iterating this shows that the k-th smallest element of $\{S_1, ..., S_n\}$ is distributed as $E_n + ... + E_{n-k+1}$ where $E_m \sim \mathsf{Exp}(m)$ independently, as required. □

We emphasise that the 'best' way to find the transition probabilities $p_{m,n}(t)$ given in Proposition 2.2.15 is to first find $p_{1,n}(t)$ using Lemma **??**. Then, decompose an SBP with initial state $m > 1$ into m independent SBPs with initial state 1. Finally, find the pdf for the sum of m independent Geometric random variables.

Exercise 2.2.19. *We found differential equations for $\mu_k(t) := \mathbb{E}(X_t^k)$ where $X \sim \mathsf{SBP}_1(\lambda)$ in Lemma 2.2.14. Another method for finding a differential equation for μ_1 is given in [Nor97, Example 2.5.1]. It is very clever, not using the KFDEs, but rather the superposition property. Read that method, then apply it to μ_2 to obtain*

$$\mu_2'(t) = \mu_2(t) + 2\mu_1(t)^2.$$

Plug in the value of $\mu_1(t)$ already found and solve this DE.

2.3 Linear Birth–Death Processes

Our final section on birth–death processes allows deaths, but restricts to linear rates.

Definition 2.3.1 (Linear Birth–Death Process)**.** A continuous-time Markov chain taking values in $\mathbb{N} \cup \{\infty\}$ is a *linear birth–death process* with birth rate λ, death rate μ, immigration rate α and emigration rate β, abbreviated $\mathsf{LBDP}(\lambda, \mu; \alpha, \beta)$, if its only non-zero transition rates are

$$q_{i,i+1} = \lambda i + \alpha \quad \text{and} \quad q_{i,i-1} = \mu i + \beta \quad \text{for all} \quad i \in \mathbb{N}.$$

We always require $\lambda, \mu, \alpha, \beta \geq 0$. Its instantaneous transition-rates matrix is

$$Q = \begin{pmatrix} -\alpha & \alpha & 0 & 0 & \cdots \\ \mu+\beta & -(\mu+\lambda)-(\alpha+\beta) & \lambda+\alpha & 0 & \cdots \\ 0 & 2\mu+\beta & -2(\mu+\lambda)-(\alpha+\beta) & 2\lambda+\alpha & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}.$$

When $\alpha = \beta = 0$, it is a *simple birth–death process*, abbreviated $\mathsf{SBDP}(\lambda, \mu)$. △

Remark. If the birth rate λ and death rate μ are non-zero, then the SBDP ($\alpha = \beta = 0$) has two communicating classes: $\{0\}$, corresponding to the absorbing state 0, and $\{1, 2, ...\}$. If also $\alpha > 0$, then the LBDP has only one communicating class. △

There are some natural and important questions to ask.

- What is the extinction probability?—ie, what is $\mathbb{P}\{X_t = 0 \text{ for some } t < \infty\}$?
- Is there explosion?—and is explosion a 0–1 event, as with SBPs?
- Under what conditions does the population reach a stochastic equilibrium?

We give some answers to these questions in the remainder of the chapter.

Proposition 2.3.2 (Extinction). *Let $X = (X_t)_{t\geq 0} \sim \mathsf{SBDP}(\lambda, \mu)$. Then,*

$$\mathbb{P}_k\{X_t = 0 \text{ for some } t < \infty\} = \begin{cases} 1 & \text{if} \quad \mu \geq \lambda, \\ (\mu/\lambda)^k & \text{if} \quad \mu \leq \lambda. \end{cases}$$

Proof. Let $\hat{X} = (\hat{X}_n)_{n\geq 0}$ be the jump chain of the SBDP. Its transition matrix is

$$\Pi = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ q & 0 & p & 0 & \cdots \\ 0 & q & 0 & p & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

where $p = \lambda/(\lambda + \mu)$ and $q = \mu/(\lambda + \mu) = 1 - p$. Evidently,

$$\{X_t = 0 \text{ for some } t > 0,\ X_0 = k\} = \{\hat{X}_n = 0 \text{ for some } n > 0,\ \hat{X}_0 = k\}.$$

The latter is the "Gambler's Ruin" problem: the probability is $\min\{(q/p)^k, 1\}$. See Example 0.5.3, or [Nor97, Example 1.3.3] for a more detailed exposition. □

Next, we calculate the mgf of the SBDP using a clever argument; cf Exercise 2.2.19.

Lemma 2.3.3. *Let $X \sim \mathsf{SBDP}(\lambda, \mu)$ with $\lambda \neq \mu$. Its (negative) mgf is given by*

$$\phi_t(r) := \mathbb{E}_1(e^{-rX_t}) = \frac{(\mu - \lambda e^{-r}) - \mu(1 - e^{-r})e^{-(\mu-\lambda)t}}{(\mu - \lambda e^{-r}) - \lambda(1 - e^{-r})e^{-(\mu-\lambda)t}}.$$

Proof. Let $r > 0$ and suppose that $X_0 = 1$. Set $\psi_r(t) := \mathbb{E}_1(e^{-rX_t})$. The function ψ_r is not the mgf: it is a function of *time* t, not the *dummy variable* r. Clearly, though, if we solve it for any (r, t), then we easily construct the mgf via $\phi_t(r) = \psi_r(t)$.

Let T_1 be the time of the first birth or death. Then, $T_1 \sim \mathsf{Exp}(\lambda + \mu)$ as $X_0 = 1$ and $q_1 = \lambda + \mu$. Averaging over T_1, using the (continuous) law of total probability,

$$\psi_r(t) = \textstyle\int_0^\infty \mathbb{E}_1(e^{-rX_t} \mid T_1 = s)(\lambda + \mu)e^{-(\lambda+\mu)s} ds.$$

We consider two cases in the integral: $s < t$ and $s \geq t$.

Suppose that $s > t$. This case is trivial as the first jump happens after time t:

$$\mathbb{E}_1(e^{-rX_t} \mid T_1 = s) = e^{-r} \quad \text{as} \quad \{T_1 = s\} \subseteq \{X_u = 1 \,\forall u \in [0,t]\}.$$

Suppose that $s \leq t$. We divide according to the first birth/death event: let

$$B \coloneqq \{\text{first event is a birth}\} \quad \text{and} \quad D \coloneqq B^c = \{\text{first event is a death}\}.$$

Then, $\mathbb{P}_k\{B \mid T_1 = s\} = \lambda/(\lambda+\mu)$ for all k and s, by independence. Then,

$$X_2 = \begin{cases} 2 & \text{on} \quad \{T_1 = s\} \cap B. \\ 0 & \text{on} \quad \{T_1 = s\} \cap D. \end{cases}$$

This, together with the Markov property applied at s, gives

$$\begin{aligned} \mathbb{E}_1(e^{-rX_t} \mid T_1 = s) &= \mathbb{E}(e^{-rX_t} \mid X_s = 2)\mathbb{P}\{B\} + \mathbb{E}(e^{-rX_t} \mid X_s = 0)\mathbb{P}\{D\} \\ &= \mathbb{E}_2(e^{-rX_{t-s}})\tfrac{\lambda}{\lambda+\mu} + \tfrac{\mu}{\lambda+\mu}. \end{aligned}$$

Next, we use a superposition property analogous to Theorem 2.2.13 for $\mathsf{SBP}(\lambda)$:

$$\text{if} \quad Y, Z \sim^{\text{iid}} \mathsf{SBDP}(\lambda,\mu), \quad \text{then} \quad X = (X_t \coloneqq Y_t + Z_t)_{t\geq 0} \sim \mathsf{SBDP}(\lambda,\mu);$$

see Example Sheet 3. So, we can separate X into independent Y and Z when $X_0 = 2$:

$$\begin{aligned} \mathbb{E}_2(e^{-rX_{t-s}}) &= \mathbb{E}(e^{-r(Y_{t-s}+Z_{t-s})} \mid Y_0 = Z_0 = 1) \\ &= \mathbb{E}_1(e^{-rY_{t-s}})\mathbb{E}_1(e^{-rZ_{t-s}}) = \psi_r(t-s)^2. \end{aligned}$$

We now combine these two cases, using the substitution $u = t - s$:

$$\begin{aligned} \psi_r(t) &= \textstyle\int_t^\infty (\lambda+\mu)e^{-(\lambda+\mu)s}e^{-r}ds \\ &\quad + \textstyle\int_0^t (\lambda+\mu)e^{-(\lambda+\mu)s}\big(\psi_r(t-s)^2\frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}\big)ds \\ &= e^{-(\lambda+\mu)t-r} + e^{-(\lambda+\mu)t}\textstyle\int_0^t e^{(\lambda+\mu)u}(\mu + \lambda\psi_r(u)^2)du. \end{aligned}$$

We turn this into a DE, then solve it. First, multiply both sides by $e^{(\lambda+\mu)t}$:

$$e^{(\lambda+\mu)t}\psi_r(t) = e^{-r} + \textstyle\int_0^t e^{(\lambda+\mu)u}(\mu + \lambda\psi_r(u)^2)du.$$

Differentiating both sides wrt t, using the Fundamental Theorem of Calculus, gives

$$e^{(\lambda+\mu)t}\big((\lambda+\mu)\psi_r(t) + \psi_r'(t)\big) = e^{(\lambda+\mu)t}\big(\mu + \lambda\psi_r(t)^2\big).$$

Simplifying and rearranging gives

$$\psi_r'(t) = \mu - (\lambda+\mu)\psi_r(t) + \lambda\psi_r(t)^2 = \big(1-\psi_r(t)\big)\big(\mu - \lambda\psi_r(t)\big).$$

Solving this DE with initial condition $\psi_r(0) = e^{-r}$ gives the result, as we now show.

This is a non-linear DE, so we cannot use the method of integrating factors to solve it. Instead, partial fractions should be used when $\lambda \neq \mu$:

$$\psi' = (1-\psi)(\mu - \lambda\psi) \iff \frac{\psi'}{1-\psi} - \frac{\lambda\psi'}{\mu - \lambda\psi} = \mu - \lambda.$$

Integrating both sides with respect to t and using $\psi(0) = e^{-r}$,

$$\log(1-\psi) - \log(\mu - \lambda\psi) = -(\mu-\lambda)t + \log(1-e^{-r}) - \log(\mu - \lambda e^{-r}).$$

Combining the logs and inverting them (ie, applying exp),

$$\frac{1-\psi}{\mu - \lambda\psi} = e^{-(\mu-\lambda)t}\frac{1-e^{-r}}{\mu - \lambda e^{-r}}.$$

Rearranging this gives the claimed formula (exercise). □

Corollary 2.3.4 (cf Proposition 2.3.2). *Let $X \sim \mathsf{SBDP}(\lambda, \mu)$ with $\lambda \neq \mu$. Then,*

$$\lim_{t\to\infty} \mathbb{P}_1\{X_t = 0\} = \begin{cases} 1 & \text{if} \quad \mu \geq \lambda, \\ \mu/\lambda & \text{if} \quad \mu < \lambda. \end{cases}$$

Proof. First, suppose that $\lambda \neq \mu$. We can write the mgf of X has

$$\mathbb{E}_1(e^{-rX_t}) = \mathbb{P}_1\{X_t = 0\} + \textstyle\sum_{k=1}^{\infty} e^{-rk}\mathbb{P}_1\{X_t = k\}.$$

Taking the limit $r \to \infty$, the infinite sum disappears:

$$0 \leq \textstyle\sum_{k\geq 1} e^{-rk}\mathbb{P}_1\{X_t = k\} \leq \sum_{k\geq 1} e^{-rk} = e^{-r}/(1-e^{-r}) \to 0.$$

Thus, using this and the (negative) mgf formula (Lemma 2.3.3), we obtain

$$\mathbb{P}_1\{X_t = 0\} = \frac{\mu - \mu e^{-(\mu-\lambda)t}}{\mu - \lambda e^{-(\mu-\lambda)t}} = \frac{\mu - \mu e^{-(\lambda-\mu)t}}{\lambda - \mu e^{-(\lambda-\mu)t}}.$$

Taking $t \to \infty$ proves the claim for $\lambda \neq \mu$: if $\mu > \lambda$, then $e^{-(\mu-\lambda)t} \to 0$; if $\mu < \lambda$, then $e^{-(\lambda-\mu)t} \to 0$. The claim for $\lambda = \mu$ holds by monotonicity, squeezing $\lambda - \mu \to 0$.[3] □

Exercise 2.3.5. *Solve the last DE in the proof of Lemma 2.3.3 when $\lambda = \mu = 1$:*

$$\text{solve} \quad f'(t) = \big(f(t) - 1\big)^2 \quad \text{for the function } f.$$

Use it to find $\lim_{t\to\infty} \mathbb{P}_1\{X_t = 0\}$ in this case, verifying Corollary 2.3.4 for $\lambda = \mu$.

[3] Formally, a *coupling* is required. This is detailed in §4.3.4; see, particularly, Example 4.3.8

3 Queueing Theory

This chapter introduces the vast area of queueing theory. We can think of a queue as produced when a sequence of customers arrives at a system and wait to be served. The inter-arrival times are typically random and the customers are served according to some scheme: typically, this is "first in, first out", abbreviated *FIFO*.

A queue is characterised by several components.

- Arrival process; often, a Poisson process
- Service time distribution; often, an Exponential distribution
- Number of servers; often, one
- Waiting capacity; often, infinite
- Queue scheme or discipline; often, FIFO

A common notation used to describe queues takes the form $\cdot/\cdot\cdot/\cdots$, such as $M/M/1$ or $M/G/1$, or sometimes $\cdot/\cdot\cdot/\cdots/\cdots\cdot$, such as $M/M/K_1/K_2$. This notation encodes, in sequence, the arrival process, the service process and the server characteristics. The first letter describes the arrival process; eg, M stands for *Markovian input* or *memoryless inter-arrival times* and G for *general input*. The second letter describes the service process: eg, M stands for *Markovian/memoryless* and G for *general service time*. The third symbol is an integer specifying the number of servers and the fourth an integer specifying the capacity of the system; the fourth is assumed infinite if omitted.

The three most common examples are these.

- $M/M/1$: Markovian input and service times, one server and infinite capacity;
- $M/M/K_1/K_2$: Markovian input and service times, K_1 servers and K_2 capacity;
- $M/G/1$: Markovian input, general service time, one server and infinite capacity.

We study these three examples. We mainly consider *equilibrium behaviour*:

- Is there a limiting equilibrium? If so, what is the equilibrium distribution?

- What are the values of various *measures of effectiveness* for these queues, such as mean queue length, probability of an empty queue or mean waiting time?

Before proceeding with the formal introduction of queues, we introduce terminology.

