CONTINUOUS-TIME MARKOV CHAINS

by

Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027-6699
Email: ww2040@columbia.edu
URL: www.columbia.edu/∼ww2040

December 19, 2006

1. Introduction

We now turn to **continuous-time Markov chains (CTMC's)**, which are a natural sequel to the study of discrete-time Markov chains (DTMC's), the Poisson process and the exponential distribution, because CTMC's combine DTMC's with the Poisson process and the exponential distribution. Most properties of CTMC's follow directly from results about DTMC's, the Poisson process and the exponential distribution. .

Like DTMC's, CTMC's are Markov processes that have a discrete state space, which we can take to be the positive integers. Just as with DTMC's, we will focus on the special case of a **finite state space**, but the theory and methods extend to infinite discrete state spaces, provided we impose additional regularity conditions. We will usually assume that the state space is the set $\{0, 1, 2, \ldots, n\}$ containing the first $n+1$ nonnegative integers for some positive integer n, but any finite set can be so labelled. Just as with DTMC's, a finite state space allows us to apply square (finite) matrices and elementary linear algebra. The main difference is that we now consider **continuous time**. We consider a stochastic process $\{X(t) : t \geq 0\}$, where time t is understood to be any nonnegative real number. The random variable $X(t)$ is the state occupied by the CTMC at time t.

As we will explain in Section 3, a CTMC can be viewed as a DTMC with altered transition times. Instead of unit times between successive transitions, the times between successive transitions are allowed to be independent exponential random variables with means that depend only on the state from which the transition is being made. Alternatively, as we explain in Subsection 3.4, a CTMC can be viewed as a DTMC (a different DTMC) in which the transition times occur according to a Poisson process. In fact, we already have considered a CTMC with just this property (but infinite state space), because the Poisson process itself is a CTMC. For that CTMC, the associated DTMC starts in state 0 and has only unit upward transitions, moving from state i to state $i+1$ with probability 1 for all i. A CTMC generalizes a Poisson process by allowing other transitions. For a Poisson process, $X(t)$ goes to infinity as $t \to \infty$. We will be interested in CTMC's that have proper limiting distributions as $t \to \infty$.

Here is how the chapter is organized: We start in Section 2 by discussing transition probabilities and the way they can be used to specify the finite-dimensional distributions, which in turn specify the probability law of the CTMC. Then in Section 3 we describe four different ways to construct a CTMC model, giving concrete examples. In Section 4 we indicate how to calculate the limiting probabilities for an irreducible CTMC. There are different ways, with the one that is most convenient usually depending on the modelling approach. In Section 5 we discuss the special case of a birth-and-death process, in which the only possible transitions are up one or down one to a neighboring state. The number of customers in a queue (waiting line) can often be modelled as a birth-and-death process. The special structure of a birth-and-death process makes the limiting probabilities even easier to compute. Finally, in Section 6 we discuss reverse-time CTMC's and reversibility. We apply those notions to start analyzing some networks of queues.

2. Transition Probabilities and Finite-Dimensional Distributions

Just as with discrete time, a continuous-time stochastic process is a **Markov process** if the conditional probability of a future event given the present state and additional information about past states depends only on the present state. A **CTMC** is a continuous-time Markov process with a discrete state space, which can be taken to be a subset of the nonnegative integers. That is, a stochastic process $\{X(t) : t \geq 0\}$ (with an integer state space) is a CTMC if

$$P(X(s+t) = j|X(s) = i, X(r) = i_r, r \in A_s \subseteq [0, s)) = P(X(s+t) = j|X(s) = i) \qquad (2.1)$$

for all states i and j and for all times $s > 0$ and $t > 0$. On the left in (2.1), we are conditioning on the values of $X(r)$ for all times r in a subset of "past" times A_s in addition to the value at the "present" time s. In general, A_s could be an arbitrary subset of $[0, s) \equiv \{r : 0 \leq r < s\}$, but to have the conditional probability in (2.1) well defined by elementary methods, we assume that A_s is a finite subset.

The conditional probabilities $P(X(s+t) = j|X(s) = i)$ are called the **transition probabilities**. We will consider the special case of **stationary transition probabilities** (sometimes referred to as homogeneous transition probabilities), occurring when

$$P(X(s+t) = j|X(s) = i) = P(X(t) = j|X(0) = i) \equiv P_{i,j}(t) \qquad (2.2)$$

for all states i and j and for all times $s > 0$ and $t > 0$; the independence of s characterizes the stationarity. We assume stationary transition probabilities unless stipulated otherwise.

Thus a key concept for CTMC's is the notion of transition probabilities. However, the transition probabilities of CTMC's are not so easy to work with. As a consequence, we usually do not directly use transition probabilities when we construct and analyze CTMC models. First, when we construct a CTMC model, we invariably do not directly define the transition probabilities (although their structure will be implied by what we do define). Second, after constructing a CTMC model, we usually do not calculate the transition probabilities. Instead, we usually calculate the associated **limiting probabilities**, denoted by α_j:

$$\alpha_j \equiv \lim_{t\to\infty} P_{i,j}(t) \equiv \lim_{t\to\infty} P(X(t) = j|X(0) = i) \ , \qquad (2.3)$$

because they are much easier to calculate, and because they usually serve as excellent approximations for the exact transition probabilities $P_{i,j}(t)$ when t is large. (We use the notation α for the limiting probability vector of the CTMC, instead of π, because we reserve π for the limiting probability vector for an associated DTMC; see Section 3.1 and Theorem 4.2.)

Consistent with what we have written in (2.3), under regularity conditions, the limiting probabilities α_j will not to depend on the initial state. Indeed, that will be true provided the CTMC is **irreducible**, which means (just as in discrete time) that it is possible with some positive probability to get from any state to any other state at some finite time, which may involve multiple transitions. (Just as in discrete time, for irreducibility, we do not require that we reach these other states in a single transition.) We assume irreducible CTMC's unless stipulated otherwise.

This chapter is largely about constructing CTMC models and calculating the limiting probability vector $\alpha \equiv (\alpha_0, \alpha_1, \ldots, \alpha_n)$. As with DTMC's, we will also want to apply the limiting probability vector α to answer a variety of related questions of interest. But, to repeat, neither constructing the CTMC model nor calculating the limiting probability vector α will directly involve the transition probabilities. Nevertheless, the transition probabilities are very important for understanding CTMC's.

Just as in discrete time, the evolution of the transition probabilities over time is described by the Chapman-Kolmogorov equations, but they take a different form in continuous time. In formula (2.4) below, we consider a sum over all possible states at some intermediate time. In doing so, we simply write a sum over integers. When we do that, we understand the sum to be over all possible states.

Lemma 2.1. (Chapman-Kolmogorov equations) *For all* $s \geq 0$ *and* $t \geq 0$*,*

$$P_{i,j}(s+t) = \sum_k P_{i,k}(s)P_{k,j}(t) \ . \tag{2.4}$$

Proof. We can compute $P_{i,j}(s+t)$ by considering all possible places the chain could be at time s. We then condition and and uncondition, invoking the Markov property to simplify the conditioning; i.e.,

$$\begin{aligned}
P_{i,j}(s+t) &= P(X(s+t)=j|X(0)=i) \\
&= \sum_k P(X(s+t)=j, X(s)=k|X(0)=i) \\
&= \sum_k P(X(s)=k|X(0)=i)P(X(s+t)=j|X(s)=k, X(0)=i) \\
&= \sum_k P(X(s)=k|X(0)=i)P(X(s+t)=j|X(s)=k) \quad \text{(Markov property)} \\
&= \sum_k P_{i,k}(s)P_{k,j}(t) \quad \text{(stationary transition probabilities)} \ . \quad \blacksquare
\end{aligned}$$

Using matrix notation, we write $P(t)$ for the square matrix of transition probabilities $(P_{i,j}(t))$, and call it the **transition function**. In matrix notation, the Chapman-Kolmogorov equations reduce to a simple relation among the transition functions involving **matrix multiplication**:

$$P(s+t) = P(s)P(t) \tag{2.5}$$

for all $s \geq 0$ and $t \geq 0$. It is important to recognize that (2.5) means (2.4). From the perspective of abstract algebra, equation (2.5) says that the transition function has a semi-group property, where the single operation is matrix multiplication. (It is not a group because an inverse is missing.)

A CTMC is well specified if we specify: (1) its initial probability distribution - $P(X(0)=i)$ for all states i - and (2) its transition probabilities - $P_{i,j}(t)$ for all states i and j and positive times t. First, we can use these two elements to compute the distribution of $X(t)$ for each t, namely,

$$P(X(t)=j) = \sum_i P(X(0)=i)P_{i,j}(t) \ . \tag{2.6}$$

However, in general, we want to do more. We want to know about the joint distributions in order to capture the dependence structure. Recall that the **probability law of a stochastic process** is understood to be the set of all its finite-dimensional distributions. A **finite-dimensional distribution** is

$$P(X(t_1)=j_1, X(t_2)=j_2, \ldots, X(t_k)=j_k) \tag{2.7}$$

for states j_i and times t_i satisfying $0 \leq t_1 < t_2 < \cdots < t_k$. The probability law is specified by all these finite-dimensional distributions, considering all positive integers k, and all sets of k states and k ordered times. It is important that we can express any finite-dimensional distribution in terms of the initial distribution and the transition probabilities. For example, assuming that $t_1 > 0$, we have

$$\begin{aligned}
&P(X(t_1)=j_1, X(t_2)=j_2, \ldots, X(t_k)=j_k) \\
&\quad = \sum_{j_0} P(X(0)=j_0)P_{j_0,j_1}(t_1)P_{j_1,j_2}(t_2-t_1) \times \cdots \times P_{j_{k-1},j_k}(t_k-t_{k-1}) \ . \quad (2.8)
\end{aligned}$$

In summary, equation (2.8) shows that we succeed in specifying the full probability law of the DTMC, as well as all the marginal distributions via (2.6), by speciying the initial probability distribution - $P(X(0) = i)$ for all i - and the transition probabilities $P_{i,j}(t)$ for all t, i and j or, equivalently, the transition function $P(t)$. However, when we construct CTMC models, as we do next, we do not directly specify the transition probabilities. We will see that, at least in principle, the transition probabilities can be constructed from what we do specify, but we usually do not carry out that step.

3. Modelling

We now turn to modelling: constructing a CTMC model. We saw that a DTMC model is specified by simply specifying its one-step transition matrix P and the initial probability distribution. Unfortunately, the situation is more complicated in continuous time.

In this section we will describe **four different approaches** to constructing a CTMC model. With each approach, we will need to specify the initial distribution, so we are focusing on specifying the model beyond the initial distribution. The four approaches are equivalent: You get to the same result from each and you can get to each from any of the others. Even though these four approaches are redundant, they are useful because they together give a different more comprehensive view of a CTMC. We see different things from different perspectives, much like the Indian fable about the blind men and the elephant, recaptured in the poem by John Godfrey Saxe (1816-1887):

The Blind Men and the Elephant

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

The First approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a wall!"

The Second, feeling of the tusk,
Cried, "Ho! what have we here
So very round and smooth and sharp?
To me tis mighty clear
This wonder of an Elephant
Is very like a spear!"

The Third approached the animal,
And happening to take
The squirming trunk within his hands,

Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a snake!"

The Fourth reached out an eager hand,
And felt about the knee.
"What most this wondrous beast is like
Is mighty plain," quoth he;
" Tis clear enough the Elephant
Is very like a tree!"

The Fifth, who chanced to touch the ear,
Said: "Een the blindest man
Can tell what this resembles most;
Deny the fact who can
This marvel of an Elephant
Is very like a fan!"

The Sixth no sooner had begun
About the beast to grope,
Than, seizing on the swinging tail
That fell within his scope,
"I see," quoth he, "the Elephant
Is very like a rope!"

And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!

For some applications, one modelling approach may be more natural than the others. Or one modelling approach may be more convenient for analyzing the model.

3.1. A DTMC with Exponential Transition Times

In order to construct a CTMC model, it is natural to build on our knowledge of DTMC's. So we first consider a way to exploit DTMC's in our construction of the CTMC model. To do so in the strongest way, we start with a DTMC having a transition matrix P, and then modify the way the transitions occur. Instead of having each transition take unit time, now we assume that each transition takes a random time. In particular, we assume that the time required to make a transition from state i has an exponential distribution with rate ν_i, and thus mean $1/\nu_i$, independent of the history before reaching state i.

This modelling approach is convenient for simulating the CTMC; we can recursively generate successive transitions. This modelling approach also avoids technical complications that arise in the conventional transition-rate approach, to be introduced in the next subsection.

This modelling approach is also appealing because many applications are naturally expressed in this way.

Example 3.1. (Pooh Bear and the Three Honey Trees) A bear of little brain named Pooh is fond of honey. Bees producing honey are located in three trees: tree A, tree B and tree C. Tending to be somewhat forgetful, Pooh goes back and forth among these three honey trees randomly (in a Markovian manner) as follows: From A, Pooh goes next to B or C with probability 1/2 each; from B, Pooh goes next to A with probability 3/4, and to C with probability 1/4; from C, Pooh always goes next to A. Pooh stays a random time at each tree. (Assume that the travel times can be ignored.) Pooh stays at each tree an exponential length of time, with the mean being 5 hours at tree A or B, but with mean 4 hours at tree C. Construct a CTMC enabling you to find the limiting proportion of time that Pooh spends at each honey tree.

Note that this problem is formulated directly in terms of the DTMC, describing the random motion at successive transitions, so it is natural to use this initial modelling approach. Here the transition matrix for the DTMC is

$$P = \begin{array}{c} A \\ B \\ C \end{array} \left(\begin{array}{ccc} 0 & 1/2 & 1/2 \\ 3/4 & 0 & 1/4 \\ 1 & 0 & 0 \end{array} \right) .$$

In the displayed transition matrix P, we have only labelled the rows. The columns are assumed to be labelled in the same order.

As specified above, the exponential times spent at the three trees have means $1/\nu_A = 1/\nu_B = 5$ hours and $1/\nu_C = 4$ hours. In the Section 4 we will see how we can calculate the limiting probabilities for this CTMC and answer the question about the long-run proportion of time that Pooh spends at each tree. ■

With this initial modelling approach, it is natural to assume, as was the case in Example 3.1, that there are no one-step transitions in the DTMC from any state immediately back to itself, but it is not necessary to make that assumption. We get a CTMC from a DTMC and exponential transition times without making that assumption.

However, to help achieve a simple relation between the first two modelling approaches, we make that assumption here: **We assume that there are no one-step transitions from any state to itself in the DTMC**; i.e., we assume that $P_{i,i} = 0$ for all i. However, we emphasize that this assumption is not critical, as we will explain after we introduce the third modelling approach. Indeed, we will want to allow transitions from a state immediately to itself in the fourth - uniformization - modelling approach. That is a crucial part of that modelling approach.

In closing this subsection, we remark that this first modelling approach corresponds to treating the CTMC as a special case of a **semi-Markov process (SMP)**. An SMP is a DTMC with independent random transition times, but it allows the distributions of the intervals between transitions to be non-exponential; see Section **??**.

3.2. Transition Rates and ODE's

A second modelling approach is based on representing the transition probabilities as the solution of a system of ordinary differential equations, which allows us to apply well-established modelling techniques from the theory of differential equations in a deterministic setting; e.g., see Simmons (1991). With this second modelling approach, we directly specify transition rates.