We typically denote the process by $X = (X_t)_{t\geq 0}$. It is a birth–death chain: a $+1/-1$ transition corresponds to the *arrival/departure* of a customer.

Inter-arrival time A: time between the arrivals of customers

Service time S: time from arrival at front of queue to departure from the system

Queue length X: the number of people waiting for service or being served—ie, the number of people in the system. Queue length 1 means that 1 person is being served and no-one is waiting for service

Queueing/waiting time T_q: time between a customer's arrival and the start of their service. If there is no-one in the system when they arrive, their queueing time is 0

Sojourn/total time T_s: entire time that the customer spends in the system—ie, the sum of their queueing and service times

Caution. There is a small conflict in terminology here: the queue length *does* include the person being served, but the queueing time *does not* include the service time.

3.1 The Single-Server Markovian Queue

This section is dedicated to the most fundamental queue: the $M/M/1$ queue.

Definition 3.1.1 ($M/M/1$ Queue)**.** The $M/M/1$ queue is characterised as follows.

- Customers arrive according to a Poisson process of rate λ, for some $\lambda > 0$. That is, the inter-arrival times of customers are independent $\mathsf{Exp}(\lambda)$-s.
- Service times are independent, Exponentially distributed with rate μ, for some $\mu > 0$, and independent of the arrival process.
- Customers are served in order of arrival.

There is a single communicating class $\{0, 1, 2, ...\}$ when $\lambda, \mu > 0$. △

Applying the results of the previous chapters gives the following theorem.

Theorem 3.1.2. *The queue length $X = (X_t)_{t\geq 0}$ is a birth–death chain with constant birth rate λ and death rate μ. See Figure 3.1 for the corresponding state diagram.*

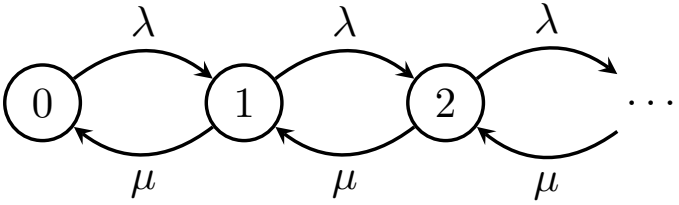

Figure 3.1. State diagram for $M/M/1$ queue

Proof. The proof follows the usual pattern. The jump times are Exponentially distributed and with constant rate—ie, their rate does not even depend on the current state, never mind the history. So, the Markov property of the process X follows from the memoryless property of the Exponential distribution. An arrival causes X to increment $+1$ and a departure -1. Hence, $q_{n,n+1} = \lambda$ and $q_{n+1,n} = \mu$ for $n \geq 0$. □

The queue uses a FIFO discipline, but this is actually somewhat inconsequential.

Exercise 3.1.3. *Consider two possible adjustments to the queue discipline.*

1. *When a customer enters the system, they join the back of the queue. When they reach the front of the queue, if they are not served within the next 1 unit of time, then they return to the back of the queue.*
2. *When a customer enters the system, they start their service immediately. If a customer was already being served, then that customer is displaced, being moved to second in the queue (ie, next to be served).*

Show that both these adaptations lead to a queue length X with the $M/M/1$ law.

3.1.1 Limiting Behaviour

We now consider the limiting behaviour of an $M/M/1$ queue. In particular, we want to understand when a limiting equilibrium exists and, if it does, what the equilibrium distribution is. Recall from §1.4 that we must search for a distribution π which solves the (global) balance equations $\pi Q = 0$—ie, $\sum_{j \in I} \pi_j q_{j,i} = 0$ for all $i \in I$. In the specific case of an $M/M/1$ queue, this boils down to the following system of equations:

$$-\lambda\pi_0 + \mu\pi_1 = 0; \quad -(\lambda + \mu)\pi_n + \lambda\pi_{n-1} + \mu\pi_{n+1} = 0 \quad \text{for} \quad n \geq 1.$$

Lemma 3.1.4. *The number of customers in an $M/M/1$ queue has an equilibrium distribution if and only if $\lambda < \mu$. When $\lambda < \mu$, this equilibrium distribution π satisfies $\pi_n = \rho^n(1-\rho)$ for all $n \geq 0$, where $\rho = \lambda/\mu < 1$ is the utilisation (factor).*

Proof. The first balance equation gives $\pi_1 = (\lambda/\mu)\pi_0 = \rho\pi_0$. Plugging this into

$$-(\lambda+\mu)\pi_1 + \lambda\pi_0 + \mu\pi_2 = 0 \quad \text{gives} \quad \mu\pi_2 = (\lambda+\mu)\tfrac{\lambda}{\mu}\pi_0 - \lambda\pi_0 = (\lambda^2/\mu)\pi_0,$$

and so $\pi_2 = (\lambda/\mu)^2\pi_0 = \rho^2\pi_0$. We now guess $\pi_n = \rho^n\pi_0$ for all $n \geq 0$ and prove it by induction. We have already verified the two base cases. By hypothesis,

$$\mu\pi_{n+1} = (\lambda+\mu)\rho^n\pi_0 - \lambda\rho^{n-1}\pi_0 = \big((\lambda+\mu)\tfrac{\lambda}{\mu} - \lambda\big)\rho^{n-1}\pi_0 = (\lambda^2/\mu)\rho^{n-1}\pi_0,$$

and so $\pi_{n+1} = (\lambda/\mu)^2\rho^{n-1}\pi_0 = \rho^{n+1}\pi_0$, completing the induction.

This defines a distribution if and only if $\sum_{n\geq 0}\rho^n < \infty$—ie, $\rho < 1$. If $\rho < 1$, then

$$\pi_0 = 1/\textstyle\sum_{n\geq 0}\rho^n = 1-\rho, \quad \text{so} \quad \pi_n = \rho^n(1-\rho) \quad \text{for all} \quad n \geq 0. \qquad \square$$

We also encountered the *detailed balance equations* in §1.4. These often make it easier to find the invariant distribution. This is the case here.

Exercise 3.1.5. *Use the detailed balance equations to prove Lemma 3.1.4.*

Remark. It is interesting to note that the equilibrium distribution is Geometric:

$$\pi_n = \rho^n(1-\rho) = \mathbb{P}\{\mathsf{Geom}_0(\rho) = n\}.$$

Hence, the queue length in equilibrium has the discrete memoryless property: knowing that the queue has length at least n does not tell you how much longer than n it is.

We see later in Burke's theorem (Theorem 3.3.1) a rather stronger version of this: in equilibrium, the *departure process*, measuring the times at which service is completed, is also a Poisson process and is independent of the *arrival process*. In particular, knowing how many customers have been served in a certain period gives no information on the number of customers queuing at the end of that period. $\triangle$

Exercise 3.1.6 (Discrete Memoryless Property)**.** *Let $N \sim \mathsf{Geom}(\rho)$ for some $\rho \in (0,1)$. Show that $(N-k \mid N \geq k) \sim \mathsf{Geom}(\rho)$ for all $k \in \mathbb{N}$.*

Exercise 3.1.7. *Let $X = (X_t)_{t\geq 0}$ denote the queue length of an $M/M/1$ queue. Spot an embedded random walk with reflecting barrier and use theory from discrete-time Markov chains to show the following.*

1. *If $\mu \leq \lambda$, then the queue length visits 0 infinitely often.*
2. *If $\mu > \lambda$, then the queue length visits 0 finitely often.*

3.1.2 Measures of Effectiveness

We now consider measures of effectiveness for the $M/M/1$ queue, such as mean queue length and waiting times. We always suppose that the queue is in equilibrium when studying these; in particular, $\rho = \lambda/\mu < 1$.

Definition 3.1.8 (Queue Lengths)**.** Let N denote the number of customers in an $M/M/1$ system in equilibrium. Let N' denote the number of customers in the queue who are not being served; so, $N' = N - \mathbf{1}\{N > 0\}$. △

Lemma 3.1.9 (Queue Lengths in Equilibrium)**.** *The following hold:*

$$\mathbb{P}\{N = 0\} = 1 - \rho \quad \textit{and} \quad \mathbb{E}(N) = \rho/(1-\rho) = \lambda/(\mu - \lambda);$$
$$\mathbb{P}\{N' = 0\} = 1 - \rho^2 \quad \textit{and} \quad \mathbb{E}(N') = \rho^2/(1-\rho) = \rho\lambda/(\mu - \lambda).$$

Proof. The equilibrium distribution π satisfies $\pi_n = \rho^n(1-\rho)$. Thus,

$$\mathbb{E}(N) = \textstyle\sum_{n\geq 0} n\pi_n = \sum_{n\geq 0} n\rho^n(1-\rho).$$

A standard trick for calculating such sums is to write the summand as a derivative:

$$\mathbb{E}(N) = \rho(1-\rho)\tfrac{d}{d\rho}\textstyle\sum_{n\geq 0}\rho^n = \rho(1-\rho)\tfrac{d}{d\rho}(1-\rho)^{-1} = \tfrac{\rho}{1-\rho} = \tfrac{\lambda}{\mu-\lambda}.$$

The probability that the *system* is empty in equilibrium is simply

$$\mathbb{P}\{N = 0\} = \pi_0 = 1 - \rho.$$

On the other hand, the probability that the *queue* is empty is

$$\mathbb{P}\{N' = 0\} = \mathbb{P}\{N \in \{0,1\}\} = \pi_0 + \pi_1 = (1-\rho) + \rho(1-\rho) = 1 - \rho^2.$$

If no customer is being served, then $N' = N$; if one is, then $N' = N - 1$. Hence,

$$N' = N - \mathbf{1}\{N > 0\} \quad \text{and so} \quad \mathbb{E}(N') = \mathbb{E}(N) - \mathbb{P}\{N > 0\} = \tfrac{\rho}{1-\rho} - \rho = \tfrac{\rho^2}{1-\rho}. \quad \square$$

Definition 3.1.10 (Waiting Times)**.** Let T_{s} denote the *sojourn time* of a new arrival to an $M/M/1$ system in equilibrium—ie, the total time the customer spends in the system, from arrival to departure. Let T_{q} denote its *queueing time*—ie, not including its service time. So, $T_{\mathsf{s}} = T_{\mathsf{q}} + S$, where S is the service time (independent of T_{q}). △

Lemma 3.1.11 (Waiting Times in Equilibrium)**.** *The following hold:*

$$\mathbb{E}(T_{\mathsf{s}}) = \tfrac{1}{\mu-\lambda} \quad \textit{and} \quad \mathbb{E}(T_{\mathsf{q}}) = \tfrac{\rho}{\mu-\lambda}.$$

Proof. Conditional on $N \geq 1$ customers in the system at arrival, the *queueing* time T_{q} can be written as a sum of $N = N' + 1$ iid $\mathsf{Exp}(\mu)$ service times:

$$T_{\mathsf{q}} = \sum_{j=1}^{N} E_\mu^{(j)} \quad \text{where} \quad E_\mu^{(1)}, E_\mu^{(2)}, \ldots \sim^{\mathsf{iid}} \mathsf{Exp}(\mu).$$

Now, conditional on $N = m \geq 1$, the length of this sum is non-random—it is m. So,

$$\mathbb{E}\big(e^{tT_{\mathsf{q}}} \mid N = n\big) = \mathbb{E}\big(\exp\big(t\textstyle\sum_{j=1}^{n} E_\mu^{(j)}\big)\big) = \mathbb{E}\big(e^{t\mathsf{Exp}(\mu)}\big)^n \quad \text{for} \quad n \geq 1$$

since the $E_\mu^{(j)}$ are iid. Now, recall the mgf of $\mathsf{Exp}(\mu)$: for $t < \mu$,

$$\mathbb{E}\big(e^{t\mathsf{Exp}(\mu)}\big) = \textstyle\int_0^\infty \mu e^{-\mu s} e^{st} ds = \frac{\mu}{\mu - t} \int_0^\infty (\mu - t) e^{-(\mu - t)s} ds = \frac{\mu}{\mu - t}.$$

Using the memoryless property for Geometrics (Exercise 3.1.6), provided $t < \mu - \lambda$,

$$\begin{aligned}
\mathbb{E}(e^{tT_{\mathsf{q}}} \mid N \geq 1) &= \textstyle\sum_{m \geq 1} \mathbb{E}(e^{tT_{\mathsf{q}}} \mid N = m) \mathbb{P}\{N = m \mid N \geq 1\} \\
&= \textstyle\sum_{m \geq 1} \big(\frac{\mu}{\mu - t}\big)^m (1 - \rho) \rho^{m-1} \\
&= \frac{(1-\rho)\mu}{\mu - t} \textstyle\sum_{m \geq 1} \big(\frac{\lambda}{\mu - t}\big)^{m-1} \\
&= \frac{\mu - \lambda}{\mu - t} \big(1 - \frac{\lambda}{\mu - t}\big)^{-1} = \frac{\mu - \lambda}{\mu - \lambda - t}.
\end{aligned}$$

The mgf characterises the distribution, so $(T_{\mathsf{q}} \mid N \geq 1) \sim \mathsf{Exp}(\mu - \lambda)$. Hence,

$$\mathbb{E}(T_{\mathsf{q}}) = \mathbb{E}(T_{\mathsf{q}} \mid N \geq 1) \mathbb{P}\{N \geq 1\} = \frac{\rho}{\mu - \lambda}$$

as $\mathbb{P}\{N \geq 1\} = \rho$ and $T_{\mathsf{q}} = 0$ if $N = 0$—the arrival is served immediately in this case.

Finally, the sojourn time is $T_{\mathsf{s}} = T_{\mathsf{q}} + S$, where $S \sim \mathsf{Exp}(\mu)$. Hence,

$$\mathbb{E}(T_{\mathsf{s}}) = \mathbb{E}(T_{\mathsf{q}}) + \mathbb{E}(\exp(\mu)) = \frac{1}{\mu} + \frac{\rho}{(1-\rho)\mu} = \frac{1}{\mu}\big(1 + \frac{\rho}{1-\rho}\big) = \frac{1}{(1-\rho)\mu} = \frac{1}{\mu - \lambda}. \qquad \square$$

Corollary 3.1.12 (Little's Law). *Little's law holds for $M/M/1$ queues:*

$$\mathbb{E}(N) = \lambda \mathbb{E}(T_{\mathsf{s}}).$$

Little's law actually holds in much more generality than $M/M/1$ queues. We encounter it again when discussing $M/G/1$ queues—ie, queues with Markovian inputs, but general service times. The same applies to the Pollaczek–Khintchine formula.