We proceed with that idea in mind, but without assuming knowledge about differential equations. We focus on the transition probabilities of the CTMC, even though they have not yet been specified. With the transition probabilities in mind, we assume that there are well-defined derivatives (from above or from the right) of the transition probabilities at 0. We assume these derivatives exist, and call them transition rates.

But first we must define **zero-time transition probabilities**, which we do in the obvious way: We let $P(0) = I$, where I is the identity matrix; i.e., we set $P_{i,i}(0) = 1$ for all i and we set $P_{i,j}(0) = 0$ whenever $i \neq j$. We are just assuming that you cannot go anywhere in zero time.

We then let the transition rate from state i to state j be defined in terms of the **derivatives**:

$$Q_{i,j} \equiv P'_{i,j}(0+) \equiv \frac{d\,P_{i,j}(t)}{dt}|_{t=0+} \; . \tag{3.1}$$

In (3.1) 0+ appears to denote the right derivative at 0 because $P_{i,j}(t)$ is not defined for $t < 0$.

This approach is used in most treatments of CTMC's, but without mentioning derivatives or right-derivatives. Instead, it is common to assume that

$$P_{i,j}(h) = Q_{i,j}h + o(h) \quad \text{as} \quad h \downarrow 0 \quad \text{if} \quad j \neq i \tag{3.2}$$

and

$$P_{i,i}(h) - 1 = Q_{i,i}h + o(h) \quad \text{as} \quad h \downarrow 0 \; , \tag{3.3}$$

where $o(h)$ is understood to be a quantity which is asymptotically negligible as $h \downarrow 0$ after dividing by h. (Formally, $f(h) = o(h)$ as $h \downarrow 0$ if $f(h)/h \to 0$ as $h \downarrow 0$.)

For a finite state space, which we have assumed, and for infinite state spaces under extra regularity conditions, we will have

$$-Q_{i,i} \equiv \sum_{j, j \neq i} Q_{i,j} \tag{3.4}$$

because the transition probabilities $P_{i,j}(t)$ sum over j to 1. Moreover, we have

$$-Q_{i,i} = \nu_i \quad \text{for all} \quad i \; , \tag{3.5}$$

because we have assumed that $P_{i,i} = 0$ in the first modelling approach.

In other words, these two assumptions mean that

$$\lim_{h \downarrow 0} \frac{P_{i,j}(h) - P_{i,j}(0)}{h} = Q_{i,j} \quad \text{for all} \quad i \quad \text{and} \quad j \; , \tag{3.6}$$

which is just what is meant by (3.1).

In summary, we first assumed that transition probabilities are well defined, at least for zero time and small positive time intervals, and then assume that they are differentiable from the right at 0. We remark that it is possible to weaken that assumption, and only assume that the transition probabilities are continuous at 0: $P(h) \to P(0) \equiv I$ as $h \downarrow 0$. Then it is possible to prove that the derivatives exist; see Sections II.1 and II.2 of Chung (1967).

Having defined the transition rates in terms of the assumed behavior of the transition probabilities in a very short (asymptotically negligible) interval of time, we can specify the CTMC model by specifying these transition rates; i.e., we specify the transition-rate matrix Q, having elements $Q_{i,j}$. (But we do not first fully define the transition probabilities themselves!) Thus, just as we specify a DTMC model via a matrix P, we can specify a CTMC model via the transition-rate matrix Q.

When specifying the transition-rate matrix Q, it suffices to specify the off-diagonal elements $Q_{i,j}$ for $i \neq j$, because the diagonal elements $Q_{i,i}$ are always defined by (3.4). The off-diagonal elements are always nonnegative, whereas the diagonal elements are always negative. Each row sum of Q is zero.

Even though this modelling approach for CTMC's is similar to what we did for DTMC's, it is more complicated, because the rate matrix Q is harder to interpret than the one-step transition matrix P. (The discussion above is intended to help interpretation.) In fact, this approach to CTMC modelling is perhaps best related to modelling with ordinary differential equations, as mentioned at the beginning of this subsection.

To construct the transition probabilities $P_{i,j}(t)$ from the transition rates $Q_{i,j} \equiv P'_{i,j}(0+)$, we apply the Chapman-Kolmogorov equations in Lemma 2.1 in order to show that the transition probabilities satisfy **two systems of ordinary differential equations (ODE's)** generated by the transition rates. In matrix notation, these will be simple first-order linear ODE's.

Theorem 3.1. (Kolmogorov forward and backward ODE's) *The transition probabilities satisfy both the Kolmogorov* ***forward*** *differential equations*

$$P'_{i,j}(t) = \sum_k P_{i,k}(t) Q_{k,j} \quad \textit{for all} \quad i \quad \textit{and} \quad j \ , \tag{3.7}$$

which in matrix notation is the matrix ODE

$$P'(t) = P(t)Q \ , \tag{3.8}$$

and the Kolmogorov ***backward*** *differential equations*

$$P'_{i,j}(t) = \sum_k Q_{i,k} P_{k,j}(t) \quad \textit{for all} \quad i \quad \textit{and} \quad j \ , \tag{3.9}$$

which in matrix notation is the matrix ODE

$$P'(t) = QP(t) \ , \tag{3.10}$$

Proof. We start with the forward equation, using matrix notation. We apply the Chapman-Kolmogorov equations in Lemma 2.1 to write

$$P(t+h) = P(t)P(h) \ ,$$

and then do an asymptotic analysis as $h \downarrow 0$. (This is tantamount to doing a careful asymptotic analysis of what happens in a small interval after time t.) We subtract $P(t)$ from both sides and divide by h, to get

$$\frac{P(t+h) - P(t)}{h} = P(t)\frac{P(h) - I}{h} \ ,$$

where I is the identity matrix. Recalling that $I = P(0)$, we can let $h \downarrow 0$ to get the desired result (3.8). To get the backward equation (3.10), we start with

$$P(t+h) = P(h)P(t)$$

and reason in the same way. (This is tantamount to doing a careful asymptotic analysis of what happens in a small interval after time 0, and then applying $P(t)$ thereafter.) ∎

To help remember which ODE is forward and which is backwards, note that $P(t)Q$ appearing on the righthand side of the forward ODE is in alphabetic order, whereas $QP(t)$

appearing on the righthand side of the backward ODE is in reverse (backward) alphabetic order.

With a finite state space, both ODE's are always well defined. With an infinite state space, there can be technical problems, because there could be infinitely many transitions in finite time. With an infinite state space, the forward ODE can be more problematic, because it presumes the process got to time t before doing the asymptotic analysis. Here we assume a finite state space, so we do not encounter those pathologies. Under regularity conditions, those pathologies will not occur with infinite state spaces either.

To obtain the transition function $P(t)$ from the transition-rate matrix Q, we can solve one of these ODE's. In preparation, we review the simple one-dimensional story. Suppose that we have an ODE $f'(t) = cf(t)$, where f is understood to be a differentiable real-valued function f with known initial value $f(0)$. If we divide both sides by $f(t)$, we get $f'(t)/f(t) = c$. Since $f'(t)/f(t)$ is the derivative of $\log f(t)$, we can integrate to get

$$\log f(t) - \log f(0) = ct \quad \text{or} \quad f(t) = f(0)e^{ct}, \quad t \geq 0 \ .$$

Thus we see that f must be an exponential function.

Closely paralleling the real-valued case, the matrix ODE's in (3.8) and (3.10) have an exponential solution, but now a matrix-exponential solution. (Since $P(0) = I$, the initial condition plays no role, just as above when $f(0) = 1$.) In particular, as a consequence of Theorem 3.1, we have the following corollary.

Theorem 3.2. (matrix exponential representation) *The transition function can be expressed as a matrix-exponential function of the rate matrix Q, i.e.,*

$$P(t) = e^{Qt} \equiv \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!} \tag{3.11}$$

This matrix exponential is the unique solution to the two ODE's with initial condition $P(0) = I$.

Proof. If we verify or assume that we can interchange summation and differentiation in (3.11), we can check that the displayed matrix exponential satisfies the two ODE's:

$$\frac{d}{dt}\sum_{n=0}^{\infty} \frac{Q^n t^n}{n!} = \sum_{n=0}^{\infty} \frac{d}{dt}\frac{Q^n t^n}{n!} = \sum_{n=0}^{\infty} \frac{nQ^n t^{n-1}}{n!} = Q\sum_{n=0}^{\infty} \frac{Q^n t^n}{n!} = Qe^{Qt} \ .$$

We give a full demonstration at the end of Subsection 3.4. ■

However, in general the transition function $P(t)$ is not elementary to compute via (3.11); see Moler and Van Loan (2003). Indeed, one of the ways to evaluate the matrix-exponential function displayed in (3.11) is to numerically solve one of the ODE's as expressed in (3.8) or (3.10).

We now illustrate this second modelling approach with an example.

Example 3.2. (Copier Breakdown and Repair) Consider two copier machines that are maintained by a single repairman. Machine i functions for an exponentially distributed amount of time with mean $1/\gamma_i$, and thus rate γ_i, before it breaks down. The repair times for copier i are exponential with mean $1/\beta_i$, and thus rate β_i, but the repairman can only work on one machine at a time. Assume that the machines are repaired in the order in which they fail. Suppose that we wish to construct a CTMC model of this system, with the goal of finding the long-run proportions of time that each copier is working and the repairman is busy. How can we proceed?

An initial question is: What should be the state space? Can we use 4 states, letting the states correspond to the subsets of failed copiers? Unfortunately, the answer is "no," because in order to have the Markov property we need to know which copier failed first when both copiers are down. However, we can use 5 states with the states being: 0 for no copiers failed, 1 for copier 1 is failed (and copier 2 is working), 2 for copier 2 is failed (and copier 1 is working), $(1,2)$ for both copiers down (failed) with copier 1 having failed first and being repaired, and $(2,1)$ for both copiers down with copier 2 having failed first and being repaired. (Of course, these states could be relabelled 0, 1, 2, 3 and 4, but we do not do that.)

From the problem specification, it is natural to work with transition rates, where these transition rates are obtained directly from the originally-specified failure rates and repair rates (the rates of the exponential random variables). In Figure 1 we display a **rate diagram** showing the possible transitions with these 5 states together with the appropriate rates. It can be helpful to construct such rate diagrams as part of the modelling process.

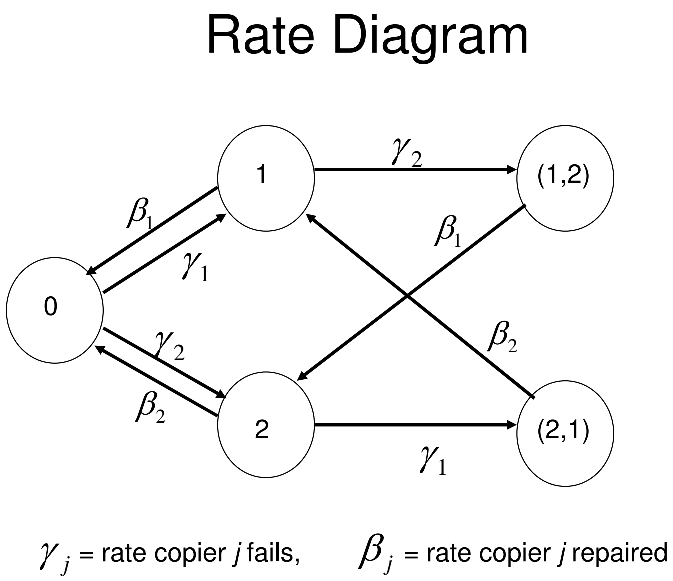

Figure 1: A rate diagram showing the transition rates among the 5 states in Example 3.2, involving copier breakdown and repair.

From Figure 1, we see that there are 8 possible transitions. The 8 possible transitions should clearly have transition rates

$$Q_{0,1} = \gamma_1, Q_{0,2} = \gamma_2, Q_{1,0} = \beta_1, Q_{1,(1,2)} = \gamma_2, Q_{2,0} = \beta_2, Q_{2,(2,1)} = \gamma_1, Q_{(1,2),2} = \beta_1, Q_{(2,1),1} = \beta_2 .$$

In other words, the rate matrix should be

$$Q = \begin{array}{c} 0 \\ 1 \\ 2 \\ (1,2) \\ (2,1) \end{array} \begin{pmatrix} -(\gamma_1+\gamma_2) & \gamma_1 & \gamma_2 & 0 & 0 \\ \beta_1 & -(\gamma_2+\beta_1) & 0 & \gamma_2 & 0 \\ \beta_2 & 0 & -(\gamma_1+\beta_2) & 0 & \gamma_1 \\ 0 & 0 & \beta_1 & -\beta_1 & 0 \\ 0 & \beta_2 & 0 & 0 & -\beta_2 \end{pmatrix} .$$

In Section 4, we will compute the limiting probability distribution of this CTMC and answer the questions posed above. ∎

3.3. Competing Clocks with Exponential Timers

We now present a third modelling approach, which is an appealing constructive alternative to the second modelling approach based on rates of unknown transition functions. This third modelling approach is even more natural for Example 3.2. This third approach also helps link the first two modelling approaches.

With this third modelling approach, movement from state to state is determined by "competing" clocks with timers that go off at random, exponentially-distributed, times. For each state i, there is a clock associated with each state j the process could possibly move to in a single transition from state i. Let C_i be the set of states that the CTMC can possibly move to from state i in a single transition. Equivalently, C_i is the **set of active clocks** in state i. (We here assume that the process does not move from state i immediately back to state i.)

Each time the CTMC moves to state i, clocks with **timers are set or reset, if necessary** to go off at random times $T_{i,j}$ for each $j \in C_i$. Each clock has an exponential timer; i.e., the random time $T_{i,j}$ is given an exponential distribution with (positive finite) rate $Q_{i,j}$ and thus mean $1/Q_{i,j}$ (depending on i and j). Moreover, we assume that these newly set times $T_{i,j}$ are mutually independent and independent of the history of the CTMC prior to that transition time. By the lack-of-memory property of the exponential distribution, resetting running timers is equivalent (leaves the probability law of the stochastic process unchanged) to not resetting the times, and letting the timers continue to run.

Example 3.3. (Copier Breakdown and Repair Revisited) At this point we should reconsider Example 3.2 and observe that it is even more natural to define the CTMC through the proposed clocks with random timers. The random times triggering transitions are the exponential times to failure and times to repair specified in the original problem formulation. However, there is a difference: In the actual system, those random times do not get reset at each transition epoch. But, because of the lack-of-memory property of the exponential distribution, a timer that is still going can be reset at any time, including one of these random transition times, without changing the distribution of the remaining time. Thus the clocks with random timers does produce a valid representation of the desired CTMC model. ∎

We now discuss the **implications** of the model specification in terms of exponential timers. As a consequence of these probabilistic assumptions, with probability 1, no two timers ever go off at the same time. (Since the exponential distribution is a continuous distribution, the probability of any single possible value is 0.) Let T_i be the time that the first timer goes off in state i and let N_i be the index j of the random timer that the first goes off; i.e.,

$$T_i \equiv \min_j \{T_{i,j}\} \tag{3.12}$$

and

$$N_i \equiv j \quad \text{such that} \quad T_{i,j} = T_i \ . \tag{3.13}$$

(The index j yielding the minimum is often called the *argmin.*) We then let the let the process move from state i next to state N_i after an elapsed time of T_i, and we repeat the process, starting from the new state N_i.