Exercise 3.1.13 (Pollaczek–Khintchine Formula). *Show that the Pollaczek–Khintchine holds for $M/M/1$ queues:*

$$\mathbb{E}(T_{\mathsf{q}}) = \frac{\lambda \mathbb{E}(S^2)}{2(1-\rho)} = \frac{\lambda(\mathbb{V}\mathrm{ar}(S) + \mathbb{E}(S)^2)}{2(1-\rho)}$$

where S has the service time distribution—so, $S \sim \mathsf{Exp}(\mu)$ for $M/M/1$ queues.

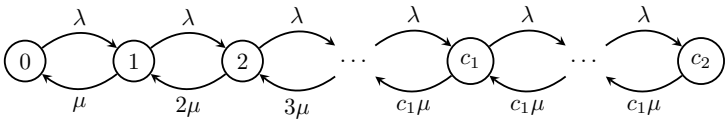

Figure 3.2. State diagram for $M/M/1$ queue

3.2 Multi-Server Queues with Finite Capacity

Multi-server queues possess $K_1 \geq 1$ servers. The capacity may be some finite integer $K_2 \in [K_1, \infty)$, in which case there is a waiting room for $K_2 - K_1 \geq 0$ additional customers, or it may be infinite. Customers who arrive to a full waiting room are turned away and never return. Always, service times and arrivals are independent.

Our interest focusses primarily on the probability that a customer is turned away when the queue is in statistical equilibrium.

Example 3.2.1 (Old-Fashioned Telephone Exchange)**.** In an old-fashioned telephone exchange, calls arrive accordion to a Poisson process of rate λ, for some $\lambda > 0$. There are c servers and calls are lost if they arrive when all servers are occupied: there is no waiting room. In the above notation, $K_1 = K_2 = K$. △

Definition 3.2.2 ($M/M/K_1/K_2$ Queue)**.** The $M/M/K_1/K_2$ queue is characterised as follows.

- Customers arrive according to a Poisson process of rate λ, for some $\lambda > 0$.
- There are K_1 servers, each serving (independently) at rate μ, for some $\mu > 0$.
- Arriving customers are directed to an arbitrary free server if one exists and to the waiting room otherwise, provided there are at most $K_2 - K_1$ customers in the waiting room already. Otherwise, the customer is turned away.
- When a server becomes free, a customer moves from the waiting room to it. △

Following arguments analogous to those for the $M/M/1$ determines the rates.

Theorem 3.2.3. *The number of customers in an $M/M/K_1/K_2$ queue is a birth–death process with the following non-zero transition rates, summarised in Figure 3.2:*

$$q_{n,n+1} = \lambda \quad \text{for} \quad n \in \{0, \ldots, K_2 - 1\};$$

$$q_{n,n-1} = \begin{cases} n\mu & \text{for} \quad n \in \{1, \ldots, K_1\}, \\ K_1\mu & \text{for} \quad n \in \{K_1 + 1, \ldots, K_2\}. \end{cases}$$

The main result of this section, which is *Erlang's formula* (Theorem 3.2.5), is for multi-server queues with no waiting room: $K_1 = K_2 = K$.

Exercise 3.2.4 (See Example Sheet 3). *Let $K \in \mathbb{N}$ and define π by*

$$\pi_n := \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \Big/ \sum_{m=0}^{K} \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \quad \text{for} \quad n \in \{0, 1, ..., K\}.$$

- *Show that π is a distribution and that it satisfied the detailed-balanced equations associated to the queue length of an $M/M/K/K$ queue.*
- *Show that the expected number of servers in use in equilibrium is $\frac{\lambda}{\mu}(1 - \pi_K)$.*

Theorem 3.2.5 (Erlang's Formula). *For an $M/M/K/K$ queue with service rate μ and arrival rate λ, the equilibrium probability of being turned away is*

$$\pi_K = \frac{1}{K!}\left(\frac{\lambda}{\mu}\right)^K \pi_0 = \frac{1}{K!}\left(\frac{\lambda}{\mu}\right)^K \Big/ \sum_{m=0}^{K} \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m.$$

Proof. A customer is turned away if and only if all K servers are in use. This has equilibrium probability π_K, which is given by Exercise 3.2.4. □

Remark. Remarkably, Erlang's formula generalises to $M/G/K/K$ queues—ie, those with Markovian input and iid, but not necessarily Exponential, service times. △

3.3 Reversibility and the Departure Process

Recall that a Markov chain is *reversible* if, when started from equilibrium, it is statistically impossible to tell whether it is being run forward or backward in time. Formally, the *detailed-balance* equations must hold:

$$\pi_i q_{i,j} = \pi_j q_{j,i} \quad \text{for all} \quad i, j \in I.$$

A key application of reversibility for queueing theory is that it allows us to study the *departure* process in equilibrium. The following theory holds for $M/M/c$ queues, for any $c \in \mathbb{N}$, as well as $M/M/\infty$ queues. We prove it for $M/M/1$ queues.

Theorem 3.3.1 (Burke's Theorem). *Suppose $\lambda < \mu$. In equilibrium, the departure process of an $M/M/1$ queue with arrival rate λ and service rate μ is a $\mathsf{PP}(\lambda)$. Further, the number of customers in the queue at a fixed time t is independent of the departure process before time t.*

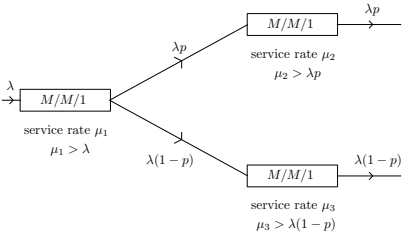

Figure 3.3. Illustration of network of queues

Proof. Let X denote the queue length. Then, X is reversible, by Exercise 3.2.4; that is, for a given $T > 0$, the processes $(X_t)_{0\le t\le T}$ and $(\hat{X}_t := X_{T-t})_{0\le t\le T}$ have the same distribution, when $X_0 \sim \pi$. Hence, $\hat{X}$ experiences jumps of size $+1$ at constant rate λ. But, $\hat{X}$ has a jump of size $+1$ at time t if and only if a customer departs the queue at time $T - t$. Therefore, departures from X becomes arrivals for $\hat{X}$. The time-reversal of a $\mathsf{PP}(\lambda)$ is a (negative) $\mathsf{PP}(\lambda)$. Hence, the departure process is a $\mathsf{PP}(\lambda)$.

For the independence, let A denote the arrival times and D the departure times. Clearly, the state of the queue up to time $T - t$ is independent of *future* arrivals:

$$(X_s)_{0\le s\le T-t} \quad \text{is independent of} \quad A \cap (T-t, T].$$

Arrivals for the forward process are departures for the backward process. So,

$$(\hat{X}_s)_{0\le s\le T-t} = (X_s)_{t\le s\le T} \quad \text{is independent of} \quad D \cap [0, t),$$

by the same logic. Hence, $(X_s)_{s\ge t}$ is independent of departures up to time t. □

It is crucial, here, that the process is in equilibrium in order to apply reversibility. Indeed, if we impose $X_0 = 5$, then the first departure will happen after $\mathsf{Exp}(\mu)$, not $\mathsf{Exp}(\lambda)$. It also imposes $\lambda < \mu$: otherwise no equilibrium state exists.

3.4 Queues in Tandem

We can use Burke's theorem to justify feeding the output (departures) of one queue into the input (arrivals) of another queue. We can even get a network of queues by sending some of the departures to one queue and some to another; see Figure 3.3.

The network need not be a tree. However, each customer should only pass through a single server once, so that arrivals and departures are independent. We next study the simplest network: a series of $M/M/1$ queues. The *Advanced Topics* for ST406 includes study of more elaborate networks of queues called *Jackson Networks* (§3.6*).

Definition 3.4.1 (Queues in Tandem)**.** We study a sequence of J queues in tandem:

- customers arrive as a $\mathsf{PP}(\lambda)$ to queue 1;
- the j-th server serves at rate μ_j;
- upon leaving queue $j \in \{1, \dots, J-1\}$, they join the queue for queue $j+1$;
- upon leaving queue J, they leave the system.

Service times are independent, including those of the same customer in different queues, as are arrivals to queue 1. △

Theorem 3.4.2. *Let $X^j = (X_t^j)_{t\geq 0}$ denote the queue length in the j-th queue, for $j \in [J]$. Suppose that $\lambda < \min_{j\in[J]} \mu_j$. Then, the invariant distribution of X is*

$$\pi(x^1, \dots, x^J) = \textstyle\prod_{j\in[J]}(1-\rho_j)\rho_j^{x^j} \quad \text{where} \quad \rho_j := \lambda/\mu_j \quad \text{for} \quad j \in [J];$$

that is, if $(\mathcal{X}^1, \dots, \mathcal{X}^J) \sim \pi$, then $\mathcal{X}^j \sim \mathsf{Geom}_0(\rho_j)$ marginally for each $j \in [J]$ and the components are jointly independent. Moreover, X is reversible wrt π.

Proof. The case $J = 1$ corresponds to a single $M/M/1$ queue; the invariant distribution was found in Lemma 3.1.4. We now prove the theorem for $J = 2$.

The possible transitions are

$$(x^1, x^2) \to \begin{cases} (x^1+1, x^2) & \text{with rate} \quad \lambda, \\ (x^1-1, x^2+1) & \text{with rate} \quad \mu_1 \quad \text{if} \quad x^1 \geq 1, \\ (x^1, x^2-1) & \text{with rate} \quad \mu_2 \quad \text{if} \quad x^2 \geq 1; \end{cases}$$

these correspond to an arrival to queue 1 entering the system, a departure from queue 1 causing an arrival to queue 2 and a departure from queue 2 leaving the system. We can check by direct computation that $\pi Q = 0$ if and only if π has the claimed form.

A more elegant and conceptual proof uses Burke's theorem. The first server behaves marginally as an $M/M/1$ queue. But, Burke's theorem says that, at equilibrium, the departure process of the first $M/M/1$ queue is a $\mathsf{PP}(\lambda)$. Hence, marginally, both are $M/M/1$ queues, so have the claimed invariant distribution. It remains to check independence. This holds because X_t^2 depends only on X_0^2 and the departure process from the first queue, which is independent of X_t^1 by Burke's theorem. □

Exercise 3.4.3. *Extend the second argument to a general number $J \geq 2$ of queues using induction and Burke's theorem (Theorem 3.3.1).*

Remark. If $X_0 \sim \pi$, then the random variables X_t^j are independent for different $j \in [J]$ for a *fixed* t. The *processes* $X^j = (X_t^j)_{t\geq 0}$ cannot be independent, though: a jump -1 for X^j with $j < J$ implies a jump $+1$ for X^{j+1}. △

The technique can establish independence of waiting times in successive queues.

Exercise 3.4.4. *Argue that the sojourn time—ie, queueing plus service time—of a customer in queue* 1 *is independent of departures from queue* 1 *prior to their departure. Deduce that, in equilibrium, the sojourn times of a customer at each of the queues are independent.*

The same idea can be extended to a tree-like network of queues. The technique is fragile, however: it does not allow a customer to leave a later queue and return to an earlier one. This scenario is studied in the *Advanced Topics* for the ST406 variant of this course. The set-up is more general, but the results are somewhat less general.

3.5 Queues with Non-Markovian Service Times

We now consider queues where the number of customers in the system is no longer Markovian. The example on which we concentrate is a modification of an $M/M/1$ queue where the service times are still iid, but no longer need be Exponential.

Definition 3.5.1 ($M/G/1$ Queue)**.** Customers arrive according to a Poisson process of rate λ. There is a single server whose service time has some specified distributed, which need not be exponential, with mean $1/\mu$. Services and arrivals are independent. △

The number of customers in the system is no longer Markovian. Eg, imagine the case in which service times are deterministically one unit of time. If we know that the queue length X satisfies $X_t = 3$ for all $t \in [1, 1.9999]$, then the chance that $X_2 = 2$ is very high: there is a departure and no arrival. On the other hand, if it were Markovian, then the overwhelmingly most likely scenario would be $X_2 = X_{1.9999} = 3$.

Not all hope is lost, however. There is an *embedded Markov chain*, obtained by considering this random process at a countable sequence of random times.

Definition 3.5.2 (Embedded Chain)**.** Given a jump process $X = (X_t)_{t\geq 0}$ on $\mathbb{N}$, with jumps of size ± 1, define Y_n to be the value of X after the n-th -1 jump. Set $Y_0 \coloneqq X_0$. Then, $Y = (Y_n)_{n\geq 0}$ is the *embedded chain*. In the case of queues, Y_n is the number of customers in the system immediately after the n-th departure. △

The embedded chain Y of an $M/G/1$ queue X is a Markov chain in discrete time.

Theorem 3.5.3. *Let X be an $M/G/1$ queue and Y its embedded chain. Then, $Y = (Y_n)_{n\geq 1}$ is a discrete-time Markov chain.*

Proof. We need to show that

$$\mathbb{P}\{Y_{n+1} = y_{n+1} \mid Y_n = y_n, ..., Y_1 = y_1\}$$

does not depend on $(y_1, ..., y_{n-1})$. This follows from Theorem 2.1.7 (Strong Poisson Increments): take T there to be the time at which the n-th service is completed; arrivals are independent of service, so we can take t to be the $(n+1)$-th time. □

Definition 3.5.4. Let A_n denote the number of customers which arrive during the service of the n-th customer. Let S denote the typical service time of a customer—ie, a random variable with distribution given by the service time—and

$$k_n \coloneqq \mathbb{P}\{n \text{ customers arrive during } S\}. \qquad \triangle$$

If there are $Y_n = y > 0$—ie, there are $y > 0$ customers in the queue immediately after the n-th service—and A_{n+1} customers arrive during the $(n+1)$-th service, then $Y_{n+1} = Y_n + A_{n+1} - 1$; the -1 term accounts for the fact that one person as *just* departed. If $Y_n = 0$, then we must wait for someone to join the queue before the $(n+1)$-th service can commence; we then proceed as if $Y_n = 1$. Hence,

$$Y_{n+1} = \begin{cases} Y_n + A_{n+1} - 1 & \text{if} \quad Y_n \geq 1, \\ A_{n+1} & \text{if} \quad Y_n = 0. \end{cases}$$

This can be written succinctly as

$$Y_{n+1} = Y_n + A_{n+1} - \mathbf{1}\{Y_n > 0\} = Y_n + A_{n+1} - 1 + \mathbf{1}\{Y_n = 0\}.$$

Thus, the transition matrix P for the embedded chain Y is

$$P = \begin{pmatrix} k_0 & k_1 & k_2 & k_3 & \cdots \\ k_0 & k_1 & k_2 & k_3 & \cdots \\ 0 & k_0 & k_1 & k_2 & \cdots \\ 0 & 0 & k_0 & k_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Eg, if there are 4 customers in the queue just after a service, then the probability that there are 6 just after the next service is k_3; indeed, 3 must have arrived in this period.