To understand the implications of these exponential clocks, recall basic properties of the exponential distribution. Recall that **the minimum of several independent exponential random variables** is again an exponential random variable with a rate equal to the sum of the rates. Hence, T_i has an exponential distribution, i.e.,

$$P(T_i \le t) = 1 - e^{-\nu_i t}, \quad t \ge 0 \ , \tag{3.14}$$

where

$$\nu_i \equiv -Q_{i,i} = \sum_{j,j\neq i} Q_{i,j} \ , \tag{3.15}$$

as in (3.4) and (3.5). (Again we use the assumption that $P_{i,i} = 0$ in the first modelling approach.)

Moreover, recall that, when considering several independent exponential random variables, each exponential random variable is **the exponential random variable yielding the minimum** with a probability proportional to its rate, so that

$$P(N_i = j) = \frac{Q_{i,j}}{\sum_{k,k\neq i} Q_{i,k}} = \frac{Q_{i,j}}{\nu_i} \quad \text{for} \quad j \neq i \ . \tag{3.16}$$

Moreover, as discussed before in relation to the exponential distribution, the random variables T_i and N_i are **independent random variables**:

$$P(T_i \le t, N_i = j) = P(T_i \le t)P(N_i = j) = (1 - e^{-\nu_i t})\left(\frac{Q_{i,j}}{\nu_i}\right) \quad \text{for all} \quad t \quad \text{and} \quad j \ .$$

After each transition, new timers are set, with the distribution of $T_{i,j}$ being the same at each transition to state i. So new timer values are set only at transition epochs. However, by the lack-of-memory property of the exponential distribution, the distribution of the remaining times $T_{i,j}$ and the associated random variables T_i and N_i would be the same any time we looked at the process in state i.

The analysis we have just done translates this clock formulation directly into a DTMC with exponential transition times, as in our first modelling approach in Subsection 3.1: The one-step transition matrix P of the DTMC is

$$P_{i,j} = P(N_i = j) = \frac{Q_{i,j}}{\sum_{k,k\neq i} Q_{i,k}} = \frac{Q_{i,j}}{\nu_i} \quad \text{for} \quad j \neq i \ , \tag{3.17}$$

with $P_{i,i} = 0$ for all i, as specified in (3.16), while the rate ν_i of the exponential holding time in state i is specified in (3.15).

Moreover, it is easy to see how to define transition rates as required for the second modelling approach. We just let $Q_{i,j}$ be the rate of the exponential timer $T_{i,j}$. We have chosen the notation to make these associations obvious. Moreover, we can use the exponential timers to prove that the transition probabilities of the CTMC are well defined and do indeed have derivatives at the origin.

The construction here makes it clear how to relate the first two modelling approaches. Given the rate matrix Q, we define the one-step transition matrix $\hat{P}$ of the DTMC by (3.17) and the rate $\hat{\nu}_i$ of the exponential transition time in state i by (3.15). That procedure gives us an underlying DTMC $\hat{P}$ with $\hat{P}_{i,i} = 0$ for all i.

These equations also tell us how to go the other way: Given (P, ν), we let

$$Q_{i,j} = \nu_i P_{i,j} \quad \text{for} \quad j \neq i \quad \text{and} \quad Q_{i,i} = -\sum_{j, j \neq i} Q_{i,j} = \nu_i \quad \text{for all} \quad i \ . \tag{3.18}$$

From this analysis, we see that the CTMC is uniquely specified by the rate matrix Q; i.e., two different Q matrices produce two different CTMC's (two different probability laws, i.e., two different f.d.d.'s). That property also holds for the first modelling approach, provided that we assume that $P_{i,i} = 0$ for all i. Otherwise, the same CTMC can be represented by different pairs (P, ν). There is only one if we require, as we have done, that there be no transitions from any state immediately back to itself.

We can also use this third modelling approach to show that the probability law of the CTMC is unaltered if there are initially one-step transitions from any state to itself. If we are initially given one-step transitions from any state to itself, we can start by removing them, but without altering the probability law of the original CTMC. If we remove a DTMC transition from state i to itself, we must compensate by increasing the transition probabilities to other states and increasing the mean holding time in state i. To do so, we first replace initial transition matrix P with transition matrix $\hat{P}$, where $\hat{P}_{i,i} = 0$ for all i. To do so without altering the CTMC, we must let the new transition probability be the old conditional probability given that there is no transition from state i to itself; i.e., we let

$$\hat{P}_{i,j} = \frac{P_{i,j}}{1 - P_{i,i}} \quad \text{for all} \quad i \quad \text{and} \quad j \ . \tag{3.19}$$

We never divide by zero, because $P_{i,i} < 1$ (assuming that the chain has more than two states and is irreducible). Since we have eliminated DTMC transitions from state i to itself, we must make the mean transition time larger to compensate. In particular, we replace $1/\nu_i$ by $1/\hat{\nu}_i$, where

$$1/\hat{\nu}_i = \frac{(1/\nu_i)}{1 - P_{i,i}} \quad \text{or} \quad \hat{\nu}_i = \nu_i (1 - P_{i,i}) \ . \tag{3.20}$$

Theorem 3.3. (removing transitions from a state back to itself) *The probability law of the CTMC is unaltered by removing one-step transitions from each state to itself, according to* (3.19) *and* (3.20)*.*

Proof. The tricky part is recognizing what needs to be shown. Since (1) the transition rates determine the transition probabilities, as shown in Subsection 3.2, (2) the transition probabilities determine the finite-dimensional distributions and (3) the finite-dimensional distributions are regarded as the probability law of the CTMC, as shown in Section 2, it suffices to show that we have the right transition rates. So that is what we show.

Applying (3.18), we see that the transition rates of the new CTMC (denoted by a hat) are

$$\hat{Q}_{i,j} \equiv \hat{\nu}_i \hat{P}_{i,j} = \nu_i (1 - P_{i,i}) \frac{P_{i,j}}{(1 - P_{i,i})} = \nu_i P_{i,j} \ , \tag{3.21}$$

just as in (3.18). ∎

In closing, we remark that this third modelling approach with independent clocks corresponds to treating the CTMC as a special case of a **generalized semi-Markov process (GSMP)**; e.g., see Glynn (1989). For general GSMP's, the clocks can run at different speeds and the timers can have nonexponential distributions.

3.4. A DTMC with Poisson Transitions

Our final modelling approach is not really a direct modelling approach, but rather an intermediate modelling approach, starting from the first modelling approach involving a DTMC with exponential transition times, that facilitates further analysis. Indeed, this modelling approach can be regarded as a special case of the first modelling approach. But it provides a different view of a CTMC.

In our first modelling approach, involving a DTMC with exponential transition times, the means of those transition times $1/\nu_i$ could vary from state to state. However, if it happened that these means were all the same, then we could represent the CTMC directly as a DTMC with transitions governed by an independent Poisson process, because in a Poisson process the times between transitions are IID exponential random variables.

Specifically, if the mean transition time is $1/\nu_0$ for all states, then we can generate all transitions from a Poisson process with rate ν_0. Let $\{Y_n : n \geq 0\}$ be the DTMC with one-step transition matrix P and let $\{N(t) : t \geq 0\}$ be an independent Poisson process with rate ν_0. Under that condition, the CTMC $\{X(t) : t \geq 0\}$ can be constructed as a **random time change** of the DTMC $\{Y_n : n \geq 0\}$ by the Poisson process $\{N(t) : t \geq 0\}$, i.e.,

$$X(t) = Y_{N(t)}, \quad t \geq 0 \ . \tag{3.22}$$

As a consequence,

$$P_{i,j}(t) \equiv P(X(t) = j|X(0) = i) = \sum_{k=0}^{\infty} P_{i,j}^k P(N(t) = k) = \sum_{k=0}^{\infty} P_{i,j}^k \frac{e^{-\nu_0 t}(\nu_0 t)^k}{k!} \ . \tag{3.23}$$

This situation may appear to be very special, but actually any finite-state CTMC can be represented in this way. We can achieve this representation by using the technique of **uniformization**, which means making the rates uniform or constant.

We make the rates uniform without changing the probability law of the CTMC by introducing one-step transitions from some states to themselves, which we can regard as **fictitious transitions**, because the process never actually moves. We can generate **potential transitions** from a Poisson process with rate λ, where λ is chosen so that

$$\nu_i \equiv -Q_{i,i} = \sum_{j,j \neq i} Q_{i,j} \leq \lambda \quad \text{for all} \quad i \ , \tag{3.24}$$

as in (3.15).

When the CTMC is in state i, each of these potential transitions is a real transition (to another state) with probability ν_i/λ, while the potential transition is a fictitious transition (a transition from state i back to state i, meaning that we remain in state i at that time) with probability $1-(\nu_i/\lambda)$, independently of past events. In other words, in each state i, we perform **independent thinning of the Poisson process** having rate λ, creating real transitions in state i according to a Poisson process having rate ν_i, just as in the original model.

The uniformization construction requires that we change the transition matrix of the embedded DTMC. The new one-step transition matrix allows transitions from a state to itself. In particular, the new one-step transition matrix $\tilde{P}$ is constructed from the CTMC transition rate matrix Q and λ satisfying (3.24) by letting

$$\tilde{P}_{i,j} = \frac{Q_{i,j}}{\lambda} \quad \text{for} \quad j \neq i \tag{3.25}$$

and

$$\tilde{P}_{i,i} = 1 - \sum_{j,j\neq i} \tilde{P}_{i,j} = 1 - \frac{\nu_i}{\lambda} = 1 + \frac{Q_{i,i}}{\lambda} = 1 - \frac{\sum_{j,j\neq i} Q_{i,j}}{\lambda} \ . \tag{3.26}$$

In matrix notation,

$$\tilde{P} = I + \lambda^{-1} Q \ . \tag{3.27}$$

Note that we have done the construction to ensure that $\tilde{P}$ is a bonafide Markov chain transition matrix; it is nonnegative with row sums 1.

Uniformization is useful because it allows us to apply properties of DTMC's to analyze CTMC's. For the general CTMC characterized by the rate matrix Q, we have transition probabilities $P_{i,j}(t)$ expressed via $\tilde{P}$ in (3.25)-(3.27) and λ as

$$P_{i,j}(t) \equiv P(X(t) = j|X(0) = i) = \sum_{k=0}^{\infty} \tilde{P}^k_{i,j} P(N(t) = k) = \sum_{k=0}^{\infty} \tilde{P}^k_{i,j} \frac{e^{-\lambda t}(\lambda t)^k}{k!} \ , \tag{3.28}$$

where $\tilde{P}$ is the DTMC transition matrix constructed in (3.25)-(3.27). We also have representation (3.22) provided that the DTMC $\{Y_n : n \geq 0\}$ is governed by the one-step transition matrix $\tilde{P}$ and the Poisson process $\{N(t) : t \geq 0\}$ has rate λ in (3.24).

But how do we know that equations (3.25) and (3.28) are really correct?

Theorem 3.4. (validity of uniformization) *The CTMC constructed via* (3.25) *and* (3.28) *leaves the probability law of the CTMC unchanged.*

Proof. We can justify the construction by showing that the transition rates are the same. Starting from (3.28), we see that, for $i \neq j$,

$$\begin{aligned} P_{i,j}(h) &= \sum_k \tilde{P}^k_{i,j} \frac{e^{-\lambda h}(\lambda h)^k}{k!} \\ &= \lambda h \tilde{P}^1_{i,j} + o(h) = \lambda h \frac{Q_{i,j}}{\lambda} + o(h) = Q_{i,j} h + o(h) \quad \text{as} \quad h \downarrow 0 \ , \end{aligned} \tag{3.29}$$

consistent with (3.2), while

$$\begin{aligned} P_{i,i}(h) - 1 &= \sum_k \tilde{P}^k_{i,i} \frac{e^{-\lambda h}(\lambda h)^k}{k!} - 1 \\ &= \tilde{P}^0_{i,i} e^{-\lambda h} + \lambda h \tilde{P}^1_{i,i} + o(h) - 1 \\ &= 1 - \lambda h + o(h) + \lambda h \left(1 + \frac{Q_{i,i}}{\lambda}\right) + o(h) - 1 \\ &= Q_{i,i} h + o(h) \quad \text{as} \quad h \downarrow 0 \ , \end{aligned} \tag{3.30}$$

consistent with (3.3). ■

We now give a full proof of Theorem 3.2, showing that the transition function $P(t)$ can be expressed as the matrix exponential e^{Qt}.

Proof of Theorem 3.2. (matrix-exponential representation) Apply (3.25) to see that $\tilde{P} = \lambda^{-1} Q + I$. Then substitute for $\tilde{P}$ in (3.28) to get

$$\begin{aligned} P(t) &= \sum_{k=0}^{\infty} \tilde{P}^k \frac{e^{-\lambda t}(\lambda t)^k}{k!} = \sum_{k=0}^{\infty} (\lambda^{-1} Q + I)^k \frac{e^{-\lambda t}(\lambda t)^k}{k!} = e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(Q + \lambda I)^k t^k}{k!} \\ &= e^{-\lambda t} e^{(Q+\lambda I)t} = e^{-\lambda t} e^{Qt} e^{\lambda t} = e^{Qt} \equiv \sum_{k=0}^{\infty} \frac{Q^k t^k}{k!} \ . \ ■ \end{aligned}$$

In the next section we will show how uniformization can be applied to quickly determine existence, uniqueness and the form of the limiting distribution of a CTMC.

4. Limiting Probabilities

Just as with DTMC's, the CTMC model specifies how the process moves locally. Just as with DTMC's, we use the CTMC model to go from the assumed local behavior to deduce global behavior. That is, we use the CTMC model to calculate its limiting probability distribution, as defined in (2.3). We then use that limiting probability distribution to answer questions about what happens in the long run. In this section we show how to compute limiting probabilities. The examples will illustrate how to apply the limiting distribution to answer other questions about what happens in the long run.

But first we want to establish a firm foundation. We will demonstrate existence and uniqueness of a limiting distribution, which justifies talking about "the" limiting distribution of an (irreducible) CTMC. We also want to show that the limiting distribution of a CTMC coincides with the (unique) stationary distribution of the CTMC. A probability vector β is a **stationary distribution** for a CTMC $\{X(t) : t \geq 0\}$ if $P(X(t) = j) = \beta_j$ for all t and j whenever $P(X(0) = j) = \beta_j$ for all j. In general the two notions - limiting distribution and stationary distribution - are distinct, but for CTMC's there is a unique probability vector with both properties.

Example 4.1. (distinction between the concepts) Before establishing positive results for CTMC's, we show that in general the two notions are distinct: there are stationary distributions that are not limiting distributions; and there are limiting distributions that are not stationary distributions.

(a) Recall that a periodic irreducible finite-state DTMC has a unique stationary probability vector, which is not a limiting probability vector; the transitions probabilities $P^k_{i,j}$ alternate as k increases, assuming a positive value at most every d steps, where d is the period of the chain. (A CTMC cannot be periodic.)

(b) To go the other way, consider a stochastic process $\{X(t) : t \geq 0\}$ with continuous state space consisting of the unit interval $[0, 1]$. Suppose that the stochastic process moves deterministically except for its initial value $X(0)$, which is a random variable taking values in $[0, 1]$. After that initial random start, let the process move deterministically on the unit interval $[0, 1]$ according to the following rules: From state 0, let the process instantaneously jump to state 1. Otherwise, let the process move according to the ODE

$$X'(t) \equiv \frac{dX(t)}{dt} = -X(t), t \geq 0 \ .$$

Then $\{X(t) : t \geq 0\}$ is a Markov process with a unique limiting distribution. In particular,

$$\lim_{t \to \infty} X(t) = 0 \quad \text{with probability } 1 \ ,$$

so that the limiting distribution is unit probability mass on 0. However, that limit distribution is not a stationary distribution. Indeed, $P(X(t) = 0) = 0$ for all $t > 0$ and all distributions of $X(0)$. If $P(X(0) = 0) = 1$, then $P(X(t) = e^{-t} \quad \text{for all} \quad t) = 1$. Even though this Markov process has a unique limiting probability distribution, there is no stationary probability vector for this Markov process. ■

But the story is very nice for irreducible finite-state CTMC's: Then there always exists a unique stationary probability vector, which also is a limiting probability vector. The situation is somewhat cleaner for CTMC's than for DTMC's, because we cannot have periodic CTMC's. That is implied by the following result.

Lemma 4.1. (positive transition probabilities) *For an irreducible CTMC, $P_{i,j}(t) > 0$ for all i, j and $t > 0$.*

Proof. The argument going forward in time is easy: By Lemma 2.1, if $P_{i,j}(s) > 0$, then

$$P_{i,j}(s+t) = \sum_k P_{i,k}(s)P_{k,j}(t) \ge P_{i,j}(s)P_{j,j}(t) \ge P_{i,j}(s)e^{Q_{j,j}t} > 0 \quad \text{for all} \quad t > 0 \ ,$$

because $P_{j,j}(t)$ is bounded below by the probability of no transition at all from state j in time t, which is $e^{Q_{j,j}t}$. (Recall that $Q_{j,j} < 0$.) More generally, we apply representation (3.28). Since the CTMC is irreducible, $P_{i,j}(t) > 0$ for some t. By representation (3.28), we thus have $\tilde{P}^k_{i,j} > 0$ for some k, implying that the embedded DTMC with transition matrix $\tilde{P}$ is irreducible. From here on, we argue by contradiction: Suppose that $P_{i,j}(t) = 0$ for some $t > 0$. Then, by representation (3.28), $\tilde{P}^k_{i,j} = 0$ for all k, which would imply that $\tilde{P}$ is reducible. Since that is a contradiction, we must have $P_{i,j}(t) > 0$ for all $t > 0$, as claimed. ■

Theorem 4.1. (existence and uniqueness) *For an irreducible finite-state CTMC, there exists a unique* ***limiting probability vector*** *α; i.e., there exists a unique probability vector α such that*

$$\lim_{t\to\infty} P_{i,j}(t) = \alpha_j \quad \textit{for all} \quad i \quad \textit{and} \quad j \ . \tag{4.1}$$

Moreover, that limiting probability vector α is the unique ***stationary probability vector****, i.e., if*

$$P(X(0) = j) = \alpha_j \quad \textit{for all} \quad j \ ,$$

then

$$P(X(t) = j) = \alpha_j \quad \textit{for all} \quad j \quad \textit{and} \quad t > 0 \ . \tag{4.2}$$

Proof. We will apply established results for DTMC's in the setting of the fourth modelling approach in Subsection 3.4; i.e., we will apply uniformization. To do so, we apply representation (3.28). From that representation and Lemma 4.1, it follows immediately that the CTMC is irreducible if and only if the embedded Markov chain with transition matrix $\tilde{P}$ is irreducible. Assuming that the CTMC is indeed irreducible, the same is true for that embedded DTMC. By making λ in (3.24) larger if necessary, we can have $\tilde{P}_{i,i} > 0$ for all i, so that the embedded DTMC with transition matrix $\tilde{P}$ can be taken to be aperiodic as well.

Given that the DTMC with transition matrix $\tilde{P}$ is irreducible and aperiodic, we know that the embedded DTMC has a unique stationary distribution $\tilde{\pi}$ satisfying

$$\tilde{\pi} = \tilde{\pi}\tilde{P} \quad \text{and} \quad \tilde{\pi}\mathbf{e} = 1 \ ,$$

with the additional property that

$$\tilde{P}^k_{i,j} \to \tilde{\pi}_j \quad \text{as} \quad k \to \infty$$

for all i and j. From representation (3.28), it thus follows that $\tilde{\pi}$ is also the limiting distribution for the CTMC; i.e., we have

$$\alpha_j = \tilde{\pi}_j \quad \text{for all} \quad j \ .$$

Here is a **detailed mathematical argument**: For any $\epsilon > 0$ given, first choose k_0 such that $|\tilde{P}^k_{i,j} - \tilde{\pi}_j| < \epsilon/2$ for all $k \ge k_0$. Then choose t_0 such that $P(N(t) < k_0) < \epsilon/4$ for all $t \ge t_0$. As a consequence, for $t > t_0$,

$$\begin{aligned} |P_{i,j}(t) - \tilde{\pi}_j| &= |P(X(t) = j|X(0) = i, N(t) < k_0) - \tilde{\pi}_j|P(N(t) < k_0) \\ &\quad + |P(X(t) = j|X(0) = i, N(t) \ge k_0) - \tilde{\pi}_j|P(N(t) \ge k_0) \\ &\le 2P(N(t) < k_0) + P(N(t) \ge k_0)\frac{\epsilon}{2} \le \frac{\epsilon}{2} + \frac{\epsilon}{2} \le \epsilon \; . \end{aligned} \tag{4.3}$$

Moreover, there can be no other stationary distribution, because any stationary distribution of the CTMC has to be coincide with the limiting distribution of the DTMC, again by (3.28). ■

We now turn to calculation. We give three different ways to calculate the limiting distribution, based on the different modelling frameworks. (We do not give a separate treatment for the competing clocks with exponential timers. We treat that case via the transition rates.) To sum row vectors in matrix notation, we right-multiply by a column vector of $1's$. Let $\mathbf{e}$ denote such a column vector of $1's$.

Theorem 4.2. (calculation)

(a) *Given a CTMC characterized as a DTMC with one-step transition matrix $\tilde{P}$ and transitions according to a Poisson process with rate λ, as in Subsection* 3.4*,*

$$\alpha_j = \tilde{\pi}_j \quad \textit{for all} \quad j \; , \tag{4.4}$$

where $\tilde{\pi}$ is the unique solution to

$$\tilde{\pi} = \tilde{\pi}\tilde{P} \quad \textit{and} \quad \tilde{\pi}\mathbf{e} = 1 \; , \tag{4.5}$$

with $\tilde{P}$ given in (3.25) *or, equivalently,*

$$\sum_i \tilde{\pi}_i \tilde{P}_{i,j} = \tilde{\pi}_j \quad \textit{for all} \quad j \quad \textit{and} \quad \sum_j \tilde{\pi}_j = 1 \; . \tag{4.6}$$

(b) *Given a CTMC characterized in terms of a DTMC with one-step transition matrix P and exponential transition times with means $1/\nu_i$, as in Subsection* 3.1*,*

$$\alpha_j = \frac{(\pi_j/\nu_j)}{\sum_k (\pi_k/\nu_k)} \; , \tag{4.7}$$

where π is the unique solution to

$$\pi = \pi P \quad \textit{and} \quad \pi\mathbf{e} = 1 \; . \tag{4.8}$$

(c) *Given a CTMC characterized by its transition-rate matrix Q, as in Subsection* 3.2*, α is the unique solution to*

$$\alpha Q = 0 \quad \textit{and} \quad \alpha\mathbf{e} = 1 \tag{4.9}$$

or, equivalently,

$$\sum_i \alpha_i Q_{i,j} = 0 \quad \textit{for all} \quad j \quad \textit{and} \quad \sum_i \alpha_i = 1 \; . \tag{4.10}$$

(d) *Given a CTMC characterized by its transition function $P(t)$, perhaps as constructed in Subsection* 3.4*, α is the unique solution to*

$$\alpha P(t) = \alpha \quad \textit{for any} \quad t > 0 \quad \textit{and} \quad \alpha\mathbf{e} = 1 \tag{4.11}$$

or, equivalently,

$$\sum_i \alpha_i P_{i,j}(t) = \alpha_j \quad \textit{for all} \quad j \quad \textit{and} \quad \sum_i \alpha_i = 1 \; . \tag{4.12}$$

Proof and Discussion. **(a) Exploiting Uniformization.** In our proof of Theorem 4.1 above, we have already shown that α coincides with $\tilde{\pi}$.

(b) Starting with the embedded DTMC. Since Theorem 4.1 establishes the existence of a unique stationary probability distribution, it suffices to show that the distribution displayed in (4.7) is that stationary distribution. Equivalently, it suffices to show that $\tilde{\pi} = \alpha$ for α in (4.7), where $\tilde{\pi}$ is the unique solution to

$$\tilde{\pi} = \tilde{\pi}\tilde{P} \quad \text{and} \quad \tilde{\pi}\mathbf{e} = 1 \ .$$

To see that is the case, observe that $\alpha_j = c\pi_j/\nu_j$ for α defined in (4.7). To show that $\alpha\tilde{P} = \alpha$, observe that

$$\begin{aligned}
(\alpha\tilde{P})_j &= \sum_i \alpha_i \tilde{P}_{i,j} = c\sum_i \frac{\pi_i}{\nu_i}\tilde{P}_{i,j} \\
&= c\left(\sum_{i,i\neq j} \frac{\pi_i}{\nu_i}\frac{Q_{i,j}}{\lambda} + \frac{\pi_j}{\nu_j}\tilde{P}_{j,j}\right) \\
&= c\left(\sum_{i,i\neq j} \frac{\pi_i}{\nu_i}\left(\frac{\nu_i}{\lambda}P_{i,j}\right) + \frac{\pi_j}{\nu_j}\left(1 - \frac{\sum_{i,i\neq j}\nu_j P_{j,i}}{\lambda}\right)\right) \\
&= c\left(\sum_{i,i\neq j} \frac{\pi_i P_{i,j}}{\lambda} + \frac{\pi_j}{\nu_j} - \frac{\pi_j\sum_{i,i\neq j} P_{j,i}}{\lambda}\right) \\
&= c\left(\sum_i \frac{\pi_i P_{i,j}}{\lambda} - \frac{\pi_j P_{j,j}}{\lambda} + \frac{\pi_j}{\nu_j} - \frac{\pi_j(1 - P_{j,j})}{\lambda}\right) \\
&= c\left(\frac{\pi_j}{\lambda} - \frac{\pi_j P_{j,j}}{\lambda} + \frac{\pi_j}{\nu_j} - \frac{\pi_j}{\lambda} + \frac{\pi_j P_{j,j}}{\lambda}\right) \\
&= c\frac{\pi_j}{\nu_j} = \alpha_j \ .
\end{aligned} \tag{4.13}$$

From (4.7), we see that $\alpha\mathbf{e} = 1$, where $\mathbf{e}$ is again a column vector of $1's$. That completes the proof.

We now give a **separate direct informal argument** (which can be made rigorous) to show that α has the claimed form. Let $Z_{i,j}$ be the time spend in state i during the j^{th} visit to state i and let $N_i(n)$ be the number of visits to state i among the first n transitions. Then the actual proportion of time spent in state i during the first n transitions, say $T_i(n)$, is

$$T_i(n) = \frac{\sum_{j=1}^{N_i(n)} Z_{i,j}}{\sum_k \sum_{j=1}^{N_k(n)} Z_{k,j}} \ , \tag{4.14}$$

However, by properties of DTMC's, $n^{-1}N_i(n) \to \pi_i$ with probability 1 as $n \to \infty$. Moreover, by the law of large numbers,

$$\frac{1}{n}\sum_{j=1}^{N_i(n)} Z_{i,j} = \left(\frac{N_i(n)}{n}\right)\left(\frac{\sum_{j=1}^{N_i(n)} Z_{i,j}}{N_i(n)}\right) \to \pi_i E[Z_{i,j}] = \pi_i/\nu_i \quad \text{as} \quad n \to \infty \ . \tag{4.15}$$

Thus, combining (4.14) and (4.15), we obtain

$$T_i(n) \to \frac{(\pi_i/\nu_i)}{\sum_k(\pi_k/\nu_k)} \quad \text{as} \quad n \to \infty \ , \tag{4.16}$$

supporting (4.7). For a full proof, we need to show that this same limit holds at arbitrary times t as $t \to \infty$. It is intuitively clear that holds, but we do not prove that directly.

(c) Starting with the transition rates. We give several different arguments, from different perspectives, to show that α is characterized as the unique solution of $\alpha Q = 0$ with $\alpha \mathbf{e} = 1$.

We first apply a **rate-conservation principle:** In steady state, the rate of transitions into state j has to equal the rate of transitions out of state j, for each state j. The steady-state rate of transitions into state j is

$$\sum_{i, i \neq j} \alpha_i Q_{i,j}$$

for the limiting probability vector α to be determined, while the steady-state rate of transitions out of state j is

$$\sum_{i, i \neq j} \alpha_j Q_{j,i} = -\alpha_j Q_{j,j} \ .$$

Setting these two steady-state rates equal yields

$$\sum_{i} \alpha_i Q_{i,j} = 0 \ ,$$

which, when applied to all j, is equivalent to $\alpha Q = 0$ in matrix notation.

Alternatively, we can **start from the ODE's.** From Theorem 4.1, we know that $P_{i,j}(t) \to \alpha_j$ as $t \to \infty$ for all i and j. Thus the right side of the backwards ODE $P'(t) = QP(t)$ converges, which implies that

$$P'_{i,j}(t) = \sum_k Q_{i,k} P_{k,j}(t) \to \sum_k Q_{i,k} \alpha_j \quad \text{as} \quad t \to \infty \ .$$

However, since $\sum_k Q_{i,k} = 0$ for all i,

$$P'_{i,j}(t) \to 0 \quad \text{as} \quad t \to \infty \quad \text{for all} \quad i \quad \text{and} \quad j \ .$$

When we apply these established limits for $P(t)$ and $P'(t)$ in the **forward ODE**, $P'(t) = P(t)Q$, we immediately obtain the desired $0 = \alpha Q$, where $\alpha \mathbf{e} = 1$.

We can instead work with the **DTMC transition matrix** $\tilde{P}$. From (3.18) and (3.25), we see that

$$Q = \lambda(\tilde{P} - I) \ . \tag{4.17}$$

Multiply on the left by α in (4.17) to get

$$\alpha Q = \lambda \left(\alpha \tilde{P} - \alpha \right) \ ,$$

which implies that $\alpha Q = 0$ if and only if $\alpha \tilde{P} = \alpha$.

(d) Starting with the transition function $P(t)$**.** This final characterization is similar to part (a). Apply the explicit expression for $P(t)$ in (3.28) with the expression $\tilde{P} = \lambda^{-1} Q + I$ to deduce that $\alpha Q = 0$ if and only if $\alpha P(t) = \alpha$. ■

To illustrate, we now return once more to Examples 3.1 and 3.2.

Example 4.2. (Pooh Bear and the Three Honey Trees Revisited) In Example 3.1 the CTMC was naturally formulated as a DTMC with exponential transition times, as in the first modelling approach in Subsection 3.1. We exhibited the DTMC transition matrix P and the mean transition times $1/\nu_i$ before. Thus it is natural to apply Theorem 4.2 (b) in order to calculate the limiting probabilities. From that perspective, the limiting probabilities are

$$\alpha_j = \frac{\pi_j(1/\nu_j)}{\sum_k \pi_k(1/\nu_k)} ,$$

where the limiting probability vector π of the discrete-time Markov chain with transition matrix P is obtained by solving $\pi = \pi P$ with $\pi \mathbf{e} = 1$, yielding

$$\pi = (\frac{8}{17}, \frac{4}{17}, \frac{5}{17}) .$$

Then final steady-state distribution, accounting for the random holding times, is

$$\alpha = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) .$$

We were then asked to find the limiting proportion of time that Pooh spends at each of the three trees. Those limiting proportions coincide with the limiting probabilities. That can be demonstrated by applying the renewal-reward theorem from renewal theory, specifically, Theorem **??**. ∎

Example 4.3. (Copier Maintenance Revisited Again) Let us return to Example 3.2 and consider the question posed there: What is the long-run proportion of time that each copier is working and what is the long-run proportion of time that the repairman is busy? To have concrete numbers, suppose that the failure rates are $\gamma_1 = 1$ per month and $\gamma_2 = 3$ per month; and suppose the repair rates are $\beta_1 = 2$ per month and $\beta_2 = 4$ per month.