Lemma 3.5.5. *We have*

$$k_n = \mathbb{E}\big((\lambda S)^n \exp(-\lambda S)/n!\big) \quad \textit{and} \quad \textstyle\sum_{n\geq 0} n k_n = \lambda \mathbb{E}(S) = \lambda/\mu.$$

Proof. Applying the tower property, conditioning on the service length S,

$$k_n = \mathbb{E}(\mathbb{P}\{n \text{ customers arrive during } S \mid S\}).$$

Now, the service and arrival times are independent. So, this inner probability is simply

$$\mathbb{P}\{n \text{ customers arrive during } S \mid S\} = \mathbb{P}\{\mathrm{Pois}(\lambda S) = n \mid S\} = (\lambda S)^n \exp(-\lambda S)/n!.$$

Plugging this into the expectation gives the claimed form of k_n.

Turning to the second part, we can sum nk_n over $n \geq 0$ (or $n \geq 1$), then exchange expectation and summation. This cancels the $\exp(-\lambda S)$, as we now show:

$$\begin{aligned}\textstyle\sum_{n\geq 0} nk_n &= \textstyle\sum_{n\geq 1} \mathbb{E}\big((\lambda S)\cdot(\lambda S)^{n-1}\exp(-\lambda S)/(n-1)!\big)\\ &= \mathbb{E}\big(\lambda S \exp(-\lambda S)\textstyle\sum_{m\geq 0}(\lambda S)^m/m!\big) = \lambda\mathbb{E}(S) = \lambda/\mu.\end{aligned}$$

We emphasise that this did not require that the system be Markovian. □

If the embedded Markov chain $(Y_n)_{n\geq 1}$ is irreducible and aperiodic, then it is natural to ask if an equilibrium distribution $(\pi_n)_{n\geq 1}$ exists. If it does, then, $\pi = \pi P$:

$$\pi_n = \pi_0 k_n + \textstyle\sum_{m=1}^{n+1} \pi_m k_{n-(m-1)} \quad \text{for} \quad n \geq 0.$$

This can be expressed in terms of *generating functions*.

Definition 3.5.6. Write K and Π for the *generating functions* of $(k_n)_{n\geq 0}$ and $(\pi_n)_{n\geq 0}$:

$$K(z) \coloneqq \textstyle\sum_{n\geq 0} k_n z^n \quad \text{and} \quad \Pi(z) \coloneqq \textstyle\sum_{n\geq 0} \pi_n z^n \quad \text{for} \quad z \in \mathbb{R}. \qquad \triangle$$

Formally differentiating, justified rigorously by Abel's theorem, gives

$$K'(1) = \textstyle\sum_{n\geq 0} nk_n = \lambda\mathbb{E}(S) = \lambda/\mu.$$

Lemma 3.5.7. *If $(Y_n)_{n\geq 1}$ admits an equilibrium distribution $(\pi_n)_{n\geq 0}$, then*

$$\Pi(z) = \frac{\Pi(0)(1-z)K(z)}{K(z)-z}.$$

Proof. Applying the equilibrium equations and manipulating the sums,

$$\begin{aligned}\Pi(z) &= \textstyle\sum_{n\geq 0} \pi_0 k_n z^n + \sum_{n\geq 0}\big(\sum_{m=1}^{n+1} \pi_m k_{n-m+1}\big) z^n\\ &= \pi_0 K(z) + \textstyle\sum_{m\geq 1} \pi_m z^{m-1} \sum_{n\geq m-1} k_{n-(m-1)} z^{n-(m-1)}\\ &= \Pi(0)K(z) + \textstyle\sum_{m\geq 1} \pi_m z^{m-1} \sum_{\ell\geq 0} k_\ell z^\ell\\ &= \Pi(0)K(z) + K(z)\big(\Pi(z)-\Pi(0)\big)/z.\end{aligned}$$

Rearranging and solving for $\Pi(z)$ gives the claimed equality. □

If the embedded Markov chain $(Y_n)_{n\geq 1}$ is positive recurrent, which is the case when $\lambda < \mu$—ie, $K'(1) < 1$—then it turns out that the $M/G/1$ queue has an equilibrium distribution. This equilibrium distribution satisfies *Little's law* and the *Pollaczek–Khintchine* formula; see Theorems 3.5.8 and 3.5.9, respectively. We explain why this is true for the remainder of the section. Recall that we already met these formulas when discussing $M/M/1$ queues.

We sketch the proof for both results under the following assumption.

> The number of customers and the sojourn time for a virtual arrival in the embedded Markov chain are the same in distribution as for a typical arrival in the $\mathsf{PP}(\lambda)$ stream of incoming customers.

This simply means that we prove the results for the embedded Markov chain and assume that they carry over to the $M/G/1$ queue itself.

Proving that the results transfer requires *ergodic theory*. It allows the use of time averages, for which the embedded Markov chain gives information, to give information about statistical averages or expectations on the $M/G/1$ queue. Ergodic theory lies (well) beyond the scope of this course, however.

Recall from Definitions 3.1.8 and 3.1.10 that N is the number of customers and T_{s} is the sojourn time—ie, the time until completion of service in the queue.

Theorem 3.5.8 (Little's Law). *Suppose that $\lambda < \mu$. Then, in equilibrium,*

$$\mathbb{E}(N) = \lambda \mathbb{E}(T_{\mathsf{s}}).$$

Proof. Consider a time at which a customer leaves the system. They leave behind N customers, all of whom arrived during the wait time T_{s} of the exiting customer, as using FIFO. During that time, conditional on the value of T_{s}, $\mathrm{Pois}(\lambda T_{\mathsf{s}})$ new customers arrived—arrivals before and after the customer who just left are independent, by Theorem 2.1.7 (Strong Poisson Increments). Therefore, the mean number of customers in the system $\mathbb{E}(N)$ can also be expressed as

$$\mathbb{E}(\mathrm{Pois}(\lambda T_{\mathsf{s}})) = \mathbb{E}(\mathbb{E}(\mathrm{Pois}(\lambda T_{\mathsf{s}}) \mid T_{\mathsf{s}})) = \lambda \mathbb{E}(T_{\mathsf{s}}). \qquad \square$$

Little's law actually holds in much more generality; see, eg, [KY14, Theorem 2.13].

We now turn to the Pollaczek–Khintchine formula. Recall that $1/\mu = \mathbb{E}(S)$ is the expectation of the typical service time S. Let $\sigma^2 \coloneqq \mathbb{V}\mathrm{ar}(S)$ be its variance. Recall also, from Definition 3.1.10, that T_{q} is the queueing (not sojourn) time.

Theorem 3.5.9 (Pollaczek–Khintchine Formula). *Suppose that $\lambda < \mu$. Let $\rho \coloneqq \lambda/\mu$ be the utilisation factor; so, $\rho < 1$. Then, in equilibrium,*

$$\mathbb{E}(T_q) = \frac{\lambda(\sigma^2 + 1/\mu^2)}{2(1-\rho)} = \frac{\lambda \mathbb{E}(S^2)}{2(1-\rho)}.$$

The proof of this theorem builds on a number of auxiliary results. We assume that the service time S has a (negative) mgf in an interval around 0: there exists $r_0 > 0$ such that $\sup_{r\in(-r_0,r_0)} \mathbb{E}(e^{-rS}) < \infty$. This allows us to swap derivatives and expectations.

Lemma 3.5.10. *Suppose that $\rho = \lambda/\mu < 1$. Then,*

$$1 - \rho = 1 - K'(1) = \pi_0.$$

Proof. The embedded Markov chain is positive recurrent and has an equilibrium distribution π since $\rho < 1$. Taking $z \uparrow 1$ in Lemma 3.5.7 gives

$$\Pi(1) = \lim_{z\uparrow 1} \frac{\Pi(0)(1-z)K(z)}{K(z) - z}.$$

Both the numerator and denominator vanish when $z = 1$. Applying l'Hôpital's rule,

$$\begin{aligned} 1 = \Pi(1) &= \pi_0 \lim_{z\uparrow 1} \frac{\frac{d}{dz}\big((1-z)K(z)\big)}{\frac{d}{dz}\big(K(z) - z\big)} \\ &= \pi_0 \lim_{z\uparrow 1} \frac{-K(z) + (1-z)K'(z)}{K'(z) - 1} = \pi_0 \frac{-1}{K'(1) - 1} = \frac{\pi_0}{1-\rho}, \end{aligned}$$

since $K(1) = \Pi(1) = 1$ and $K'(1) = \rho$. Rearranging proves the claim. □

The next result is given as an exercise.

Exercise 3.5.11. *Use differentiation and l'Hôpital's rule (twice) to show that*

$$\mathbb{E}(N) = \Pi'(1) = \frac{2\Pi(0)K'(1) + K''(1)}{2(1 - K'(1))}.$$

There is one final ingredient required to prove Pollaczek–Khintchine formula.

Lemma 3.5.12. *Suppose that $\rho = \lambda/\mu < 1$. Then,*

$$K''(1) = \lambda^2 \mathbb{E}(S^2) = \lambda^2\big(\sigma^2 + 1/\mu^2\big).$$

Proof. The distribution of the service time S is characterised by its (negative) mgf

$$\phi(r) \coloneqq \mathbb{E}(e^{-rS}) \quad \text{for} \quad r \in \mathbb{R}.$$

Differentiating twice and setting $r = 0$ gives

$$\phi''(0) = \lim_{r\downarrow 0} \tfrac{d^2}{dr^2}\mathbb{E}(e^{-rS}) = \lim_{r\downarrow 0} \mathbb{E}(S^2 e^{-rS}) = \mathbb{E}(S^2).$$

Let A be the typical number of customers arriving during the service of a single customer; so, $K(z) = \mathbb{E}(z^A)$. Then, $(A \mid S = t) \sim \text{Pois}(\lambda t)$, and hence

$$\begin{aligned} K(z) = \mathbb{E}(z^A) &= \mathbb{E}(\mathbb{E}(z^A \mid S)) = \mathbb{E}(\mathbb{E}(z^{\text{Pois}(\lambda S)} \mid S)) \\ &= \mathbb{E}\big(\textstyle\sum_{m\geq 0} e^{-\lambda S}(\lambda S)^m z^m/m!\big) \\ &= \mathbb{E}\big(e^{-\lambda S}\textstyle\sum_{m\geq 0}(z\lambda S)^m/m!\big) \\ &= \mathbb{E}(e^{-\lambda S}e^{z\lambda S}) = \mathbb{E}(e^{-(1-z)\lambda S}) = \rho\big((1-z)\lambda\big). \end{aligned}$$

Consequently,

$$K''(1) = \lambda^2\, \phi''\big((1-z)\lambda\big)\big|_{z=1} = \lambda^2\phi''(0) = \lambda^2\mathbb{E}(S^2). \qquad \square$$

We can now prove the Pollaczek–Khintchine formula.

Proof of Theorem 3.5.9. Using the previous three results,

$$\mathbb{E}(N) \overset{3.5.11}{=} \frac{\Pi(0)K'(1)}{1-K'(1)} + \frac{K''(1)}{2(1-K'(1))} \overset{3.5.10,\,3.5.12}{=} \rho + \frac{\lambda^2\mathbb{E}(S^2)}{2(1-\rho)}.$$

By Little's law (Theorem 3.5.8) and the fact that $\rho = \lambda/\mu = \lambda\mathbb{E}(S)$, we deduce that

$$\mathbb{E}(T_{\mathsf{s}}) = \frac{1}{\lambda}\mathbb{E}(N) = \mathbb{E}(S) + \frac{\lambda(\mathbb{V}\text{ar}(S) + \mathbb{E}(S)^2)}{2(1-\rho)}.$$

The result follows from the fact that $T_{\mathsf{s}} = S + T_{\mathsf{q}}$. $\square$

Example 3.5.13. Jobs arrive at a computer's central processing unit (CPU) according to a Poisson process of rate $\lambda = \frac{1}{2}$. The CPU serves at an average of one unit of time per job. If the service times are Exponential—ie, we are in an $M/M/1$-queue set-up—then $\rho = \frac{1}{2}$ and the expected queueing time is

$$\mathbb{E}(T_{\mathsf{q}}) = \lambda/(1-\rho) = 1.$$

On the other hand, if the service times are not Exponential and have variance σ^2—for reference, $\mathbb{V}\text{ar}(\mathsf{Exp}(1)) = 1$—then we still have $\rho = \lambda = \frac{1}{2}$, but now

$$\mathbb{E}(T_{\mathsf{q}}) = \tfrac{1}{2}(1+\sigma^2).$$

Reducing the variance thus reduces the expected queueing time. $\triangle$

Using (negative) mgfs, also known as *Laplace transforms*, we can find statistics such as the expectation or variance of the length of a *busy period* in an $M/G/1$ queue.

Definition 3.5.14 (Busy Period)**.** Suppose that a customer arrives at an empty queue. The *busy period* is the interval of time from that customer's arrival until the next time the queue is empty. We denote the *length of the busy period* by B. $\triangle$

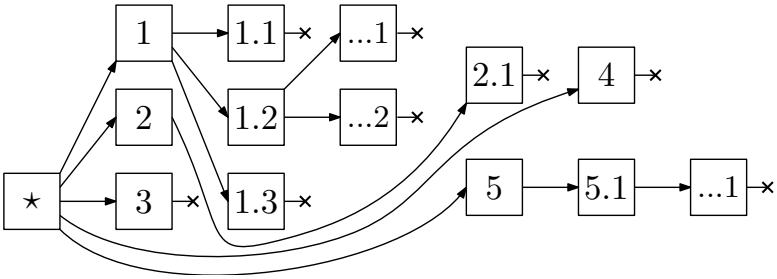

Figure 3.4. Depth-first search (lexicographic) for busy period:

- $\star$ served first and $(1, 2, 3, 4, 5)$ arrive during their service;
- 1 served next and $(1.1, 1.2, 1.3)$ arrive during their service;
- 1.1 served next and no-one arrives during their service;
- 2 (next lowest lexicographically) served next; etc.

We now find an expression for the law of B which is useful for evaluating the mgf. The motivation is "run until the first person is served and see what happens"—or, equivalently, "take one step of the embedded chain and see what happens".

Lemma 3.5.15. *Let S denote the service time of the customer who arrives at the empty queue. Then, conditional on $S = t$, we have*

$$B \sim t + B_1 + \ldots + B_A$$

where A is the number of customers arriving whilst the first customer is served and $B_1, B_2, \ldots \sim^{\text{iid}} B$, independently of A. Note that $(A \mid S = t) \sim \text{Pois}(\lambda t)$.

Proof. The key observation is that in order to determine the busy time B, it does not matter in which order the customers are served. All that matters is that *somebody* is being served whilst the system is non-empty. We use a different queue discipline.

We assume that each of the customers arriving during the first service starts a separate queue. Once the service of the first customer has been completed, move onto a customer who arrived during the service, iterating the process: any arrivals during the service of the second customer join the queue started by the second customer.

One the queue started by the second customer is emptied, which takes time $B_1 \sim B$, we proceed in the same way for another customer who arrived during the original service. The busy period ends once all these sub-busy periods have been handled. □

Remark. The previous proof uses a *depth-first search* (DFS); see Figure 3.4.

- If no customers arrive during the first busy period, then stop.
- If customers do arrive, then handle each of their busy periods in turn. △

We use this to find the (negative) mgf, then differentiate to find moments.

Lemma 3.5.16. *Let $\psi(r) \coloneqq \mathbb{E}(e^{-rB})$ and $\phi(r) \coloneqq \mathbb{E}(e^{-rS})$ for $r \in \mathbb{R}$. Then,*

$$\psi(r) = \psi\big(r + \lambda(1 - \phi(r))\big).$$

Proof. Using the previous lemma and the tower property, we have

$$\psi(r) = \mathbb{E}(\mathbb{E}(e^{-rB} \mid S)) = \mathbb{E}(e^{-rS}\mathbb{E}(\psi(r)^A \mid S)).$$

Now, $(A \mid S) \sim \text{Pois}(\lambda S)$, so $\mathbb{E}(\psi^A \mid S) = \exp(\lambda S(\psi - 1))$ for all $\psi \in \mathbb{R}$. Hence,

$$\psi(r) = \mathbb{E}\big(\exp\big(-rS + \lambda S(\psi(r) - 1)\big)\big) = \phi\big(r + \lambda(1 - \psi(r))\big). \qquad \square$$

This is only an implicit relation for $\psi(r)$, which is not easy to solve. However, we can find the expectation and variance of the busy-period length by differentiating.

Lemma 3.5.17. *Suppose that $\rho < 1$. Then,*

$$\mathbb{E}(B) = \mathbb{E}(S)/(1 - \rho) \quad \textit{and} \quad \mathbb{E}(B^2) = \mathbb{E}(S^2)/(1 - \rho)^3.$$

Proof. Derivatives of the (negative) mgf evaluated 0 correspond to moments:

$$\begin{aligned}
&\phi'(0) = -\mathbb{E}(S) \quad \text{and} \quad \phi''(0) = \mathbb{E}(S^2), \quad \text{so} \quad \mathbb{V}\text{ar}(S) = \phi''(0) - \phi'(0)^2;\\
&\psi'(0) = -\mathbb{E}(B) \quad \text{and} \quad \psi''(0) = \mathbb{E}(B^2), \quad \text{so} \quad \mathbb{V}\text{ar}(B) = \psi''(0) - \psi'(0)^2.
\end{aligned}$$

Differentiating the formula from the previous lemma gives the following:

$$\begin{aligned}
\psi'(r) &= \big(1 - \lambda\psi'(r)\big)\phi'\big(r + \lambda(1 - \psi(r))\big);\\
\psi''(r) &= \big(1 - \lambda\psi'(r)\big)^2\phi''\big(r + \lambda(1 - \psi(r))\big) - \lambda\psi''(r)\phi'\big(r + \lambda(1 - \psi(r))\big).
\end{aligned}$$

Plugging $r = 0$ into the first equation, using $\phi(0) = \psi(0) = 1$, we obtain

$$\mathbb{E}(B) = -\psi'(0) = -\big(1 - \lambda\psi'(0)\big)\phi'(0) = \big(1 + \lambda\mathbb{E}(B)\big)\mathbb{E}(S);$$

rearranging,

$$\mathbb{E}(B) = \frac{1}{1 - \lambda/\mu}\mathbb{E}(S) = \frac{1}{1 - \rho}\mathbb{E}(S) = \frac{1}{\lambda}\frac{\rho}{1 - \rho}.$$

Plugging all this into the second equation,

$$\mathbb{E}(B^2) = \psi''(0) = \big(1 + \lambda\mathbb{E}(B)\big)^2\mathbb{E}(S^2) + \lambda\mathbb{E}(B^2)\mathbb{E}(S);$$

rearranging,

$$\mathbb{E}(B^2) = \frac{(1 + \lambda\mathbb{E}(B))^2}{1 - \lambda\mathbb{E}(S)}\mathbb{E}(S^2) = \frac{1}{(1 - \rho)^3}\mathbb{E}(S^2). \qquad \square$$

The additive formulation of Lemma 3.5.15 actually gives the expectation directly.

Lemma 3.5.18. *Suppose that* $\rho := \lambda/\mu = \lambda\mathbb{E}(S) < 1$. *Then,*

$$\mathbb{E}(B) = \mathbb{E}(S)/(1-\rho).$$

Proof. Conditioning on A and using the previous formula,

$$\mathbb{E}(B \mid A) = \mathbb{E}(S \mid A) + A\mathbb{E}(B),$$

since $B_1, B_2, \ldots \sim^{\text{iid}} B$, independently of A. Taking expectation over A,

$$\mathbb{E}(B) = \mathbb{E}(\mathbb{E}(B \mid A)) = \mathbb{E}(S) + \mathbb{E}(A)\mathbb{E}(B).$$

Now, $(A \mid S = t) \sim \text{Pois}(\lambda t)$. So, $\mathbb{E}(A) = \mathbb{E}(\mathbb{E}(\text{Pois}(\lambda S) \mid S)) = \lambda\mathbb{E}(S)$. Hence,

$$\mathbb{E}(B) = \mathbb{E}(S)\big(1 + \lambda\mathbb{E}(B)\big), \quad \text{so} \quad \mathbb{E}(B) = \mathbb{E}(S)/(1-\rho). \qquad \square$$

The second moment can be found via a similar method. The variance can be calculated in a similar manner, using the Law of Total Variance.

Exercise 3.5.19. *Use the Law of Total Variance to find* $\mathbb{V}\text{ar}(B)$.

3.6* Jackson Networks—Advanced Topics for ST406

This section discusses *Jackson networks*: after passing through one queue, the customer joins another or leaves the system. These were alluded to at the end of §3.3, after Burke's theorem. The special case of *queues in tandem* was studied in §3.4. *Migration processes*, where the $M/M/1$ queues below are generalised, are studied in [KY14, Chapter 2]. This section can be seen as a warm-up for that chapter.

> **The material in this section (§3.6*) is only examinable for the ST406 fourth-year variant, not the ST333 third-year variant.**
>
> To emphasise, this section is not the *only* examinable content for the *ST406 Advanced Topics*; parts of [KY14, Chapter 2] are as well.

Burke's theorem (Theorem 3.3.1) says that the output of an $M/M/1$ queue is a Poisson process and the queue length at time t is independent of the departure process up to time t. We used this to analyse queues in tandem in §3.4. This could be extended to any directed-tree-like structure. It was a fairly fragile technique, though: a customer was not allowed to enter the same queue twice.

We develop a set of tools that does not give such fine-detail results as before, but can tolerate more general flow patterns. We focus on *Jackson networks*.

We start with an informal definition of a Jackson network. Consider a network of N single-server, Markovian queues. The arrival rate into queue i from outside the system is λ_i; the service rate of each queue is μ_i. Upon completion of service, each customer can either exit the system or move to another queue: the customer moves to queue j with probability $p_{i,j}$ and exists with probability $p_{i,0} = 1 - \sum_{j\in[N]} p_{i,j}$.

We now give the formal definition.

Definition 3.6.1. Let $N \in \mathbb{N}$. A *Jackson network* is a Markov chain on $\Omega = \mathbb{N}_0^N$; if $n \in \Omega$, then n_i denotes the number of customers in queue i. Let $e_i \in \Omega$ be defined by $e_{i,j} \coloneqq (e_i)_j \coloneqq \mathbf{1}\{i = j\}$—the i-th unit vector. The non-zero transition rates are

$$\begin{cases} q(n, n + e_i) = \lambda_i, \\ q(n, n + e_j - e_i) = \mu_i p_{i,j} \quad \text{if} \quad n_i \geq 1, \\ q(n, n - e_i) = \mu_i p_{i,0} \quad \text{if} \quad n_i \geq 1. \end{cases}$$

We assume that $p_{i,0} > 0$ and $p_{i,i} = 0$ for all $i \in [N]$. We also assume that $(\lambda_i)_{i\in[N]}$ and $(p_{i,j})_{i,j\in[N]}$ are such that the Markov chain is irreducible. △

What can be said about equilibrium for Jackson networks? The interaction between the queues destroys the independence that we had in the queues-in-tandem case. Nevertheless, we are going to see some surprisingly explicit and simple answers.

The key is to introduce the quantity $\bar{\lambda}_i$ ($i \in [N]$) which will be the *effective rate* at which customers enter queue i—that is, the rate from outside (ie, λ_i) plus the rate from the other queues. These will satisfied the so-called *traffic equations*.

Definition 3.6.2. A vector $\bar{\lambda} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_N) \in \mathbb{R}_{\geq 0}^N$ satisfy the *traffic equations* if

$$\bar{\lambda}_i = \lambda_i + \sum_{j:j\neq i} \bar{\lambda}_j p_{j,i} \quad \text{for all} \quad i \in [N]. \qquad \triangle$$

We make this guess based on Burke's theorem: the effective output rate of a queue should be the same as the effective input rate, in equilibrium. Thus, if $(\bar{\lambda}_i)_{i\in[N]}$ are the effective input rates, then they should satisfy the traffic equations.

Importantly, there exists a unique solution to the traffic equations.

Lemma 3.6.3. *There exists a unique solution to the traffic equations.*

Proof. We prove existence first. The matrix $P \coloneqq (p_{i,j})_{i,j\in[N]}$ with $p_{0,0} \coloneqq 1$ defines a stochastic matrix on $\{0, \ldots, N\}$. The corresponding discrete-time Markov chain Z is eventually absorbed at 0; so, the number V_i of visits to i by Z satisfies $\mathbb{E}(V_i) < \infty$.

Impose the initial law $\mathbb{P}\{Z_0 = i\} = \lambda_i/\lambda$ for $i \geq 1$, where $\lambda \coloneqq \sum_{i\geq 1} \lambda_i$. Then,

$$\mathbb{E}(V_i) = \mathbb{P}\{Z_0 = i\} + \sum_{n\geq 0} \mathbb{P}\{Z_{n+1} = i\}$$

$$
\begin{aligned}
&= \lambda_i/\lambda + \textstyle\sum_{n\geq 0} \sum_{j\in[N]} \mathbb{P}\{Z_n = j,\ Z_{n+1} = i\} \\
&= \lambda_i/\lambda + \textstyle\sum_{j\in[N]} \sum_{n\geq 0} \mathbb{P}\{Z_n = j\} p_{j,i} \\
&= \lambda_i/\lambda + \textstyle\sum_{j\in[N]} \mathbb{E}(V_j) p_{j,i}.
\end{aligned}
$$

So, multiplying through by $\lambda > 0$, if $\bar{\lambda}_i = \lambda \mathbb{E}(V_i)$, then the traffic equations are solved:

$$
\bar{\lambda}_i = \lambda_i + \textstyle\sum_{j\in[N]} \bar{\lambda}_j p_{j,i} \quad \text{for all} \quad i \in [N].
$$

We now prove uniqueness. Let λ' be a solution to the traffic equations:

$$
\lambda'_i = \lambda_i + \textstyle\sum_{j\in[N]} \lambda'_j p_{j,i} \quad \text{for all} \quad i \in [N].
$$

We need to show that $\lambda' = \bar{\lambda}$. To this end, let $\Delta \coloneqq \lambda'_i - \bar{\lambda}_i$ for $i \in [N]$: we show that $\Delta = 0$. Cancelling the λ_i term, we need to show that the only solution to

$$
\Delta_i = \textstyle\sum_{j\in[N]} \Delta_j p_{j,i} \quad \text{for all} \quad i \in [N] \quad \text{is} \quad \Delta = 0.
$$

But, $p_{i,0} > 0$ for all i, so the matrix $P_{\setminus 0} \coloneqq (p_{i,j})_{i,j\geq 1}$ is *sub*-stochastic: its row-sums are strictly less than 1. Thus, the operator norm[1] $\|P_{\setminus 0}\| < 1$. Hence, if $\Delta \neq 0$, then

$$
\|\Delta\| = \|\Delta P_{\setminus 0}\| \leq \|\Delta\| \|P_{\setminus 0}\|_{\mathsf{op}} < \|\Delta\|,
$$

a contradiction. Hence, $\Delta = 0$, so the traffic equations have a unique solution. □

We now come to the main theorem of this section. It frequently appears in lists of the most useful mathematical results for industry.

Theorem 3.6.4 (Jackson's Theorem, 1957). *Assume that the traffic equations have a solution $(\bar{\lambda}_i)_{i\geq 1}$ such that $\bar{\lambda}_i < \mu_i$ for all $i \geq 1$. Define $\bar{\rho}_i \coloneqq \bar{\lambda}_i/\mu_i$ for $i \geq 1$—the effective utilisation. Then, the Jackson network has invariant distribution π given by*

$$
\pi(n) \coloneqq \textstyle\prod_{i\geq 1} (1 - \bar{\rho}_i) \bar{\rho}_i^{n_i} \quad \textit{for} \quad n \in \Omega.
$$

So, the queue lengths in equilibrium are independent and Geometrically distributed.

This theorem was proved relatively recently—1957, compared with Erlang whose studies, including $M/M/K$ queues, took place in the early 1900s. There are two likely reasons. One is that the system is not reversible. Indeed, there is no assumption that $p_{i,j} > 0$ if and only if $p_{j,i} > 0$, or even that $\lambda_i > 0$. This always makes computations

[1] the *operator norm* is a norm on linear operators A:

$$
\|A\|_{\mathsf{op}} \coloneqq \inf\{\lambda \geq 0 \mid \|Av\| \leq \lambda \|v\| \text{ for all } v \in V\};
$$

when A is a square matrix, it is equal to the modulus of the largest eigenvalue in modulus

vastly more complicated, a priori. When reversibility does not hold, typically, a distribution is proposed and checked, rather than solving $\pi Q = 0$ to find π. Second, it is a pretty bold proposal that queue lengths are independent at equilibrium!

We are going to see that there is a partial form of reversibility.