We first substitute the specified numbers for the rates γ_i and β_i in the rate matrix Q in (4.3), obtaining

$$Q = \begin{array}{c} 0 \\ 1 \\ 2 \\ (1,2) \\ (2,1) \end{array} \left(\begin{array}{ccccc} -4 & 1 & 3 & 0 & 0 \\ 2 & -5 & 0 & 3 & 0 \\ 4 & 0 & -5 & 0 & 1 \\ 0 & 0 & 2 & -2 & 0 \\ 0 & 4 & 0 & 0 & -4 \end{array} \right) .$$

Then we solve the system of linear equations $\alpha Q = 0$ with $\alpha \mathbf{e} = 1$, which is easy to do with a computer and is not too hard by hand. Just as with DTMC's, one of the equations in $\alpha Q = 0$ is redundant, so that with the extra added equation $\alpha \mathbf{e} = 1$, there is a unique solution. Performing the calculation, we see that the limiting probability vector is

$$\alpha \equiv (\alpha_0, \alpha_1, \alpha_2, \alpha_{(1,2)}, \alpha_{(2,1)}) = \left(\frac{44}{129}, \frac{16}{129}, \frac{36}{129}, \frac{24}{129}, \frac{9}{129} \right) .$$

Thus, the long-run proportion of time that copier 1 is working is $\alpha_0 + \alpha_2 = 80/129 \approx 0.62$, while the long-run proportion of time that copier 2 is working is $\alpha_0 + \alpha_1 = 60/129 \approx 0.47$. The long-run proportion of time that the repairman is busy is $\alpha_1 + \alpha_2 + \alpha_{(1,2)} + \alpha_{(2,1)} = 1 - \alpha_0 = 85/129 \approx 0.659$,

Now let us consider **an alternative repair strategy**: Suppose that copier 1 is more important than copier 2, so that it is more important to have it working. Toward that end, suppose the repairman always work on copier 1 when both copiers are down. In particular, now suppose that the repairman stops working on copier 2 when it is down if copier 1 also subsequently fails, and immediately shifts his attention to copier 1, returning to work on copier 2 after copier 1 has been repaired. How do the long-run proportions change?

With this alternative repair strategy, we can revise the state space. Now it does suffice to use 4 states, letting the state correspond to the set of failed copiers, because now we know what the repairman will do when both copiers are down; he will always work on copier 1. Thus it suffices to use the single state $(1,2)$ to indicate that both machines have failed. There now is only one possible transition from state $(1,2)$: $Q_{(1,2),2} = \mu_1$. We display the revised rate diagram in Figure 2 below.

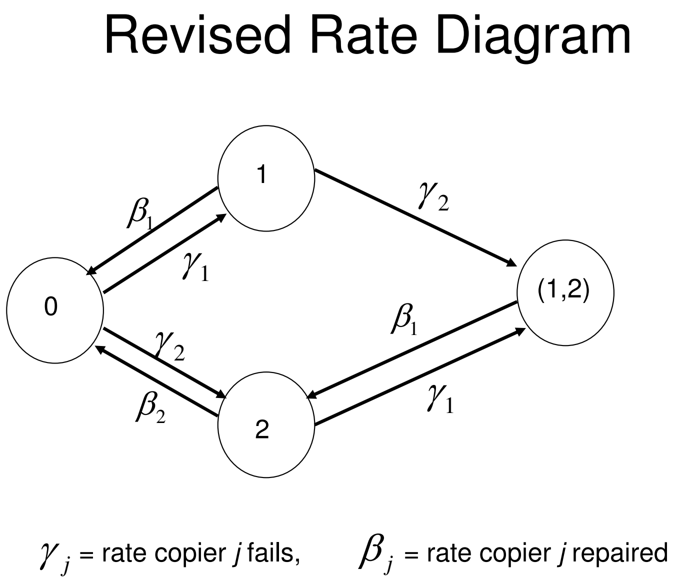

Figure 2: A revised rate diagram showing the transition rates among the 4 states in Example 3.3, where the repairman always works on copier 1 first when both have failed.

The associated rate matrix is now

$$Q = \begin{array}{c} 0 \\ 1 \\ 2 \\ (1,2) \end{array} \left(\begin{array}{cccc} -(\gamma_1+\gamma_2) & \gamma_1 & \gamma_2 & 0 \\ \beta_1 & -(\gamma_2+\beta_1) & 0 & \gamma_2 \\ \beta_2 & 0 & -(\gamma_1+\beta_2) & \gamma_1 \\ 0 & 0 & \beta_1 & -\beta_1 \end{array} \right)$$

or, with the numbers assigned to the parameters,

$$Q = \begin{array}{c} 0 \\ 1 \\ 2 \\ (1,2) \end{array} \left(\begin{array}{cccc} -4 & 1 & 3 & 0 \\ 2 & -5 & 0 & 3 \\ 4 & 0 & -5 & 1 \\ 0 & 0 & 2 & -2 \end{array} \right) .$$

Just as before, we obtain the limiting probabilities by solving $\alpha Q = 0$ with $\alpha \mathbf{e} = 1$. Now we obtain

$$\alpha \equiv (\alpha_0, \alpha_1, \alpha_2, \alpha_{(1,2)}) = \left(\frac{20}{57}, \frac{4}{57}, \frac{18}{57}, \frac{15}{57}\right) .$$

Thus, the long-run proportion of time that copier 1 is working is $\alpha_0 + \alpha_2 = 38/57 = 2/3 \approx 0.67$, while the long-run proportion of time that copier 2 is working is $\alpha_0 + \alpha_1 = 24/57 \approx 0.42$. The new strategy has increased the long-run proportion of time copier 1 is working from 0.62 to 0.67, at the expense of decreasing the long-run proportion of time copier 2 is working from 0.47 to 0.42. The long-run proportion of time the repairman is busy is $1 - \alpha_0 = 37/57 \approx 0.649$, which is very slightly less than before.

We conclude by making some **further commentary**. We might think that the revised strategy is wasteful, because the repairman quits working on copier 2 when copier 1 fails after copier 2 previously failed. By shifting to work on copier 1, we might think that the repairman is being inefficient, "wasting" his expended effort working on copier 2, making it more likely that both copiers will remain failed. In practice, under other assumptions, that might indeed be true, but here because of the lack-of-memory property of the exponential distribution, the expended work on copier 2 has no influence on the remaining required repair times. From a pure efficiency perspective, it might be advantageous to give one of the two copiers priority at this point, but not because of the expended work on copier 2. On the other hand, we might prefer the original strategy from a "fairness" perspective. In any case, the CTMC model lets us analyze the consequences of alternative strategies. As always, the relevance of the conclusions depends on the validity of the model assumptions. But even when the model assumptions are not completely realistic or not strongly verified, the analysis can provide insight. ▪

5. Birth-and-Death Processes

Many CTMC's have transitions that only go to neighboring states, i.e., either up one or down one; they are called birth-and-death processes. Motivated by population models, a transition up one is called a **birth**, while a transition down one is called a **death**. The birth rate in state i is denoted by λ_i, while the death rate in state i is denoted by μ_i. The rate diagram for a birth-and-death process (with state space $\{0, 1, \ldots, n\}$) takes the simple linear form shown in Figure 3.

Thus, for a birth-and-death process, the CTMC transition rates take the special form

$$Q_{i,i+1} = \lambda_i, \quad Q_{i,i-1} = \mu_i \quad \text{and} \quad Q_{i,j} = 0 \quad \text{if} \quad j \notin \{i-1, i, i+1\}, \quad 1 \le i \le n-1 , \tag{5.1}$$

with

$$Q_{0,1} = \lambda_0, \quad Q_{0,j} = 0 \quad \text{if} \quad j \notin \{0, 1\}, \quad Q_{n,n-1} = \mu_n \quad \text{and} \quad Q_{n,j} = 0 \quad \text{if} \quad j \notin \{n-1, n\} . \tag{5.2}$$

As before, the row sums of Q are zero.

A further special case is a **pure-birth process**, which only has transitions up one (equivalently, all death rates are 0). We have already encountered a special pure-birth process (on the nonnegative integers) - the Poisson process - which has constant birth rate, i.e., $\lambda_i = \lambda$ for all i. Similarly, a pure-death process has only transitions down one. For a finite state space, a pure-death process is equivalent to a pure-birth process, because we can just relabel the states.

The special structure of a birth-and-death process makes it easier to calculate the limiting probabilities. First, we observe that the global-balance equations (flow into state j equals flow out of state j), captured by the equation $\alpha Q = 0$, can be replaced by more elementary detailed-balance equations.

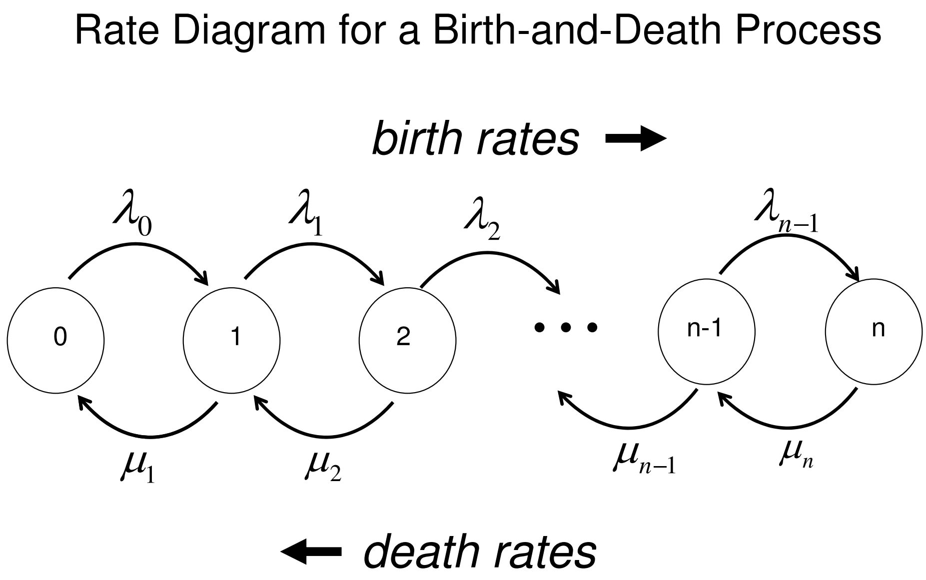

Figure 3: A rate diagram showing the transition rates for a birth-and-death process.

Theorem 5.1. (detailed-balance equations) *For a birth-and-death process, the limiting probability vector α is the unique solution to the **detailed-balance equations***

$$\alpha_j \lambda_j = \alpha_{j+1}\mu_{j+1} \quad \textit{for all} \quad j, \quad 0 \le j \le n-1 \ , \tag{5.3}$$

with $\alpha \mathbf{e} = 1$.

Proof. We give two different proofs: First, just as for a general CTMC, we can apply a **rate-conservation principle**, but now in a more special form. Because the birth-and-death process can move only to neighboring states, we can deduce that the steady-state rate of transitions up from state j to state $j+1$, $\alpha_j \lambda_j$, must equal the steady-state rate of transitions down from state $j+1$ to state j, $\alpha_{j+1}\mu_{j+1}$. That is, there is rate conservation between any two neighboring states. That yields the detailed-balance equations in (5.3). The rate-conservation principle itself follows from a simple observation. In the time interval $[0,t]$, the number of transitions up from state j to state $j+1$ can differ by at most by one from the number of transitions down from state $j+1$ to state j.

From the perspective of a general CTMC, we already have established that α is the unique solution to $\alpha Q = 0$ with $\alpha \mathbf{e} = 1$. For a birth-and-death process, the j^{th} equation in the system $\alpha Q = 0$ is

$$(\alpha Q)_j = \alpha_{j-1}\lambda_{j-1} - \alpha_j(\lambda_j + \mu_j) + \alpha_{j+1}\mu_{j+1} = 0 \quad \text{for} \quad 1 \le j \le n-1 \ , \tag{5.4}$$

with

$$(\alpha Q)_0 = -\alpha_0\lambda_0 + \alpha_1\mu_1 = 0 \quad \text{and} \quad (\alpha Q)_n = \alpha_{n-1}\lambda_{n-1} - \alpha_n\mu_n = 0 \ . \tag{5.5}$$

From (5.3)-(5.5), we see that the sum of the first j equations of the form (5.4) and (5.5) from $\alpha Q = 0$ coincides with the j^{th} detailed-balance equation in (5.3), while the difference between the j^{th} and $(j-1)^{\text{st}}$ detailed-balance equations coincides with the j^{th} equation from (5.4) and (5.5). Hence the two characterizations are equivalent. ∎

In fact, it is not necessary to solve a system of linear equations each time we want to calculate the limiting probability vector α, because we can analytically solve the detailed-balance equations to produce an explicit formula.

Theorem 5.2. (limiting probabilities) *For a birth-and-death process with state space* $\{0, 1, \ldots, n\}$,

$$\alpha_j = \frac{r_j}{\sum_{k=0}^{n} r_k} \quad 0 \le j \le n \ , \tag{5.6}$$

where

$$r_0 = 1 \quad \text{and} \quad r_j = \frac{\lambda_0 \times \lambda_1 \times \cdots \times \lambda_{j-1}}{\mu_1 \times \mu_2 \times \cdots \times \mu_j} \ . \tag{5.7}$$

Proof and Discussion. By virtue of Theorem 5.1, it suffices to solve the detailed-balance equations in (5.3). We can do that **recursively**:

$$\alpha_j = \frac{\lambda_{j-1}}{\mu_j}\alpha_{j-1} \quad \text{for all} \quad j \ge 1 \ ,$$

which implies that

$$\alpha_j = r_j \alpha_0 \quad \text{for all} \quad j \ge 1 \ ,$$

for r_j in (5.7). We obtain the final form (5.6) when we require that $\alpha \mathbf{e} = 1$. ∎

Example 5.1. (a small barbershop) Consider a small barbershop, where there are only two barbers, each with his own barber chair. Suppose that there is only room for at most 5 customers, with 2 in service and 3 waiting. Assume that potential customers arrive according to a Poisson process at rate $\lambda = 6$ per hour. Customers arriving when the system is full are **blocked and lost**, leaving without receiving service and without affecting future arrivals. Assume that the duration of each haircut is an independent exponential random variable with a mean of $\mu^{-1} = 15$ minutes. Customers are served in a first-come first-served manner by the first available barber.

We can ask a variety of **questions**: (a) What is the long-run proportion of time there are two customers in service plus two customers waiting? (b) What is the (long-run) proportion of time each barber is busy? We might then go on to ask how this long-run behavior would change if we changed the number of barbers or the number of waiting spaces.