Definition 3.6.5 (Partial Balance)**.** A measure π and a matrix Q on a state space Ω are in *partial balance* if, for all $x \in \Omega$, we can find a partition of $\Omega \setminus \{x\}$, say into $\Omega_1^x, \Omega_2^x, \ldots$, such that

$$\sum_{y \in S_i^x} \pi(x) q(x,y) = \sum_{y \in S_i^x} \pi(y) q(y,x) \quad \text{for all} \quad i \geq 1. \qquad \triangle$$

Remark. *Global* balance means that the total probability flux into and out of a state is the same. *Detailed* balance requires equal flux between any pair of states. *Partial* balance means that, for each state, there is a subset of the states for which the total flux between that state and the subset is equal in each direction. $\triangle$

Partial balance implies global balance, as the name suggests.

Exercise 3.6.6. *If π and Q are in partial balance, then they are in global balance.*

We are going to show that partial balance holds for the Jackson network. This includes choosing appropriate partitions. Ignoring the outside, transitions occur between queues j and k with $j \neq k$. The checking global balance requires summing over all pairs $(j,k) \in [N]^2$ with $j \neq k$. In essence, we are going to fix a queue j and sum over all k. This will show that the flux in and out of a given queue is equal.

Proof of Theorem 3.6.4. Let us define $\pi(n) \coloneqq \prod_{i \geq 1} \bar{\rho}_i^{n_i}$ for $n \in \Omega$; this is a constant multiple off what is in the theorem. Let us then define

$$\hat{q}(n,m) \coloneqq \tfrac{\pi(m)}{\pi(n)} q(m,n) \quad \text{for} \quad n, m \in \Omega;$$

these are the transitions for the time-reversal of the Jackson network.

We now choose the partitions for partial balance. Let

$$\mathcal{A} \coloneqq \{e_i \mid i \in [N]\};$$

thus, if $n \in \Omega$ and $m \in \mathcal{A}$, then $n + m$ denotes any possible state after the arrival of a customer from outside to some queue. Let

$$\mathcal{D}_j \coloneqq \{e_i - e_j \mid i \in [N] \setminus \{j\}\} \cup \{-e_j\} \quad \text{for} \quad j \in [N];$$

thus, if $n \in \Omega$ and $m \in \mathcal{D}_j$, then $n + m$ denotes any possible state after the departure of a customer from queue j. We show, for all $n \in \Omega$, that

$$\sum_{m \in \mathcal{E}} q(n, n+m) = \sum_{m \in \mathcal{E}} \hat{q}(n, n+m) \quad \text{if} \quad \mathcal{E} = \mathcal{A} \text{ or } \mathcal{E} = D_j \text{ for some } j. \qquad (\star)$$

These are a set of partial balance equations, so this implies that π is invariant.

We first handle arrivals: $\mathcal{E} = \mathcal{A}$. The left-hand side of ($\star$) is

$$\textstyle\sum_{m\in\mathcal{A}} q(n, n+m) = \sum_{i\geq 1} \lambda_i.$$

To evaluate the right-hand side of ($\star$), we first calculate

$$\hat{q}(n, n+e_i) = \tfrac{\pi(n+e_i)}{\pi(n)} \cdot q(n+e_i, n) = \bar{\rho}_i \cdot \mu_i p_{i,0} = \bar{\lambda}_i p_{i,0}.$$

Plugging this into the right-hand side of ($\star$), we get

$$\begin{aligned}\textstyle\sum_{m\in\mathcal{A}} \hat{q}(n, n+m) &= \textstyle\sum_{i\geq 1} \bar{\lambda}_i p_{i,0} = \sum_{i\geq 1} \bar{\lambda}_i \big(1 - \sum_{j\geq 1} p_{i,j}\big) \\ &= \textstyle\sum_{i\geq 1} \bar{\lambda}_i - \sum_{j\geq 1} \sum_{i\geq 1} \bar{\lambda}_i p_{i,j} = \sum_{i\geq 1} \big(\bar{\lambda}_i - \sum_{j\geq 1} \bar{\lambda}_j p_{j,i}\big),\end{aligned}$$

swapping the i–j indices in the double sum to get $\sum_{i,j\geq 1} \bar{\lambda}_i p_{i,j} = \sum_{i,j\geq 1} \bar{\lambda}_j p_{j,i}$. We now apply the traffic equations (Definition 3.6.2) to this last sum over j to get

$$\textstyle\sum_{m\in\mathcal{A}} \hat{q}(n, n+m) = \sum_{i\geq 1} \big(\bar{\lambda}_i - (\bar{\lambda}_i - \lambda_i)\big) = \sum_{j\geq 1} \lambda_j.$$

We now turn to departures: $\mathcal{E} = \mathcal{D}_j$ for arbitrary $j \in [N]$. We have

$$q(n, n+m) = \mu_j p_{j,0} \quad \text{if} \quad m = -e_j \quad \text{and} \quad q(n, n+m) = \mu_j p_{j,i} \quad \text{if} \quad m = e_i - e_j.$$

The left-hand side of ($\star$) is

$$\textstyle\sum_{m\in\mathcal{D}_j} q(n, n+m) = \mu_j p_{j,0} + \sum_{i:i\neq j} \mu_j p_{j,i} = \mu_j;$$

this makes sense as the service rate at queue j is μ_j. Now,

$$\hat{q}(n, n+e_i-e_j) = \tfrac{\pi(n+e_i-e_j)}{\pi(n)} \cdot q(n+e_i-e_j, n) = \tfrac{\bar{\rho}_i}{\bar{\rho}_j} \cdot \mu_i p_{i,j} = \mu_j \bar{\lambda}_i p_{i,j} / \bar{\lambda}_j$$

and

$$\hat{q}(n, n-e_j) = \tfrac{\pi(n-e_j)}{\pi(n)} \cdot q(n-e_j, n) = \lambda_j / \bar{\rho}_j = \mu_j \lambda_j / \bar{\lambda}_j.$$

Using this and the traffic equations, we deduce that the right-hand side of ($\star$) is

$$\textstyle\sum_{m\in\mathcal{D}_j} \hat{q}(n, n+m) = \big(\lambda_j + \sum_{i:i\neq j} \bar{\lambda}_i p_{i,j}\big) \cdot \mu_j / \bar{\lambda}_j = \mu_j.$$

This completes the proof of partial balance and hence of invariance. □

In equilibrium, the departure process from an $M/M/1$ queue is a Poisson process, by Burke's theorem. This used reversibility. A Jackson network is not reversible, however its time-reversal is also a Jackson network. We use this to show that the departures to outside—those governed by $p_{i,0}$—form independent Poisson processes.

Corollary 3.6.7. *Consider a Jackson network with arrival rates $(\lambda_i)_{i\geq 1}$, service rates $(\mu_i)_{i\geq 1}$ and transition probabilities $(p_{i,j})_{i,j\geq 1}$:*

$$q(n, n+e_i) = \lambda_i, \qquad q(n, n+e_j-e_i) = \mu_i p_{i,j}, \qquad q(n, n-e_i) = \mu_i p_{i,0}.$$

Its time-reversal is itself a Jackson network with arrival rates $(\hat{\lambda}_i := \bar{\lambda}_i p_{i,0})_{i\geq 1}$, service rates $(\hat{\mu}_i := \mu_i)_{i\geq 1}$, transition probabilities $(\hat{p}_{i,j} := \bar{\lambda}_j p_{j,i}/\bar{\lambda}_i)_{i,j\geq 1}$ and exit probabilities $(\hat{p}_{i,0} := 1 - \sum_{j\geq 1} \hat{p}_{i,j} = \lambda_i/\bar{\lambda}_i)_{i\geq 1}$:

$$\hat{q}(n, n+e_i) = \bar{\lambda}_i p_{i,0}, \quad \hat{q}(n, n+e_j-e_i) = \mu_i \bar{\lambda}_j p_{j,i}/\bar{\lambda}_i, \quad \hat{q}(n, n-e_i) = \mu_i \lambda_i/\bar{\lambda}_i.$$

Moreover, (λ, p) and $(\hat{\lambda}, \hat{p})$ satisfy the same traffic equations.

Proof. We need to determine the time-reversed rates

$$\hat{q}(n,m) := \tfrac{\pi(m)}{\pi(n)} q(m,n) \quad \text{for} \quad m, n \in \Omega.$$

But, we already did this in the previous proof, obtaining the required expressions. The fact that $\hat{p}_{i,0} = 1 - \sum_{j\geq 1} \hat{p}_{i,j} = \lambda_i/\bar{\lambda}_i \in [0,1]$ implies that the $\hat{p}_{i,j}$-s are probabilities.

Departures become arrivals in reverse-time and vice versa. The final part follows as the effective arrival rate equals the effective departure rate in equilibrium. □

Corollary 3.6.8. *At equilibrium, the processes of departures (to outside) form independent Poisson processes; the rate of departures from queue i is $\bar{\lambda}_i p_{i,0}$. Further, the state of the Jackson network at time t is independent of the departures up to time t.*

Proof. The departure process (to outside) from queue i for the Jackson network X is the arrival process to queue i (from outside) in the time-reversal $\hat{X}$. But, as we just showed, these are independent Poisson processes, the i-th with rate $\bar{\lambda}_i p_{i,0}$. Hence, the departures form independent Poisson processes of the claimed rates.

The final independence follows analogously to that in Burke's theorem. Indeed, the length of the queues at time 0 is independent of the arrival process between times 0 and t. By time reversal, X_t is independent of departures up to time t. □

Remark. These last two corollaries have intuitive justifications, too, via time-reversal arguments and the fact that $\bar{\lambda}$ is the vector of effective arrival rates.

- In equilibrium, the rate at which customers arrive at a particular queue must be the same as the rate at which they depart. Arrivals become departures when time is reversed, and vice versa. So, the effective arrival/departure rate is $\bar{\lambda}_k$ for queue k both in forward- and reverse-time.
- An external arrival in reverse-time is an external departure in forward-time. So, the external-arrival rate in reverse-time is the effective arrival rate times the departure probability in forward-time: $\hat{\lambda}_i = \bar{\lambda}_i p_{i,0}$. Analogously, $\lambda_i = \bar{\lambda}_i \hat{p}_{i,0}$.

- A transition from queue i to queue j in reverse-time is a transition from j to i in forward-time. So, a j-to-i transition happens at rate $\bar{\lambda}_j p_{j,i}$ in forward-time. This must be balanced in reverse-time: $\bar{\lambda}_i \hat{p}_{i,j} = \bar{\lambda}_j p_{j,i}$.
- The service rate at queue i is the reciprocal of the mean time between when a customer starts and finishes being served. Start and finish swap roles in forward- versus reverse-time. So, the service rate is unchanged: $\hat{\mu}_i = \mu_i$. $\triangle$

4 Epidemic Models

An epidemic is a widespread occurrence of an infectious disease in a community.

The cycle of events for an individual in an epidemic can be broken down as follows.

1. Start off as *susceptible*.
2. Be exposed to infection: become *infected* and *infectious*.
3. Exhibit symptoms: become *symptomatic*.
4. Be *removed* from the system: become immune, or die.

The intervals between events have the following names.

- Between exposure to infection and becoming infected: *latent period*
- Between exposure to infection and exhibiting symptoms: *incubation period*
- Between becoming infectious and removal: *infectious period*

Infectious individuals are sometimes called *infectives*, and susceptible *susceptibles*.

We consider a number of simplifications.

- Infections arise from contact with other infectives and the whole population mixes homogeneously: any individual is equally likely to interact with any other
- An individual becomes infectious and symptomatic *immediately* after infection
- Removed individuals cannot become susceptible again

Based on these simplifications, we consider three different models.

§4.1 Deterministic without removals: ***S****usceptible* → ***I****nfected*

§4.2 Stochastic without removals: ***S****usceptible* → ***I****nfected*

§4.3 Stochastic with removals: ***S****usceptible* → ***I****nfected* → ***R****emoved*

These simplifications and models are, of course, far from the truth for any real epidemic. However, the theory has to start somewhere! Dealing with simplified situations provides insights so long as one remembers to be thoughtful about their application.

The classic text on epidemics is the book by Bailey [Bai75].

4.1 Deterministic SI Model

Deterministic models are used to model population averages. They are easier to implement and can be useful for getting an overall picture. However, they are often not as accurate as stochastic models, particularly in the crucial starting stages. For example, stochastic models often have some probability that an epidemic actually gets going; deterministic models, naturally, have no such probabilistic events.

We start with a simple example in order to build some intuition. Suppose that the infectious period extends indefinitely—there are no removals—and that the overall population size n is large, with just one initial infective. Let X_t and Y_t denote the number of susceptible and infectives at time t, respectively; so, $Y_0 = 1$ and $X_t + Y_t = n$.

Example 4.1.1. Under the assumption of homogeneous mixing, we expect the number of newly infected individuals in a short interval of length Δt to be given by

$$\Delta X_t = -\alpha X_t Y_t \Delta t$$

where α is the rate at which individuals come in contact with each other. Indeed, each of the Y_t infectives is mixing with each of the X_t susceptibles, giving rise to $X_t \cdot Y_t$ possible pairs of contacts; the mixing rate is α, so a proportion $\alpha \Delta t$ interact. Thus,

$$\tfrac{d}{dt} X_t = -\alpha X_t Y_t = -\alpha X_t (n - X_t) \quad \text{with} \quad X_0 = n - 1,$$

by taking $\Delta t \to 0$. Solving this DE, using partial fractions, gives

$$X_t = \frac{n(n-1)}{n - 1 + e^{\alpha n t}} \quad \text{and} \quad Y_t = n - X_t = \frac{n e^{\alpha n t}}{n - 1 + e^{\alpha n t}} = \frac{n}{(n-1)e^{-\alpha n t} + 1}.$$

So, $X_t \to 0$ as $t \to \infty$. Thus, eventually, the whole population will be infected.

It is often informative to look at the *epidemic curve*:

$$\frac{dY_t}{dt} = -\frac{dX_t}{dt} = \alpha X_t Y_t = \frac{n^2 (n-1) e^{\alpha n t}}{(n - 1 + e^{\alpha n t})^2}.$$

This models the *rate* of infections. Figure 4.1 shows two epidemic curves—one with $n = 40$ and one with $n = 60$—with $\alpha = 1$.

- The infection rate gets higher as the population size n grows.
- The time until almost all the population is infected *decreases* as n increases.

Can you explain why these hold in words? △

One obvious limitation of this model is that X_t is modelled as a positive real, tending to 0 as $t \to \infty$—but never equal to 0. In reality, $X_t \in \mathbb{N} = \{0, 1, 2, \ldots\}$. To address this issue, we next model X_t as a random process, taking values in $\mathbb{N}$.

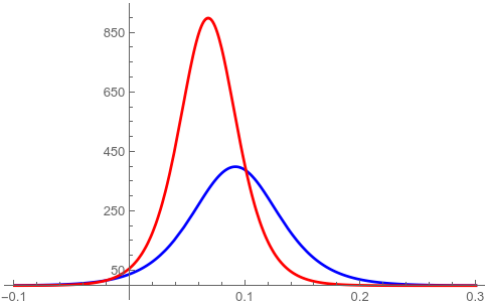

Figure 4.1. The epidemic curve for a rate-1, deterministic epidemic with no removals, starting from a single infective: blue curve has population size $n = 40$; red curve has $n = 60$

4.2 Stochastic SI Model

Let X_t and Y_t be the number of susceptibles and infectives at time t, respectively. We assume there are no removals. Let $n = X_t + Y_t$ be the total population size.

There are $X_t Y_t$ possible pairs of contacts at a given time t. The deterministic model said that a *proportion* $\alpha\Delta t$ of these interact in a (short) interval of length Δt. The stochastic model says that each interacts *with probability* $\alpha\Delta t$, independently. Thus, a proportion $\alpha\Delta t$ interact *on average* in the stochastic model.

Definition 4.2.1 (SI Model)**.** Let $Y = (Y_t)_{t\geq 0}$ be the (non-linear) birth process with

$$q_{i,i+1} = \begin{cases} \alpha i(n-i) & \text{if} \quad 0 < i < n, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, it increases by 1 each time an individual gets infected, which happens at rate $\alpha X_t Y_t = \alpha Y_t(n - Y_t)$ when $X_t > 0$. This is homogeneous mixing at rate α. $\triangle$

Example 4.2.2. In the stochastic SI model with homogeneous mixing at rate α, what is the expected time until everyone becomes infected?—ie, what is

$$\mathbb{E}(\tau_1) \quad \text{where} \quad \tau_1 := \inf\{t \geq 0 \mid Y_t = n\}?$$

Without loss of generality, we take $\alpha = 1$, scaling time by α at the end. The jump from i to $i + 1$ takes time $\mathsf{Exp}(i(n - i))$, independent of everything else. So,

$$\tau_1 = T_1 + \ldots + T_{n-1} \quad \text{where} \quad T_i \sim \mathsf{Exp}\big(i(n-1)\big) \quad \text{independently.}$$

Hence,

$$\mathbb{E}(\tau_1) = \sum_{i=1}^{n-1} \mathbb{E}(T_i) = \sum_{i=1}^{n-1} \tfrac{1}{i(n-i)}.$$

Using partial fractions,

$$\mathbb{E}(\tau_1) = \tfrac{1}{n} \sum_{i=1}^{n-1} \big(\tfrac{1}{i} + \tfrac{1}{n-i}\big) = \tfrac{2}{n} \sum_{i=1}^{n-1} \tfrac{1}{i}.$$

Using the standard expression for partial sums of the harmonic series,

$$\mathbb{E}(\tau_1) = \tfrac{2}{n}\big(\log n + \gamma + \mathcal{O}(1/n)\big) \approx 2 \log n/n,$$

where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant. △

Exercise 4.2.3. *Compute the variance* $\mathbb{V}\mathrm{ar}(\tau_1)$ *of the total time until the epidemic runs its course. How does it behave as* $n \to \infty$*? Does* τ_1 *concentrate as* $n \to \infty$*?*

Example 4.2.4. What is the law of the *half life* of the epidemic?—ie, what is the law of

$$\tau_{1/2} := \inf\{t \geq 0 \mid Y_t/n > \tfrac{1}{2}\}?$$

As in the previous exercise, letting $m := \lfloor n/2 \rfloor$,

$$\tau_{1/2} = T_1 + \ldots + T_m \quad \text{where} \quad T_i \sim \mathsf{Exp}\big(i(n-i)\big) \quad \text{independently.}$$

Exercise 2.2.6 showed that the law of $\tau_{1/2}$ is given by

$$\mathbb{P}\{\tau_{1/2} > t\} = \mathbb{P}\{T_1 + \ldots + T_m > t\} = \sum_{i=1}^{m} e^{-\lambda_i t} \prod_{j:j\neq i} \tfrac{\lambda_j}{\lambda_j - \lambda_i}$$

where $\lambda_k := k(n-k)$. We also observe that, if n is even, then $m = n/2$ and

$$T_1 + \ldots + T_m =^d T_n + \ldots + T_{m+1} \quad \text{since} \quad \lambda_k = \lambda_{n-k}.$$

So, the lifetime τ_1 has the same law as the sum of two independent half-lifes $\tau_{1/2}$. △

4.3 Stochastic SIR Model

The final section of the course concerns the stochastic SIR model: infectious individuals are removed from the system—either by dying or becoming immune—autonomously (independent of everything else) at rate β; immunity is never lost.

There is a deterministic model, akin to that of §4.1 except with removals at rate βi if there are i infectives. We leave analysis of this deterministic model as an exercise.

Exercise 4.3.1. *Let* S_t*,* I_t *and* R_t *denote the number of susceptible, infected and removed individuals at time* t*, respectively. Let* n *denote the population size; so,* $n = S_t + I_t + R_t$ *for all* $t \geq 0$*. Derive and solve a* DE *in the deterministic case.*

We emphasise that immunity is indefinite: a removed individual *cannot* become susceptible or infected ever again. We now define the process precisely.

4.3.1 Definition and Markov Property

We start with the precise definition of the stochastic SIR dynamics.

Definition 4.3.2 (SIR Model)**.** Let S_t, I_t and R_t denote the number of susceptible, infected and removed individuals at time t, respectively. The dynamics are as follows:

- each infective *infects* each susceptible one at rate α independently;
- each infective is *removed* at rate β independently.

We denote such a model $\mathsf{SIR}(\alpha, \beta)$. Let n denote the total population size; so, $n = S_t + I_t + R_t$ for all $t \geq 0$. Such independent removal is called *autonomous* removal. $\triangle$

Exercise 4.3.3. *Argue that $(S_t, I_t, R_t)_{t\geq 0}$ is a Markov chain on $\mathbb{N}^3$ with non-zero rates*

$$(s,i,r) \to \begin{cases} (s-1, i+1, r) & \text{at rate} \quad \alpha si, \\ (s, i-1, r+1) & \text{at rate} \quad \beta i. \end{cases}$$

Give 'physical' interpretations for each of the transitions.

The rates in the previous exercise are often written in *interaction* form:

$$\begin{aligned} (S,\, I) &\to (S-1,\, I+1) \quad \text{at rate} \quad \alpha SI; \\ (I,\, R) &\to (I-1,\, R+1) \quad \text{at rate} \quad \beta I. \end{aligned}$$

The process $(I_t)_{t\geq 0}$ looks like a birth–death process: indeed, it increases/decreases by 1 in each step. However, it is not actually a Markov chain, considered alone. In order to know the rate at which it goes up/down, the number of susceptible and removed individuals is needed. Looking into the past gives an estimate on the previous jump rates, and hence number of susceptibles. Since $R_t = n - S_t - I_t$, we can reduce the number of degrees of freedom from three to two, so $(S_t, I_t)_{t\geq 0}$ is a Markov chain.

The main question we address in this section is that of relating $(I_t)_{t\geq 0}$ to a birth–death process in a useful way. Particularly, what can we say about the probability of an epidemic becoming widespread? Naturally, we do not expect $I_t = n$ for some $t \geq 0$ when starting with $I_0 = 1$: this would require all the other individuals to get infected before any are removed. So, for example, what is

$$\mathbb{P}_1\{\sup_{t\geq 0} I_t \geq \gamma n\} \quad \text{for} \quad \gamma > 0 \quad \text{independent of } n.$$

The subscript in $\mathbb{P}$ indicates I_0. We always start with no-one removed, so $S_0 + I_0 = n$.

We answer this by drawing a comparison with birth–death processes: we make a simultaneous 'microscopic' construction of a birth–death process and an epidemic with removals in such a way that we can gets bounds on the above probability.

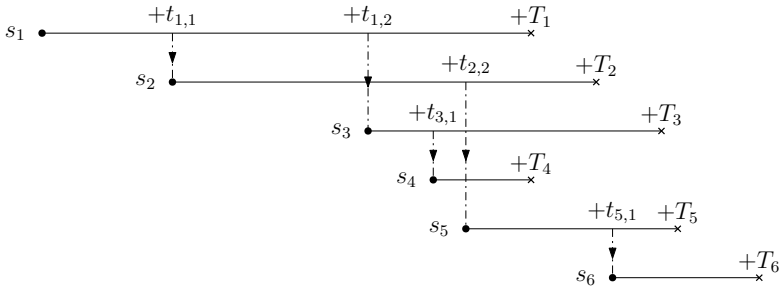

Figure 4.2. The microscopic construction of birth–death process

4.3.2 Microscopic Construction of a Birth–Death Process

We use the following construction of a birth–death process.

Definition 4.3.4 (Microscopic Birth–Death)**.** Let X_t and Y_t denote the number of infected and removed individuals at time t, respectively.

Suppose that we have an infinite population, indexed by $\mathbb{N} = \{1, 2, \ldots\}$. Let $(t_{i,n})_{n\geq 1}$ be a sequence of iid $\mathsf{Exp}(\alpha)$-s and $T_i \sim \mathsf{Exp}(\beta)$ independently, for each $i \in \mathbb{N}$ independently. The sequence $(t_{i,n})_{n\geq 1}$ controls the infections that individual i makes, whilst T_i is the length of its infectious period.

We represent collections of individuals as subsets of $\mathbb{N} = \{1, 2, \ldots\}$. Let $\{1, \ldots, i_0\}$ represent the set of originally-infected individuals and $\{i_0 + 1, \ldots\}$ represent the remainder, which are initially susceptible. Infections and removals follow these rules.

- If individual i becomes infected at time s_i, then remove i at time $s_i + T_i$. So, s_i is the *infection time* and $s_i + T_i$ the *removal time* for i.
- At times $\tau_{i,j} = s_i + \sum_{n=1}^{j} t_{i,n}$ between the infection and removal times—ie, in $[s_i, s_i + T_i)$—use individual i to infect the individual ℓ with the lowest number who is susceptible at time $\tau_{i,j}$. Then, $s_\ell = \tau_{i,j} = s_i + \sum_{n=1}^{j} t_{i,n}$.

Set $s_i := 0$ for $i \leq i_0$, corresponding to individuals who are initially infected. △

The above construction is illustrated in Figure 4.2. We now show that this construction does indeed give a birth–death process with birth rate α and death rate β.

Theorem 4.3.5. *The process $X = (X_t)_{t\geq 0}$ from Definition 4.3.4 is an $\mathsf{SBDP}(\alpha, \beta)$:*

$$q_{n,n+1} = \alpha n, \quad q_{n,n-1} = \beta n \quad \text{and} \quad q_n = (\alpha + \beta)n \quad \text{for all} \quad n.$$

Proof. We need to show that the times between jumps are independent and Exponentially distributed, with rates which depend only on the current state, not the past. The memoryless property is crucial in ensuring this in the statements below.

Infections. By the memoryless property, each currently-infected individual is waiting an $\mathsf{Exp}(\alpha)$ time until it causes an infection. So, the next infection is at the minimum of these $\mathsf{Exp}(\alpha)$-s, which is distributed as $\mathsf{Exp}(\alpha n)$ if there are n currently-infected individuals. Hence, the time-t infection rate is αn if $X_t = n$.

Removals. Similarly, the time until removal for each infected individual is still $\mathsf{Exp}(\beta)$. So, the next removal is at the minimum of these $\mathsf{Exp}(\beta)$-s. Again, minimum of n $\mathsf{Exp}(\beta)$-s is $\mathsf{Exp}(\beta n)$. So, the time-t removal rate is βn if $X_t = n$. □

4.3.3 Microscopic Construction of the SIR Model

We now construct the SIR model in an analogous manner. We have to reduce the infection rate as the epidemic progresses, to take account of there being fewer susceptibles. We attach to each infection incident a 'censoring' random variable to do this.

Definition 4.3.6 (Microscopic SIR)**.** Let S_t, I_t and R_t denote the number of susceptible, infected and removed individuals at time t, respectively.

Suppose that we have a population of size n, indexed by $\{1, \ldots, n\}$. Let $(t_{i,n})_{n\geq 1}$ be a sequence of iid $\mathsf{Exp}(\alpha)$-s, $(U_{i,j})_{j\geq 1}$ be a sequence of iid $\mathrm{Unif}([0,1])$-s and $T_i \sim \mathsf{Exp}(\beta)$ independently, for each i independently. The sequence $(t_{i,n})_{n\geq 1}$ controls the infections that individual i makes, whilst T_i is the length of its infectious period.

We represent collections of individuals as subsets of $\mathbb{N} = \{1, 2, \ldots\}$. Let $\{1, \ldots, i_0\}$ represent the set of originally-infected individuals and $\{i_0 + 1, \ldots\}$ represent the remainder, which are initially susceptible. Infections and removals follow these rules.

- If individual i becomes infected at time s_i, then remove i at time $s_i + T_i$.
- At times $\tau_{i,j} = s_i + \sum_{n=1}^{j} t_{i,n} \in [s_i, s_i + T_i)$, use individual i to *attempt* to infect the individual ℓ with the lowest number who is susceptible at time $\tau_{i,j}$. This particular infection attempt is successful if and only if $(n - i_0)U_{i,j} \leq S_{\tau_{i,j}}$. This last event has probability $S_{\tau_{i,j}}/(n - i_0)$ conditional on $(S_u, I_u, R_u)_{u\leq \tau_{i,j}}$.

Set $s_i := 0$ for $i \leq i_0$, corresponding to individuals who are initially infected. △

The above construction is illustrated in Figure 4.3. We now show that this construction does indeed give an SIR process with infection rate $\frac{\alpha}{n-i_0}$ and death rate β.

Theorem 4.3.7. *The process $(S_t, I_t, R_t)_{t\geq 0}$ from Definition 4.3.6 is an $\mathsf{SIR}(\frac{\alpha}{n-i_0}, \beta)$:*

$$\begin{aligned} (S,\, I) &\to (S-1,\, I+1) \quad \text{at rate} \quad \alpha SI/(n-i_0); \\ (I,\, R) &\to (I-1,\, R+1) \quad \text{at rate} \quad \beta I. \end{aligned}$$

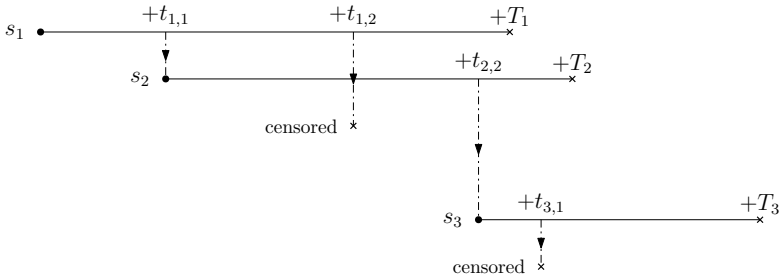

Figure 4.3. The microscopic construction of an SIR process. Notice how the same random variables are used: eg, individual 3 has the same lifetime (T_3) in both cases. It would be tempting to simply delete the lines, and their descendants, from BD corresponding to censored infection attempts to obtain SIR. But, this would mean that 3 has lifetime T_5 in SIR. This would be a legitimate construction, by iid nature of the random variables, but it is not what we do. This is crucial for monotonicity later

Proof. They key difference here compared with the last proof is that some contacts are censored. The censoring is done by a Unif$([0,1])$ random variable, which is independent of everything else. This does not affect the Exp(β)-length removal times.