We start by **constructing the model**. Let $Q(t)$ denote the number of customers in the system at time t. Then the stochastic process $\{Q(t) : t \ge 0\}$ is a birth-and-death process with six states: $0, 1, 2, 3, 4, 5$. Indeed, this is a standard **queueing model**, commonly referred to as **the M/M/2/3 queue**. The first M means a Poisson arrival process (M for Markov), the second M means IID exponential service times (M again for Markov), the 2 is for 2 servers, and the 3 is for 3 additional waiting spaces.) It is common to use λ to denote the arrival rate and μ the service rate of each server.

We can represent the CTMC in terms of competing exponential timers, as in Subsection 3.3. The possible triggering events are an arrival (birth), causing the state to go up 1, or a departure (death), causing the state to go down 1. It is of course important that these are independent exponential random variables.

The blocking alters the arrival process. The blocking means that no arrivals can enter in state 5. By making the state space $\{0, 1, \ldots, 5\}$, we have accounted for the blocking. Since the interarrival times of a Poisson process have an exponential distribution, there are active clocks with exponential timers corresponding to the event of a new arrival in states 0-4. The arrival clock in state i has mean $1/\lambda_i = 1/\lambda$, where $\lambda = 6$ per hour is the arrival rate of the Poisson process. Hence the birth rates are $\lambda_i = 6$, $0 \leq i \leq 4$. We have $\lambda_5 = 0$, because there are no arrivals when the system is full.

Since the service times are independent exponential random variables, the active clocks corresponding to departures also can be represented as exponential random variables. (Recall that the minimum of independent exponential variables is again exponential with a rate equal to the sum of the rates.) There are active clocks with exponential timers corresponding to the event of a new departure in states 1-5. The departure clock in state i has mean $1/\mu_i$, where μ_i is the death rate to be determined. Since the mean service time is $1/\mu = 15$ minutes, the service rate for each barber is $\mu = 1/15$ per minute or $\mu = 4$ per hour. However, we must remember that the service rate applies to each server separately. Since we are measuring time in hours, the death rates are $\mu_1 = \mu = 4$, $\mu_2 = \mu_3 = \mu_4 = \mu_5 = 2\mu = 8$. We have $\mu_0 = 0$ since there can be no departures from an empty system.

Given the birth rates and death rates just defined, we can draw the rate diagram for the six states $0, 1, \ldots, 5$, as in Figure 3. The associated rate matrix is now

$$Q = \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} -6 & 6 & 0 & 0 & 0 & 0 \\ 4 & -10 & 6 & 0 & 0 & 0 \\ 0 & 8 & -14 & 6 & 0 & 0 \\ 0 & 0 & 8 & -14 & 6 & 0 \\ 0 & 0 & 0 & 8 & -14 & 6 \\ 0 & 0 & 0 & 0 & 8 & -8 \end{pmatrix}$$

We now can apply Theorem 5.2 to calculate the **limiting probabilities**. From (5.6) and (5.7),

$$\alpha_j = \frac{r_j}{r_0 + r_1 + \cdots r_5}, \quad 0 \leq j \leq 5 \,,$$

where $r_0 = 1$ and

$$r_j = \frac{\lambda_0 \times \lambda_1 \times \cdots \times \lambda_{j-1}}{\mu_1 \times \mu_2 \times \cdots \times \mu_j} \,, \quad 1 \leq j \leq 5 \,.$$

Here

$$\begin{aligned} r_0 &= 1 = \frac{512}{512} \\ r_1 &= \frac{6}{4} = \frac{3}{2} = \frac{768}{512} \\ r_2 &= \frac{6 \times 6}{4 \times 8} = \frac{36}{32} = \frac{9}{8} = \frac{576}{512} \\ r_3 &= \frac{6 \times 6 \times 6}{4 \times 8 \times 8} = \frac{27}{32} = \frac{432}{512} \\ r_4 &= \frac{27 \times 6}{32 \times 8} = \frac{81}{128} = \frac{324}{512} \\ r_5 &= \frac{81 \times 6}{128 \times 8} = \frac{243}{512} \end{aligned}$$

Hence,

$$\alpha_0 = \frac{512}{2855}, \quad \alpha_1 = \frac{768}{2855}, \quad \alpha_2 = \frac{576}{2855}, \quad \alpha_3 = \frac{432}{2855},$$

$$\alpha_4 \quad = \quad \frac{324}{2855} \quad \text{and} \quad \alpha_5 = \frac{243}{2855} \ .$$

This particular calculation is admittedly a bit tedious, but it is much better than solving the system of equations based on $\alpha Q = 0$, which would be required for a general CTMC.

Given the steady-state probability vector $\alpha \equiv (\alpha_0, \ldots, \alpha_5)$, you can then **answer the questions posed**: (a) The long-run proportion of time there are two customers in service plus two customers waiting is α_4. (b) The (long-run) proportion of time each barber is busy is $(1/2)\alpha_1 + \alpha_2 + +\alpha_3 + \alpha_4 + \alpha_5$. (The two barbers are each busy half of the time that only one barber is busy.) Finally, we could see how these answers would change if we changed the number of barbers or the number of waiting spaces. We would just perform a similar analysis with the alternative model(s). ∎

Example 5.2. (customers that balk and abandon) Consider the same barbershop with two barbers and three waiting spaces, as specified above, but in addition suppose that customers may elect to balk or abandon. In particular, suppose that an arriving customer finding both barbers busy, but an available waiting space, will elect to stay, independently of all past events, with probability 2/3; otherwise, the arrival will **balk**, i.e., refuse to join and instead immediately leave, without affecting future arrivals. Moreover, suppose that each arriving customer who is not blocked and who elects to wait is only willing to wait a certain time before starting service; otherwise the customer will **abandon**, i.e., leave without receiving service and without affecting future arrivals. Let the amount of patience of successive customers, i.e., these successive times to abandon, be IID exponential random variables with mean $\theta^{-1} = 10$ minutes.

Again we can ask a variety of **questions**: (a) What is the rate of customer abandonment? (b) What is the long-run proportion of potential arrivals that enter and then abandon? (c) What proportion of potential customers enter upon arrival (i.e., neither balk nor are blocked)? (d) What proportion of potential customers are served?

Even though it may not be entirely evident at first, the stochastic process representing the number of customers in the system over time is again a birth-and-death process. Again we can represent the CTMC in terms of competing exponential timers, as in Subsection 3.3. The possible triggering events are an arrival (birth), causing the state to go up 1, or a departure (death), causing the state to go down 1, where the departure may be due to service completion or abandonment. As noted before, the blocking means that no arrivals can enter in state 5. The balking alters the arrival process further. The balking corresponds to performing an independent thinning of the external Poisson arrival process in states 2-4. In those states, the actual arrivals form a Poisson process with arrival rate $\lambda \times (2/3) = 6 \times (2/3) = 4$ per hour. Since the interarrival times of a Poisson process have an exponential distribution, there are active clocks with exponential timers corresponding to the event of a new arrival in states 0-4. The arrival clock in state i has mean $1/\lambda_i$, where λ_i is the birth rate to be determined. The birth rates in these states are: $\lambda_0 = \lambda_1 = \lambda = 6$ per hour and $\lambda_2 = \lambda_3 = \lambda_4 = 6 \times (2/3) = 4$ per hour. (The reduction is due to the balking. We have $\lambda_5 = 0$, because there are no arrivals when the system is full.)

Since the service times and times to abandon are independent exponential random variables, the active clocks corresponding to departures also can be represented as exponential random variables. As before, there are active clocks with exponential timers corresponding to the event of a new departure in states 1-5. The departure clock in state i has mean $1/\mu_i$, where μ_i is the death rate to be determined. As before, the service rate for each barber is $\mu = 4$ per hour. Since the mean time to abandon is $1/\theta = 10$ minutes for each customer, the individual abandonment rate is $\theta = 1/10$ per minute or 6 per hour. However, we must remember that the service rate applies to each server separately, while the abandonment rate applies to each waiting customer separately. Thus the death rates are $\mu_1 = \mu = 4$, $\mu_2 = 2\mu = 8$, $\mu_3 = 2\mu + \theta = 8 + 6 = 14$, $\mu_4 = 2\mu + 2\theta = 8 + 12 = 20$, $\mu_5 = 2\mu + 3\theta = 8 + 18 = 26$. ($\mu_0 = 0$.)

Given the new birth rates and death rates just defined, we can draw the new rate diagram for the six states $0, 1, \ldots, 5$, as in Figure 3. The new rate matrix is

$$Q = \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} -6 & 6 & 0 & 0 & 0 & 0 \\ 4 & -10 & 6 & 0 & 0 & 0 \\ 0 & 8 & -12 & 4 & 0 & 0 \\ 0 & 0 & 14 & -18 & 4 & 0 \\ 0 & 0 & 0 & 20 & -24 & 4 \\ 0 & 0 & 0 & 0 & 26 & -26 \end{pmatrix}$$

We now can apply Theorem 5.2 to calculate the **limiting probabilities**. From (5.6),

$$\alpha_i = \frac{r_i}{r_0 + r_1 + \cdots r_5}, \quad 0 \le i \le 5 \; ,$$

where $r_0 = 1$ and

$$r_i = \frac{\lambda_0 \times \lambda_1 \times \cdots \times \lambda_{i-1}}{\mu_1 \times \mu_2 \times \cdots \times \mu_i} \; , \quad 1 \le i \le 5 \; .$$

Here

$$\begin{aligned} r_0 &= 1 = \frac{3640}{3640} \\ r_1 &= \frac{6}{4} = \frac{3}{2} = \frac{5460}{3640} \\ r_2 &= \frac{6 \times 6}{4 \times 8} = \frac{36}{32} = \frac{9}{8} = \frac{4095}{3640} \\ r_3 &= \frac{6 \times 6 \times 4}{4 \times 8 \times 14} = \frac{36}{112} = \frac{9}{28} = \frac{1170}{3640} \\ r_4 &= \frac{9 \times 4}{28 \times 20} = \frac{9}{140} = \frac{234}{3640} \\ r_5 &= \frac{9 \times 4}{140 \times 26} = \frac{18}{1820} = \frac{9}{910} = \frac{36}{3640} \end{aligned}$$

Hence,

$$\begin{aligned} \alpha_0 &= \frac{3640}{14635}, \quad \alpha_1 = \frac{5460}{14635}, \quad \alpha_2 = \frac{4095}{14635}, \quad \alpha_3 = \frac{1170}{14635}, \\ \alpha_4 &= \frac{234}{14635} \quad \text{and} \quad \alpha_5 = \frac{36}{14635} \; . \end{aligned}$$

Given the steady-state probability vector $\alpha \equiv (\alpha_0, \ldots, \alpha_5)$, you can then **answer the other questions**: (a) The rate of customer abandonments is $\theta\alpha_3 + 2\theta\alpha_4 + 3\theta\alpha_5 = 6\alpha_3 + 12\alpha_4 + 18\alpha_5$. (b) The long-run proportion of potential customers that enter and abandon is the rate customers abandon, just determined, divided by the arrival rate, i.e.,

$$\frac{(\theta\alpha_3 + 2\theta\alpha_4 + 3\theta\alpha_5)}{\lambda} = \frac{(6\alpha_3 + 12\alpha_4 + 18\alpha_5)}{6} = \alpha_3 + 2\alpha_4 + 3\alpha_5 \; .$$

Questions (c) and (d) are more tricky, because they ask about the proportion of customers having a specified experience, instead of the long-run proportion of time. However, it turns out that these notions agree in this problem, because the arrival process of potential customers is a Poisson process. There is a principle called **Poisson Arrivals See Time Averages (PASTA)** that implies that the proportion of customers that see some state upon arrival coincides with the proportion of time the process spends in that state, provided that the arrival process is Poisson (and other regularity conditions hold, which they do here; e.g., see Section 5.16 of Wolff (1989), Melamed and Whitt (1990) or Stidham and El Taha (1999)). Hence, consistent with intuition, the long-run proportion of all potential customers that are blocked coincides with the long-run proportion of time that the system is full, which is α_5. (But that property would not remain true if we made the arrival process non-Poisson.) Similarly, the long-run proportion of customers that balk is 1/3 times the long-run proportion of time that the system is in one of the states 2, 3 or 4, which is $(1/3) \times (\alpha_2 + \alpha_3 + \alpha_4)$. (c) Hence the long-run proportion of potential customers enter upon arrival (i.e., neither balk nor are blocked) is $1 - (1/3) \times (\alpha_2 + \alpha_3 + \alpha_4) - \alpha_5$.

(d) We can find the long-run proportion of potential customers served in two different ways:

Method 1. The long-run proportion of customers served is 1 minus the sum of the proportions that balk, abandon and are blocked. We can thus apply the answers to previous questions. The answer is

$$1 - \frac{\alpha_2 + \alpha_3 + \alpha_4}{3} - \alpha_5 - (\alpha_3 + 2\alpha_4 + 3\alpha_5) \ ,$$

Rewriting, we get

$$1 - \frac{\alpha_2}{3} - \frac{4\alpha_3}{3} - \frac{7\alpha_4}{3} - 4\alpha_5 = \frac{11,020}{14,635} = 0.753 \ .$$

Method 2. The long-run proportion of customers served can be represented as the overall service completion rate divided by the external arrival rate, counting all potential arrivals. The denominator - the arrival rate - is $\lambda = 6$ per hour. The service completion rate is

$$(\alpha_1 \times 4) + (\alpha_2 + \alpha_3 + \alpha_4 + \alpha_5) \times 8 \ ,$$

because the service rate is 4 in state 1, while the service rate is $2 \times 4 = 8$ in states $2-5$. Hence, the long-run proportion of customers served is

$$\frac{2\alpha_1}{3} + \frac{4(\alpha_2 + \alpha_3 + \alpha_4 + \alpha_5)}{3} = \frac{11,020}{14,635} \ .$$

Even though the two formulas are different, they give the same answer.

Finally, just as before, we could see how these answers would change if we changed the number of barbers or the number of waiting spaces. We would just perform a similar analysis with the alternative model(s). ■

In many applications it is natural to use birth-and-death processes with **infinite state spaces**. As in other mathematical settings, we primarily introduce infinity because it is more convenient. With birth-and-death processes, an infinite state space often simplifies the form of the limiting probability distribution. We illustrate by giving a classic queueing example.

Example 5.3. (the M/M/1 Queue) One of the most elementary queueing models is the $M/M/1$ queue, which has a single server and unlimited waiting room. As with the $M/M/s/r$ model considered in Example 5.1 (with $S = 2$ and $r = 3$), customers arrive in a Poisson process with rate λ and the service times are IID exponential random variables with mean $1/\mu$. The number of customers in the system at time t as a function of t, say $Q(t)$, is then a birth-and-death process. However, since there is unlimited waiting room, the state space is infinite.

With an infinite state space, we must guard against pathologies. In order to have a proper stationary distribution (which coincides with the limiting distribution), it is necessary to require that the arrival rate λ be less than the maximum possible rate out, μ. Equivalently, we require that the traffic intensity $\rho \equiv \lambda/\mu$ be strictly less than 1.

When we apply the extension of Theorem 5.2 to infinite state spaces, under the assumption that $\rho < 1$, we get

$$\alpha_j = \frac{r_j}{\sum_{k=0}^{\infty} r_k}, \quad \text{where} \quad r_j = \rho^j, \quad j \geq 0 \;, \tag{5.8}$$

which implies that α is the **geometric distribution**; i.e.,

$$\lim_{t\to\infty} P(Q(t) = j | Q(0) = i) = \alpha_j = (1-\rho)\rho^j, \quad j \geq 0 \;, \tag{5.9}$$

which has mean $\rho/(1-\rho)$.