Without censoring, the infection time is the first arrival of a collection of independent Poisson processes. The censoring simply thins the Poisson processes; but, we know that this simply leads to another Poisson process. Thus, the new infection rate is α times the conditional probability $S_t/(n-i_0)$ that the Poisson incident is kept. This thinning probability depends only on the current state S_t, not on the history.

Alternatively, looking at the rates as derivatives, we see that the probability of having an infection in the SIR model by time δ is equal, up to an $o(\delta)$ error, to that in the BD model multiplied by the probability that the infection is accepted. Hence,

$$q^{\mathsf{SIR}}_{(s,i),(s-1,i+1)} = q^{\mathsf{BD}}_{i,i+1} \cdot \mathbb{P}_{(s,i)}\{\text{accept infection}\} = \alpha i s/(n-i_0). \qquad \square$$

Remark. We emphasise that having a state-dependent probability of accepting an infection does not break the Markov property. If the probability depends on the history—eg, if a second attempt by the same individual were more likely to succeed than the first, all else the same—then the Markov property would clearly be broken.

This can be thought of as a 'thinning of Poisson process of infections'. Such a viewpoint can be helpful for intuition, but care must be taken: the thinning probability changes, and in a way which depends on the process. $\triangle$

4.3.4 Coupling

We have now seen how to build an epidemic and a birth–death process in an essentially analogous way. There is nothing stopping us from using the same random variables to construct both *simultaneously*. The only thing that is different is that the epidemic requires the extra uniform random variables to thin the infections.

We can use this single source of randomness to prove results about the sizes of the two processes. A similar approach was used in Example Sheet 2 to compare Poisson processes. We consider basic birth–death models as a warm-up.

Example 4.3.8. Let $X^\lambda = (X_t^\lambda)_{t\geq 0} \sim \mathsf{SRW}_{\mathbb{N}}(\lambda, 1)$ on $\mathbb{N} = \{0, 1, ...\}$ with $\lambda > 0$:

$$q^\lambda_{x,x+1} = \lambda \quad \text{and} \quad q^\lambda_{x+1,x} = 1 \quad \text{for} \quad x \geq 0.$$

We construct couplings which establishes the following monotone properties:

$$\lambda \mapsto \mathbb{P}_x\{X_t^\lambda = 0\} \quad \text{is weakly decreasing} \quad \text{for all} \quad x \geq 0.$$
$$x \mapsto \mathbb{P}_x\{X_t^\lambda = 0\} \quad \text{is weakly decreasing} \quad \text{for all} \quad \lambda > 0.$$

Fix $\lambda \geq \mu > 0$. Let $X \coloneqq X^\lambda$ and $Y \coloneqq X^\mu$. Let $A \sim \mathsf{PP}(\lambda)$. Define B by thinning the Poisson process A, keeping with probability $\mu/\lambda < 1$. Then, $B \sim \mathsf{PP}(\mu)$—it *is not* independent of its 'parent' process A. Let $D \sim \mathsf{PP}(1)$. Use the following dynamics:

- on incidents of D, step both X and Y down by 1, with a barrier at 0;
- on incidents of A, step X up by 1; on incidents of B, step Y up by 1.

These dynamics have $X \sim \mathsf{SRW}_{\mathbb{N}}(\lambda, 1)$ and $Y \sim \mathsf{SRW}_{\mathbb{N}}(\mu, 1)$—not independently.

The incidents of B are a subset of those of A. Thus, if Y moves up, then so does X. The two move down together, unless one is already at 0, the lowest point. This means that if Y starts below X, then there is no way that it can ever get above: if $X_0 \geq Y_0$, then $X_t \geq Y_t$ for all $t \geq 0$. In particular, $X_t = 0$ implies $Y_t = 0$. Thus,

$$\mathbb{P}_1\{X_t = 0\} \leq \mathbb{P}_1\{Y_t = 0\}.$$

For the second statement, we move $X, Y \sim \mathsf{SRW}_{\mathbb{N}}(\lambda, 1)$ together: if one is at 0 and the other is not, then an attempted move down by the one at 0 is censored. This way, if $X_0 \geq Y_0$, then $X_t \geq Y_t$ for all $t \geq 0$. In particular, $X_t = 0$ implies $Y_t = 0$. △

Exercise 4.3.9 (Coupling for $\mathsf{SBDP}(\lambda, \mu)$)**.** *We showed in Corollary 2.3.4 that*

$$\lim_{t\to\infty} \mathbb{P}_1\{X_t = 0\} = \begin{cases} 1 & \text{if} \quad \mu \geq \lambda, \\ \mu/\lambda & \text{if} \quad \mu < \lambda. \end{cases}$$

We proved this for $\lambda \neq \mu$ by using the (negative) mgf from Lemma 2.3.3.

Use a monotone coupling to deduce that $\lambda = \mu$ case from the $\lambda \neq \mu$ case.

We now move onto the main purpose of our couplings: to show that the number of individuals infected in the birth–death chain *dominates* that in the epidemic. This is intuitive: the two process are the same, except that some infections are censored.

Theorem 4.3.10 (BD Dominates SIR). *Let* $(X, Y) \sim \mathsf{SBDP}(\alpha, \beta)$ *and* $(S, I, R) \sim \mathsf{SIR}(\frac{\alpha}{n-i_0}, \beta)$ *with* $X_0 = i_0 = i_0$ *and* $Y_0 = 0 = R_0$. *Then,* $I + R \lesssim X + Y$*: there exists a coupling of* (X, Y) *and* (S, I, R) *such that*

$$I_t + R_t \leq X_t + Y_t \quad \text{for all} \quad t \geq 0.$$

Proof. The coupling we use is given by Definitions 4.3.4 and 4.3.6: use the same source of randomness for both processes, except for the extra, iid Unif$([0, 1])$-s.

We would like to say, "Every accepted infection in SIR happens at the same time as one in BD. Hence, SIR $\leq$ BD." However, this is not true: individuals get infected at different times in the two models, so the (attempted) infection times $\tau_{i,j}$ are different in the two models. This is crucial, though: each individual *attempts* the same number of infections in each model, but with delays in SIR. See Figure 4.4.

We look at times s_i at which individuals get infected:

$$s_i^{\mathsf{SIR}} = \inf\{t \geq 0 \mid I_t + R_t \geq i\} \quad \text{and} \quad s_i^{\mathsf{BD}} = \inf\{t \geq 0 \mid X_t + Y_t \geq i\}.$$

The result follows if $s_i^{\mathsf{SIR}} \geq s_i^{\mathsf{BD}}$—ie, it takes longer for the i-th individual to be infected in the SIR than in the BD model—for all i. We prove this by induction on i.

Clearly, $s_i^{\mathsf{SIR}} = 0 = s_i^{\mathsf{BD}}$ for all $i \leq i_0$. Let $j > i_0$ and suppose that $s_i^{\mathsf{SIR}} \geq s_i^{\mathsf{BD}}$ for all $i \leq j - 1$. Remember that individuals are infected in numerical order. Let

$$N_{j-1}(t) \coloneqq \#\{k \in \mathbb{N} \mid \text{individual } k \text{ infected by one of } 1, \ldots, j-1 \text{ by time } t\}.$$

Key is the coupling between the (attempted) infection times: for individual i, they are at $\tau_{i,j} = s_i + \sum_{n=1}^{j} t_{i,n}$ for j such that $s_i + \tau_{i,j} \in [s_i, s_i + T_i)$; see Figure 4.4. This means that, by time t, the number of (attempted) infections by i in SIR is at most that for BD; some are rejected in SIR. Summing over all $i \in \{1, \ldots, j-1\}$ gives

$$N_{j-1}^{\mathsf{SIR}}(t) \leq N_{j-1}^{\mathsf{BD}}(t) \quad \text{for all} \quad t \geq 0.$$

But,

$$s_j^{\mathsf{SIR/BD}} = \inf\{t \geq 0 \mid N_{j-1}^{\mathsf{SIR/BD}}(t) = j - i_0\}.$$

Hence, $s_j^{\mathsf{SIR}} \geq s_j^{\mathsf{BD}}$, which completes the inductive step and the proof. □

Example 4.3.11. Suppose that we say there is an *epidemic* if the total number ever infected exceeds a predetermined threshold γn, with $\gamma \in (0, 1)$. From the coupling,

$$\max_{t \geq 0}\{X_t + Y_t\} < \gamma n \implies \max_{t \geq 0}\{I_t + R_t\} < \gamma n.$$

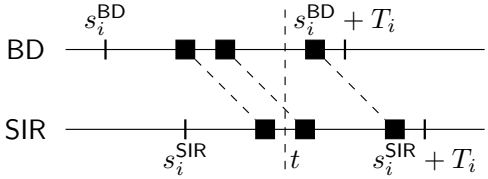

Figure 4.4. Infection attempts are shown on a timeline by filled squares; only those in $[s_i, s_i + T_i)$ are shown, the boundary of which is shown by vertical lines. This interval is different in BD vs SIR. The infection attempts in both occur at the same times after s_i; their coupling is illustrated via the diagonal dashed lines. The number of infection attempts in SIR by t (at the dotted vertical line) is at most that in BD; some are rejected in SIR

Thus, taking the limit $n \to \infty$ and using Proposition 2.3.2, we have

$$\lim_{n\to\infty} \mathbb{P}\{\text{no epidemic}\} \geq \mathbb{P}\{\mathsf{SBDP}(\alpha, \beta) \text{ dies out}\} = \min\big\{\big(\tfrac{\beta}{\alpha}\big)^{i_0},\, 1\big\}.$$

We can actually get a lower bound on the limiting probability of having an epidemic, too. Before passing the threshold γn, at least $(1-\gamma)n$ are susceptible. Thus,

$$\alpha I_t S_t/(n - i_0) \geq (1-\gamma)\alpha n I_t/(n - i_0).$$

Adjusting the censoring, we get a lower bound

$$I + R \gtrsim X' + Y' \quad \text{where} \quad (X', Y') \sim \mathsf{SBDP}(\alpha(1-\gamma), \beta).$$

This gives the complementary bound

$$\lim_{n\to\infty} \mathbb{P}\{\text{no epidemic}\} \leq \mathbb{P}\{\mathsf{SBDP}(\alpha(1-\gamma), \beta) \text{ dies out}\} = \min\big\{\big(\tfrac{\beta}{\alpha(1-\gamma)}\big)^{i_0},\, 1\big\}. \quad \triangle$$

Exercise 4.3.12. *Verify the lower bound $I + R \gtrsim X' + Y'$, checking the rates.*

The previous example is formalised in the following theorem.

Theorem 4.3.13 (Whittle's Threshold Theorem, [Whi55]). *Let $(S, I, R) \sim \mathsf{SIR}(\frac{\alpha}{n-i_0}, \beta)$ with $(S_0, I_0, R_0) = (n - i_0, i_0, 0)$. Let $\gamma \in (0, 1)$, independent of n. Say that there is an epidemic if the total number of infected ever exceeds γn. Then,*

$$\min\big\{\big(\tfrac{\beta}{\alpha}\big)^{i_0},\, 1\big\} \leq \lim_{n\to\infty} \mathbb{P}\{\text{no epidemic}\} \leq \min\big\{\big(\tfrac{\beta}{\alpha(1-\gamma)}\big)^{i_0},\, 1\big\}.$$

Furthermore, if we let $\gamma = \gamma_n$ depend on n, then

$$\lim_{n\to\infty} \mathbb{P}\{\text{no epidemic}\} = \min\big\{\big(\tfrac{\beta}{\alpha}\big)^{i_0},\, 1\big\} \quad \text{if} \quad \lim_{n\to\infty} \gamma_n = 0 \quad \text{and} \quad \lim_{n\to\infty} \gamma_n n = \infty.$$

Example 4.3.14 (Influenza in a Finite Population)**.** Suppose that an influenza outbreak proceeds as an SIR model. Suppose that the infectious period averages three days and the initial rate of infection is such that an individual comes in contact with one person per day, on average. So, we are studying $\mathsf{SIR}(\alpha = 1, \beta = 1/3)$.

Suppose that there is initially one infected person. Show that the probability of having an epidemic does not exceed $\frac{2}{3}$ in the limit as the population size $n \to \infty$.

From Whittle's threshold theorem,

$$\lim_{n\to\infty} \mathbb{P}\{\text{epidemic}\} = 1 - \lim_{n\to\infty} \mathbb{P}\{\text{no epidemic}\} \leq 1 - \frac{\beta}{\alpha} = 1 - \frac{1}{3} \leq \frac{2}{3}. \qquad \triangle$$

Coupling theory is an *extremely* important tool in probability theory. In particular, the idea of coming up with a *monotone* coupling between two random processes X and Y on $\mathbb{R}$ such that $X_t \geq Y_t$ for all $t \geq 0$ is a technique which comes up all the time in research. In epidemic and queueing theory, these results typically rely heavily on Poisson thinning, and sometimes on Poisson superposition.

Bibliography

[Bai75] N. T. J. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications. 2nd edition.* Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975

[Bré20] P. Brémaud. *Markov Chains—Gibbs Fields, Monte Carlo Simulation and Queues*. Vol. 31. Texts in Applied Mathematics. Springer, Cham, 2020, pp. xvi+557

[GW14] G. Grimmett and D. Welsh. *Probability: An Introduction*. 2nd ed. Oxford University Press, Oxford, 2014, pp. x+270

[Jar79] R. G. Jarrett. "A Note on the Intervals Between Coal-Mining Disasters". In: *Biometrika* 66.1 (Apr. 1979), pp. 191–193. eprint: https://academic.oup.com/biomet/article-pdf/66/1/191/600109/66-1-191.pdf

[KY14] F. Kelly and E. Yudovina. *Stochastic Networks*. Vol. 2. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge, 2014

[Nor97] J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1997

[Whi55] P. Whittle. "The Outcome of a Stochastic Epidemic—A Note on Bailey's Paper". In: *Biometrika* 42 (1955), pp. 116–122