If instead we considered the $M/M/1/r$ model (the $M/M/1$ model with a finite waiting room), which has 1 server and r extra waiting spaces, then the birth-and-death process has $r+2$ states, from 0 to $r+1$. The limiting distribution then becomes the **truncated geometric distribution**:

$$\alpha_j = \frac{(1-\rho)\rho^j}{(1-\rho^{r+2})}, \quad 0 \leq j \leq r+1 \;. \tag{5.10}$$

The geometric distribution in (5.9) is more appealing than the truncated geometric distribution in (5.10) because of its cleaner form. However, the finite-waiting room model applies without constraint on ρ; a proper limiting distribution exists for $\rho \geq 1$ as well as for $\rho < 1$. ■

6. Reverse-Time CTMC's and Reversibility

Just as for DTMC's, an important concept for CTMC's is reversibility. A stochastic process $\{X(t) : -\infty < t < \infty\}$ is said to be **reversible** if it has the same probability law as $\{X(-t) : -\infty < t < \infty\}$. Thus a CTMC is reversible if we get the same CTMC if we run time backwards.

Just as for DTMC's, we can start by constructing the **reverse-time CTMC** associated with a given CTMC with transition-rate matrix Q and transition function $P(t)$. We obtain the reverse-time Markov chain from the original (forward) CTMC by reversing time. The reverse-time transition probabilities describe where the process came from instead of where it is going. That is, we let

$$\overleftarrow{P}_{i,j}(t) \equiv P(X(s) = j | X(s+t) = i) \;. \tag{6.1}$$

We can then then apply basic properties of conditional probabilities to express these reverse-time transition probabilities in terms of given forward-time transition probabilities; i.e.,

$$\overleftarrow{P}_{i,j}(t) \equiv P(X(s) = j | X(s+t) = i) = \frac{P(X(s) = j, X(s+t) = i)}{P(X(s+t) = i)} = \frac{P(X(s) = j) P_{j,i}(t)}{P(X(s+t) = i)} \;. \tag{6.2}$$

Unfortunately, however, when we do this, we see that in general we do not obtain bonafide Markov transition probabilities; if we sum over j in (6.2), the transition probabilities do not necessarily sum to 1, as required. We need to assume more. What we assume in addition is that **the Markov chain is in equilibrium**. That is, we assume that the given (forward) CTMC with transition function $P(t)$ is irreducible with initial distribution equal to its stationary distribution α. Then $P(X(t) = j) = \alpha_j$ for all t. Under this extra equilibrium condition, equation (6.2) can be expressed as

$$\overleftarrow{P}_{i,j}(t) \equiv P(X(s) = j | X(s+t) = i) = \frac{P(X(s) = j, X(s+t) = i)}{P(X(s+t) = i)} = \frac{\alpha_j P_{j,i}(t)}{\alpha_i} \ . \tag{6.3}$$

Theorem 6.1. (reverse-time CTMC) *If an irreducible CTMC is put in equilibrium by letting its initial distribution be its stationary distribution α, then the transition probabilities in* (6.3) *are bonafide transition probabilities, with the same stationary probability distribution α, and the stochastic process satisfies the reverse-time Markov property*

$$\begin{aligned} P(X(s) = j | X(s+t) = i, X(u) = i_u, u \in A_u \subseteq (s+t, \infty)) &= P(X(s) = j | X(s+t) = i) \\ &\equiv \overleftarrow{P}_{i,j}(t) \ . \end{aligned} \tag{6.4}$$

The reverse-time Markov property in (6.4) says that the conditional probability of a "past" state j at time s, given the "present" state i at time $s+t$ plus the states at "future" times u in the set A_u depends only on the present state i at time $s+t$.

Proof. First it is clear that the alleged transition probabilities in (6.3) are nonnegative. Since $\alpha P(t) = \alpha$ for each t, by Theorem 4.2 (d), the row sums of $\overleftarrow{P}$ now do indeed sum to 1, as required. To see that the Markov property in (6.4) does indeed hold, apply properties of conditional probabilities to rewrite (6.4) as

$$\begin{aligned} & P(X(s) = j | X(s+t) = i, X(u) = i_u, u \in A_u \subseteq (s+t, \infty)) \\ & = \frac{P(X(s) = j, X(s+t) = i, X(u) = i_u, u \in A_u \subseteq (s+t, \infty))}{P(X(s+t) = i, X(u) = i_u, u \in A_u \subseteq (s+t, \infty))} \\ & = \frac{P(X(s) = j) P_{j,i}(t) P(X(u) = i_u, u \in A_u \subseteq (s+t, \infty) | X(s+t) = i)}{P(X(s+t) = i) P(X(u) = i_u, u \in A_u \subseteq (s+t, \infty) | X(s+t) = i)} \\ & = \frac{P(X(s) = j) P_{j,i}(t)}{P(X(s+t) = i)} = \frac{\alpha_j P_{j,i}(t)}{\alpha_i} \ , \end{aligned}$$

where in the last line we have first cancelled the common term $P(X(u) = i_u, u \in A_u \subseteq (s+t, \infty) | X(s+t) = i)$ from the numerator and the denominator and then exploited the equilibrium property. ▪

From (6.3), we see right away (by looking at the derivative at 0) that the reverse-time CTMC associated with a CTMC having transition-rate matrix Q itself has transition-rate matrix $\overleftarrow{Q}$, where

$$\overleftarrow{Q}_{i,j} = \frac{\alpha_j Q_{j,i}}{\alpha_i} \ . \tag{6.5}$$

Note that any irreducible CTMC in equilibrium (initialized with its stationary distribution α) has an associated reverse-time CTMC with transition-rate matrix $\overleftarrow{Q}$. But that does not make the CTMC reversible. That is a stronger property: A CTMC is said to be time-reversible or just **reversible** if the reverse-time CTMC coincides with the original CTMC in equilibrium, i.e., if $\overleftarrow{Q} = Q$ or, equivalently, if

$$\alpha_i Q_{i,j} = \alpha_j Q_{j,i} \quad \text{for all} \quad i \quad \text{and} \quad j \ . \tag{6.6}$$

For CTMC's, reversibility as defined at the beginning of this section is equivalent to equations (6.6). From a rate-conservation perspective, reversibility holds for a CTMC if and only if the steady-state rate of transitions from state i to state j equals the steady-state rate of transitions from state j to state i for all states i and j. Thus reversibility is characterized by the **detailed-balance equations** in (6.6), which are generalizations of the detailed-balance equations for birth-and-death processes in (5.3). We summarize these properties in the following theorem.

Theorem 6.2. (reversibility of CTMC's) *A CTMC with transition rate matrix Q is reversible with stationary probability vector α if and only if the detailed balance equations in* (6.6) *hold.*

Proof. We have seen that the detailed balance equations in (6.6) imply reversibility, given that α is the stationary vector. Summing those equations on either i or j gives $\alpha Q = 0$, implying that α must in fact be the stationary probability vector. ■

As a consequence, we immediately have the following result.

Theorem 6.3. (reversibility of birth-and-death processes) *All birth-and-death processes are reversible.*

By the statement in Theorem 6.3, we mean that the birth-and-death process is reversible *provided that it has a stationary probability vector α and is initialized by that stationary probability vector α.*

We now observe that reversibility is inherited by truncation. We say that a CTMC with state space S and rate matrix Q is *truncated* to the subset $A \subset S$ if we disallow all transitions out of the subset A; i.e., if we set $Q_{i,j} = 0$ for $i \in A$ and $j \in S - A$. We obtain a new CTMC with state space A by letting $Q^{(A)}_{i,j} = Q_{i,j}$ for all i and j in A with $i \neq j$ and $Q^{(A)}_{i,i} = -\sum_{j \in A} Q_{i,j}$ for all $i \in A$.

Theorem 6.4. (truncation) *If a reversible CTMC with trate matrix Q and stationary probability vector α is truncated to a subset A, yielding the rate matrix $Q^{(A)}$ defined above, and remains irreducible, then the truncated CTMC with the rate matrix $Q^{(A)}$ is also reversible and has stationary probability vector*

$$\alpha^{(A)}_j = \frac{\alpha_j}{\sum_{k \in A} \alpha_k}, \quad \textit{for all} \quad j \in A \; . \tag{6.7}$$

Proof. It is elementary that the truncated CTMC with the probability vector $\alpha^{(A)}$ in (6.7) satisfies the detailed balance equations in (6.6) if the original CTMC does, which holds by Theorem 6.2. ■

We have seen an instance of Theorem 6.3 when we looked at the $M/M/1/r$ queueing model in (5.10), following Example 5.3.

We now apply reversibility to get something new and interesting. For that, we consider the $M/M/s$ queue with s servers, unlimited waiting room, Poisson arrival process with arrival rate λ and IID exponential service times with mean $1/\mu$. The number in system over time is a birth-and-death process with constant birth rate λ and death rates $\mu_j = \min\{j, s\}\mu$. Since there is an infinite state space, we require that $\rho \equiv \lambda/s\mu < 1$ in order to have a proper limiting distribution.

Theorem 6.5. (departures from an M/M/s queue) *The departure process from an $M/M/s$ queue in equilibrium* (*with $\rho \equiv \lambda/s\mu < 1$*) *is a Poisson process with departure rate equal to the arrival rate λ.*

Proof. Having $\rho \equiv \lambda/s\mu < 1$ ensures that a proper limiting probability vector α exists. Put the system in equilibrium by letting the initial distribution be that limiting distribution. By Theorem 6.3, the CTMC is reversible. Thus, in equilibrium, the process counting the number of customers in the system at any time in reverse time has the same probability law as the original process. However, reversing time changes departures (jumps down) into arrivals (jumps up) and vice versa. So the departures must form a Poisson process with rate λ. ■

Theorem 6.5 is quite remarkable. It takes some effort to even directly show that the time between successive departures in an $M/M/1$ queue in equilibrium has an exponential distribution with mean $1/\lambda$. That is an instructive exercise.

We can do more: We can establish an even more surprising result. Let $Q(t)$ be the number in system at time t, either waiting or being served, and let $D(t)$ count the number of departures in the interval $[0,t]$. We can show that the distribution of the queue length at time t is independent of the departures prior to time t, for the $M/M/s$ queue in equilibrium!

Theorem 6.6. (more about departures from an M/M/s queue) *For an $M/M/s$ queue in equilibrium (with $\rho \equiv \lambda/s\mu < 1$), the number in system at time t, $Q(t)$, is independent of $\{D(s) : 0 \le s \le t\}$, the departure process before time t.*

Proof. Just as for Theorem 6.5, having $\rho \equiv \lambda/s\mu < 1$ ensures that a proper limiting probability vector α exists. As before, put the system in equilibrium by letting the initial distribution be the limiting distribution. By Theorem 6.3, the CTMC is reversible. Thus, in equilibrium, the process in reverse time has the same probability law as the original process. With any one ordering of time, the arrival process after time t is independent of the queue length at time t. by the independent-increments property of the Poisson process. Since arrivals and departures switch roles when we reverse time, we deduce the asserted conclusion as well. ■

We now go on to consider networks of queues. We first combine Theorems 6.5 and 6.6 with Example 5.3 to obtain the limiting distribution for the number of customers at each station for two single-server queues in series. In particular, consider an $M/M/1$ model with arrival rate λ and service rate μ_1, where $\rho_1 \equiv \lambda/\mu_1 < 1$. Let all departures from this $M/M/1$ queue proceed next to a second single-server queue with unlimited waiting room and IID exponential service times with individual service rate μ_2, where also $\rho_2 \equiv \lambda/\mu_2 < 1$. This model is often referred to as the $M/M/1 \rightarrow /M/1$ tandem queue. Let $Q_i(t)$ be the number of customers at station i at time t, either waiting or being served.

Theorem 6.7. (the limiting probabilities for then M/M/1 → /M/1 tandem queue) *For the $M/M/1 \rightarrow /M/1$ tandem queue in equilibrium (with $\rho_i \equiv \lambda/s\mu_i < 1$ for each i), the departure processes from the two queues are Poisson processes with rate λ and*

$$\lim_{t\to\infty} P(Q_1(t) = j_1, Q_2(t) = j_2) = (1-\rho_1)\rho_1^{j_1}(1-\rho_2)\rho_2^{j_2} \quad \textit{for all} \quad j_1 \quad \textit{and} \quad j_2 \ . \tag{6.8}$$

Theorem 6.7 concludes that the limiting probabilities for the two queues in series are the same as for two independent $M/M/1$ queue, each with the given arrival rate λ. (However, the two stochastic processes $\{Q_1(t) : t \ge 0\}$ and $\{Q_2(t) : t \ge 0\}$. starting in equilibrium, are not independent. The result is for a single time t.) We say that the limiting distribution has **product form**. That product form means that the two marginal distributions are independent.

Proof. Suppose that we initialize the system with the alleged limiting probability distribution. Since it is the product of two geometric distributions, the two marginal distributions are independent. Hence we can first focus on the first queue. By Theorem 6.5, the departure process from this first queue is Poisson with rate λ. Hence, the second queue by itself is an $M/M/1$ queue in equilibrium. Hence each queue separately has the displayed geometric stationary distribution, which coincides with its limiting distribution. Now, considering the system in equilibrium, by Theorem 6.6, at any time t, the random number $Q_1(t)$ is independent of the departure process from the first queue up to time t. That implies that $Q_2(t)$ and $Q_1(t)$ must be independent for each t, which implies the product-form limiting distribution in (6.8). ■

Note that the entire $\mathbf{M/M/1} \rightarrow /\mathbf{M/1}$ tandem queue is not itself reversible; it is possible to go from state (i, j) to state $(i-1, j+1)$ (with a departure from queue 1), but it is not possible to go back. So the detailed-balance conditions in (6.6) cannot hold. We established Theorem 6.7 by exploiting the reversibility of only the first station by itself. However, there is an alternative way to prove Theorem 6.7 exploiting only reverse-time CTMC's, which has other applications, e.g., to treat networks of Markovian queues that are not acyclic (do not have flow in one direction only).

Theorem 6.8. (using detailed-balance equations to find limiting probabilities) *Let Q be the transition-rate matrix of an irreducible CTMC. If we can find numbers α_j and $\overleftarrow{Q}_{i,j}$ such that*

$$\alpha_i Q_{i,j} = \alpha_j \overleftarrow{Q}_{j,i} \quad \textit{for all} \quad i \neq j \tag{6.9}$$

and

$$-Q_{i,i} \equiv \sum_{j, j \neq i} Q_{i,j} = \sum_{j, j \neq i} \overleftarrow{Q}_{i,j} \equiv - \overleftarrow{Q}_{i,i} \quad \textit{for all} \quad i \;, \tag{6.10}$$

then $\overleftarrow{Q}_{i,j}$ are the transition rates for the reverse-time CTMC associated with Q and α is the limiting probability vector for both CTMC's.

Proof. Add over i with $i \neq j$ in (6.9) and apply (6.10) to get

$$\sum_{i, i \neq j} \alpha_i Q_{i,j} = \sum_{i, i \neq j} \alpha_j \overleftarrow{Q}_{j,i} = \alpha_j \sum_{i, i \neq j} Q_{j,i} \;,$$

which implies that $\alpha Q = 0$. Hence α is the limiting distribution for the CTMC with transition-rate matrix Q. Consequently, $\overleftarrow{Q}_{i,j}$ are the transition rates associated with the reverse-time CTMC based on Q in equilibrium. That implies (by summing on j in (6.9)) that α is the limiting distribution for the reverse-time CTMC as well. ■

We now apply Theorem 6.8 to give an alternative proof of Theorem 6.7, which again has the advantage that it does not require directly solving the equation $\alpha Q = 0$.

Alternative proof of Theorem 6.7. As for many harder problems, the first step is to guess the form of the limiting distribution; i.e., we guess that α has the product form in (6.8). We then guess that the reverse-time CTMC should be itself a $M/M/1 \rightarrow /M/1$ tandem queue with arrival rate λ and the given service rates. Going forward from state (i, j), we have three possible transitions: (1) to state $(i+1, j)$ due to an arrival, (2) to state $(i-1, j+1)$ due to departure from queue 1 and (3) to $(i, j-1)$ due to departure from queue 2. We have possible flows in the other direction for the reverse-time CTMC to state (i, j) in three possible ways: (1) from state $(i+1, j)$ due to a departure from queue 1 (in original order), (2) from state $(i-1, j+1)$ due to departure from queue 2 and (3) to $(i, j-1)$ due to an arrival from outside at queue 2. We thus have three equations to check in order to verify (6.9):

$$(1-\rho_1)\rho_1^i(1-\rho_2)\rho_2^j\lambda \quad = \quad (1-\rho_1)\rho_1^{i+1}(1-\rho_2)\rho_2^j\mu_1$$

$$
\begin{aligned}
(1-\rho_1)\rho_1^i(1-\rho_2)\rho_2^j\mu_1 &= (1-\rho_1)\rho_1^{i-1}(1-\rho_2)\rho_2^{j+1}\mu_2 \\
(1-\rho_1)\rho_1^i(1-\rho_2)\rho_2^j\mu_2 &= (1-\rho_1)\rho_1^{i}(1-\rho_2)\rho_2^{j-1}\lambda \ ,
\end{aligned} \tag{6.11}
$$

which are easily seen to be satisfied. It is also easy to see that (6.10) holds. For a state (i, j) with $i > 0$ and $j > 0$, the total rate out to a new state is $\lambda + \mu_1 + \mu_2$, corresponding to the possibilities of an arrival or a service completion at one of the two queues. For the state $(0, j)$ with $j > 0$, the total rate out to a new state is $\lambda + \mu_2$ in both cases, excluding the possibility of a service completion from queue 1 in both cases. The case $(i, 0)$ is similar. For the state $(0, 0)$, the total rate out to new states is λ in both cases, corresponding to an arrival. ∎

Essentially the same argument applies to treat a general **Markovian open network of single-server queues**. Here is the new model: Let there be m single-server queues, each with unlimited waiting room. Let there be an external Poisson arrival process at queue i with rate $\lambda_{e,i}$. Let the service times at queue i be exponential with mean $1/\mu_i$. Let the arrival processes be mutually independent. Let all the service times be mutually independent and independent of the arrival processes. Let there be **Markovian routing**, independent of the arrival and service processes; i.e., let each customer, immediately after completing service at queue i, go next to queue j with probability $P_{i,j}$, independently of all previous events. Let each customer depart from the network from queue i with probability $1 - \sum_{j=1}^{m} P_{i,j}$. If we include outside of the network as a single state $m + 1$, then the routing is characterized by a DTMC. In this routing DTMC we assume that $P_{m+1,m+1} = 1$, making the outside state absorbing. Moreover, we assume that all other states are transient states. That is, we assume that each arriving customer will eventually depart from the system.

Consider the vector valued process $Q(t) \equiv (Q_1(t), Q_2(t), \ldots, Q_m(t))$, where $Q_i(t)$ is the number of customers at queue i, either waiting or being served, at time t. It is easy to see that the stochastic process $\{Q(t) : t \geq 0\}$ is a CTMC. The possible events are an arrival from outside or a service completion at one of the two queues. Those all governed by the specified rates. We will show that this CTMC also has a product-form limiting distribution, under regularity conditions. Given the possibility of feedback now, it is even more remarkable that the marginal distributions of the limiting probability distribution should be independent.

To characterize the limiting behavior, we first need to find the **total arrival rate** at each queue, i.e., the sum of the external and internal arrival rates. In order to find the total arrival rate to each queue, we need to solve a system of linear equations, **the traffic-rate equations**:

$$
\lambda_j = \lambda_{e,j} + \sum_{i=1}^{m} \lambda_i P_{i,j} \quad \text{for} \quad 1 \leq j \leq m \ , \tag{6.12}
$$

or, equivalently, in matrix notation,

$$
\Lambda = \Lambda_e + \Lambda P \ , \tag{6.13}
$$

which implies that

$$
\Lambda = \Lambda_e (I - P)^{-1} \ . \tag{6.14}
$$

The inverse $(I-P)^{-1}$ is the fundamental matrix of the absorbing routing DTMC. In Section **??** it was shown that the inverse is well defined.

In order for the solution of the traffic-rate equations to be valid total arrival rates, we have to be sure that the net arrival rate is less than the maximum possible service rate at each queue. As before, let the **traffic intensity** at queue j be

$$
\rho_j = \frac{\lambda_j}{\mu_j}, \quad 1 \leq j \leq m \ . \tag{6.15}
$$

We assume that $\rho_j < 1$ for all j, $1 \le j \le m$.

With those assumptions, the CTMC $\{Q(t) : t \ge 0\}$ has a product-form limiting distribution.

Theorem 6.9. (Markovian open network of single-server queues) *Consider the Markovian open network of single server queues defined above, where the inverse $(I - P)^{-1}$ is well defined for the routing matrix P and $\rho_i < 1$ for each i. Then the limiting distribution has product form, with geometric marginals of the $M/M/1$ queue; i.e.,*

$$\lim_{t\to\infty} P(Q(t) = (j_1, \ldots, j_m)) = \alpha(j_1, \ldots, j_m) = \prod_{k=1}^{m} (1 - \rho_k)\rho_k^{j_k} \ . \tag{6.16}$$

Proof. A direct proof is to guess that the solution if of product form, as in (6.16), and then simply verify that $\alpha Q = 0$. That verification step is simplified by applying Theorem 6.8. To do so, we need to define the candidate reverse-time CTMC with transition rates $\overleftarrow{Q}$. Just as with Theorem 6.7, we guess that the reverse-time CTMC itself corresponds to an open network of single-server queues, with the same service-time distributions at the queues, but we need to guess the appropriate external arrival rates $\overleftarrow{\lambda}_{e,i}$ and routing probabilities $\overleftarrow{P}_{i,j}$. The idea is to guess those quantities by seeing what is required to have the flow rates balance in equilibrium.

First, the flow rate through each queue should be the same in either direction, so we should have

$$\overleftarrow{\lambda}_i = \lambda_i \quad \text{for all} \quad i \ . \tag{6.17}$$

Next, the reverse-time external arrival rate at queue i should be equal to the forward-time departure rate from queue i, i.e.,

$$\overleftarrow{\lambda}_{e,i} = \lambda_i(1 - \sum_{j=1}^{m} P_{i,j}) \quad \text{for all} \quad i \ . \tag{6.18}$$

To complete the construction, note that the stationary flow from queue i to queue j in forward time should equal the stationary reverse flow in reverse time, i.e.,

$$\lambda_i P_{i,j} = \overleftarrow{\lambda}_j \overleftarrow{P}_{j,i} \ . \tag{6.19}$$

As a consequence, we have

$$\overleftarrow{P}_{j,i} = \frac{\lambda_i P_{i,j}}{\overleftarrow{\lambda}_j} \ . \tag{6.20}$$

Combining equations (6.17), (6.18) and (6.20), we have defined the reverse-time model elements $\overleftarrow{\lambda}_{e,i}$, $\overleftarrow{\lambda}_i$ and $\overleftarrow{P}_{j,i}$ for all i and j in terms of the corresponding forward-time modelling elements.

For the reverse-time queueing network, we should have an analog of the traffic-rate equations in (6.21) in reverse time, namely,

$$\overleftarrow{\lambda}_j = \overleftarrow{\lambda}_{e,j} + \sum_{i=1}^{m} \overleftarrow{\lambda}_i \overleftarrow{P}_{i,j} \quad \text{for} \quad 1 \le j \le m \ , \tag{6.21}$$

or, equivalently, in matrix notation,

$$\overleftarrow{\Lambda} = \overleftarrow{\Lambda}_e + \overleftarrow{\Lambda} \overleftarrow{P} \ . \tag{6.22}$$

And, indeed, it is easy to check that these reverse-time traffic rate equations are valid, by applying the definitions above.

It then remains to verify that these guesses yield the right answer; i.e., we need to verify equations (6.9) and (6.10) in this setting, remembering that the states now correspond to m-tuples $(i_1, \ldots, i_m)$. That is straightforward, paralleling the proof of Theorem 6.7. As before, all transitions are triggered by arrivals and service completions (followed by a random routing). ■

From the reverse-time construction just completed, we also can deduce the following corollary.

Corollary 6.1. (departure processes from an open network of queues) *Under the assumptions of Theorem 6.9, the m departure processes from the network from the individual queues are m independent Poisson processes, with the departure process from queue i having rate*

$$\delta_i \equiv \lambda_i (1 - \sum_{j=1}^{m} P_{i,j}) \ . \tag{6.23}$$

Moreover, the total departure process is a Poisson process with rate

$$\delta \equiv \sum_{j=1}^{m} \delta_j = \sum_{j=1}^{m} \lambda_{e,j} \ . \tag{6.24}$$

Proof. For (6.24), recall that the superposition of independent Poisson processes is Poisson with a rate equal to the sum of the rates. To see that the total rate in equals the total rate out, as we would expect, compare the sum over i of (6.23) to the sum over j of (6.21). ■

In closing, we remark that Theorem 6.9 and Corollary 6.1 extend to Markovian open networks of **multi-server queues**, where there may be different numbers of servers at each queue. Again there is a product-form limiting distribution, but then the marginals have the limiting distribution of the $M/M/s$ queue, where s is the number of servers at that queue. Indeed, what we have presented is only the beginning of a rich theory of stochastic networks; e.g., see Chen and Yao (2001), Kelly (1979), Serfozo (1999), van Dijk (1993), Walrand (1988), Whittle (1986).

7. Time-Dependent Transition Probabilities

Example 7.1. (A two-state CTMC) In the special case of a two-state CTMC it is easy to calculate the transition probabilities. Let the states be 0 and 1. To simplify the notation, let $Q_{0,1} = \lambda$ and $Q_{1,0} = \mu$. Then $Q_{0,0} = -\lambda$ and $Q_{1,1} = -\mu$. With that notation, we can express two of the four ODE's as

$$P'_{0,0}(t) = -\lambda P_{0,0}(t) + \lambda P_{1,0}(t) \tag{7.1}$$

and

$$P'_{1,0}(t) = \mu P_{0,0}(t) - \mu P_{0,1}(t). \tag{7.2}$$

Multiplying equation (7.1) by μ and equation (7.2) by λ and then adding, we obtain

$$\mu P'_{0,0}(t) + \lambda P'_{1,0}(t) = 0 \ . \tag{7.3}$$

Integrating equation (7.3) from 0 to t yields

$$\mu P_{0,0}(t) + \lambda P_{1,0}(t) - \mu P_{0,0}(0) - \lambda P_{1,0}(0) = 0 \ .$$

Since $P_{0,0}(0) = 1$ and $P_{1,0}(0) = 0$, we obtain

$$\mu P_{0,0}(t) + \lambda P_{1,0}(t) = \mu \ .$$

Substituting this into equation (7.1), we obtain

$$P'_{0,0}(t) = \mu - (\mu + \lambda) P_{0,0}(t) \ . \tag{7.4}$$

We can eliminate the constant term if we consider

$$f(t) = P_{0,0}(t) - \frac{\mu}{\lambda + \mu} \ ,$$

because

$$f'(t) = P'_{0,0}(t) = \mu - (\mu + \lambda) P_{0,0}(t) = -(\lambda + \mu) f(t) \ ,$$

which implies first that

$$\frac{f'(t)}{f(t)} = -(\lambda + \mu)$$

and then

$$f(t) = C e^{-(\lambda+\mu)t} \ ,$$

for some constant C. Continuing, we get

$$P(t) = \begin{pmatrix} P_{0,0}(t) & P_{0,1}(t) \\ P_{1,0}(t) & P_{1,1}(t) \end{pmatrix} = \begin{pmatrix} \frac{\mu}{(\lambda+\mu)} + \frac{\lambda}{(\lambda+\mu)} e^{-(\lambda+\mu)t} & \frac{\lambda}{(\lambda+\mu)} - \frac{\lambda}{(\lambda+\mu)} e^{-(\lambda+\mu)t} \\ \frac{\mu}{(\lambda+\mu)} - \frac{\mu}{(\lambda+\mu)} e^{-(\lambda+\mu)t} & \frac{\lambda}{(\lambda+\mu)} + \frac{\mu}{(\lambda+\mu)} e^{-(\lambda+\mu)t} \end{pmatrix} \ .$$

From the specific form of the transition function, we can see the limit and the rate of convergence:

$$P_{i,0}(t) \to \frac{\mu}{\lambda + \mu} \quad \text{as} \quad t \to \infty \quad \text{for each} \quad i$$

and

$$P_{0,0}(t) - \frac{\mu}{\lambda + \mu} = \frac{\lambda}{(\lambda + \mu)} e^{-(\lambda+\mu)t} \quad \text{for all} \quad t \ .$$

Consistent with intuition, $P_{0,0}(t)$ decreases steadily down from $P_{0,0}(0) = 1$ to the limiting steady-state value $\mu/(\lambda + \mu)$ as t increases.

8. Issues

strong Markov property
infinite state spaces
infinitely many transitions in finite time
PASTA, used in BD barbershop example
transition probabilities for BD processes - Keilson
subset property of reversibility
closed Jackson networks
first passage times - inversion approach

NOTE: Put renewal-reward theorem in early chapter on SLLN and CLT. But leave hard renewal theory for later.

References

[1] Chen, H. and D. D. Yao. 2001. *Fundamentals of Queueing Networks*, Springer.

[2] Chung, K. L. 1967. *Markov Chains with Stationary Transition Probabilities*, second edition, Springer.

[3] El Taha, M. and S. Stidham, Jr. 1999. *Sample-Path Analysis of Queueing Systems*, Kluwer.

[4] Glynn, P. W. 1989. A GSMP formalism for discrete-event systems. *Proceedings IEEE* 77, 14–23.

[5] Kelly, F. P. 1979. *Reversibility and Stochastic Networks*, Wiley.

[6] Keilson, J. 1979. *Markov Chain Models – Rarity and Exponentiality*, Springer.

[7] Melamed, B. and W. Whitt. 1990. On arrivals that see time averages. *Operations Research* 38, 156-172.

[8] Moler, C. and C. Van Loan. 2003. Nineteen dubious ways to compute the exponetial of a matrix, twenty-five years later. *SIAM Review* 45, 3–49.

[9] Serfozo, R. F. 1999. *Introduction to Stochastic Networks*, Springer.

[10] Simmons, G. F. 1991. *Differential Equations with Applications and Historical Notes*, second edition, McGraw-Hill.

[11] van Dijk, N. 1993. *Queueing Networks and Product Forms*, Wiley.

[12] Walrand, J. 1988. *Introduction to Queueing Networks*, Prentice-Hall.

[13] Whittle, P. 1986. *Systems in Stochastic Equilibrium*, Wiley.

[14] Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*, Prentice-Hall.