25

Continuous-Time Markov Chains - Introduction

Prior to introducing continuous-time Markov chains today, let us start off with an example involving the Poisson process. Our particular focus in this example is on the way the properties of the exponential distribution allow us to proceed with the calculations. This will give us a good starting point for considering how these properties can be used to build up more general processes, namely continuous-time Markov chains.

Example: (Ross, p.338 #48(a)). Consider an n-server parallel queueing system where customers arrive according to a Poisson process with rate λ, where the service times are exponential random variables with rate μ, and where any arrival finding all servers busy immediately departs without receiving any service. If an arrival finds all servers busy, find

(a) the expected number of busy servers found by the next arrival.

Solution: Let T_k denote the expected number of busy servers found by the next arrival for a k-server system when there are currently k servers busy. Equivalently, let it denote the expected number of busy servers found by the next arrival if there are currently k servers busy. The two descriptions of T_k are equivalent because of the memoryless property of the exponential service and interarrival times and because between the current time and the time of the next arrival we can ignore the $n-k$ idle servers when considering the expected number of busy servers found by the next arrival.

First, T_0 is clearly 0 because if there are currently 0 busy servers the next arrival will find 0 busy servers for sure. Next, consider T_1. If there is currently 1 busy server, the next arrival finds 1 busy server if the time to the next arrival is less than the remaining service time of the busy server. By memorylessness, the time to the next arrival is Exponential(λ) and the remaining service time is Exponential(μ). Therefore, the probability that the next arrival finds 1 server busy is $\lambda/(\lambda+\mu)$, and

$$T_1 = (1)\frac{\lambda}{\lambda+\mu} + (0)\frac{\mu}{\lambda+\mu} = \frac{\lambda}{\lambda+\mu}.$$

In general, consider the situation where we currently have k servers busy. We can obtain an expression for T_k by conditioning on what happens first. Let us see how the properties of the exponential distribution allow us to proceed with this argument. When there are currently k servers busy, we have $k+1$ independent exponential "alarm clocks" going: k Exponential(μ) clocks, one for each remaining service time, and 1 Exponential(λ) clock for the time till the next arrival. For our purposes we wish to condition on whether a service completion happens first or the next arrival happens first. The time till

the next service completion is the minimum of the k Exponential(μ) clocks, and this has an Exponential($k\mu$) distribution. Thus, the probability that the next thing to happen is a service completion is the probability that an Exponential($k\mu$) random variable is less than an Exponential(λ) random variable, and this probability is $k\mu/(k\mu+\lambda)$. Similarly, the probability that the next thing to happen is the next arrival is $\lambda/(k\mu+\lambda)$.

Now, if the first thing to happen is the next customer arrival, then the expected number of busy servers found by the next arrival is k. On the other hand, if the first thing to happen is a service completion, then the expected number of busy servers found by the next arrival is T_{k-1}.

The reason this latter conditional expectation is given by T_{k-1}, and really the main thing I wish you to understand in this example, is that the memorylessness of the exponential interarrival time and all the exponential service times allows us to say that once we have conditioned on the first thing to happen being a service completion, then we can essentially *restart* the exponential clock on the interarrival time and the exponential clocks on the $k-1$ service times still going. Thus, *probabilistically* we are in exactly the conditions defining T_{k-1}.

We have

$$T_k = T_{k-1}\frac{k\mu}{k\mu+\lambda} + k\frac{\lambda}{k\mu+\lambda}.$$

Solving for T_n is now a matter of solving the recursion given by the above expression.

Starting with T_2, we have

$$\begin{aligned} T_2 &= T_1 \frac{2\mu}{2\mu+\lambda} + \frac{2\lambda}{2\mu+\lambda} \\ &= \left(\frac{\lambda}{\mu+\lambda}\right)\left(\frac{2\mu}{2\mu+\lambda}\right) + \frac{2\lambda}{2\mu+\lambda}. \end{aligned}$$

Continuing (since the pattern isn't so obvious yet),

$$\begin{aligned} T_3 &= T_2 \frac{3\mu}{3\mu+\lambda} + \frac{3\lambda}{3\mu+\lambda} \\ &= \left(\frac{\lambda}{\mu+\lambda}\right)\left(\frac{2\mu}{2\mu+\lambda}\right)\left(\frac{3\mu}{3\mu+\lambda}\right) + \left(\frac{2\lambda}{2\mu+\lambda}\right)\left(\frac{3\mu}{3\mu+\lambda}\right) + \frac{3\lambda}{3\mu+\lambda}. \end{aligned}$$

In general, we can observe the following patterns for T_n:

- T_n will be a sum of n terms.
- The ith term will be a product of $n+1-i$ factors.
- the ith term will have a factor $i\lambda/(i\mu+\lambda)$ for $i = 1, \ldots, n$.
- The ith term will have $n-i$ remaining factors that are given by $(i+1)\mu/((i+1)\mu+\lambda), \ldots, n\mu/(n\mu+\lambda)$, for $i = 1, \ldots, n-1$, while the nth term has no remaining factors.

Based on these observations, we can write

$$T_n = \frac{n\lambda}{n\mu+\lambda} + \sum_{i=1}^{n-1} \frac{i\lambda}{i\mu+\lambda} \prod_{j=i+1}^{n} \frac{j\mu}{j\mu+\lambda}$$

as our final expression. □

We saw in the last example one way to think about how the process which keeps track of the number of busy servers evolves, based on the exponential service and interarrival times. We make the following observations.

(i) When there are i busy servers (at any time), for $i < n$, there are $i + 1$ independent exponential alarm clocks running, with i of them having rate μ and 1 of them having rate λ. The time until the process makes a jump is exponential whose rate is the sum of all the competing rates: $i\mu + \lambda$. If there are n busy servers then only the n exponential clocks corresponding to the service times can trigger a jump, and the time until the process makes a jump is exponential with rate $n\mu$.

(ii) When the process jumps from state i, for $i < n$, it jumps to state $i + 1$ with probability $\lambda/(i\mu + \lambda)$ and jumps to state $i - 1$ with probability $i\mu/(i\mu + \lambda)$. If there are n busy servers the process jumps to state $n - 1$ with probability $n\mu/n\mu = 1$.

(iii) When the process makes a jump from state i we can start up a whole new set of clocks corresponding to the state we jumped to. This is because even though some of the old clocks that had been running before we made our jump but did not actually trigger the current jump might still trigger the next jump, we can either reset these clocks or, equivalently, replace them with new clocks.

Note that every time we jump to state i, regardless of when the time is, the distribution of how long we stay in state i and the probabilities of where we jump to next when we leave state i are the same. In other words, the process is *time-homogeneous*.

We may generalize the preceding process which tracks the number of busy servers in our opening example in a fairly straightforward manner. First, we can generalize the state space $\{0, 1, \ldots, n\}$ in that example to any arbitrary countable state space S. In addition, we can generalize (i), (ii) and (iii) on the preceding page to the following:

(I) Every time the process is in state i there are n_i independent exponential clocks running, such that the first one to go off determines the state the process jumps to next. Let the rates of these n_i exponential clocks be $q_{i,j_1}, \ldots, q_{i,j_{n_i}}$, such that $j_1, \ldots, j_{n_i}$ are the n_i states that the process can possibly jump to next. The time until the process makes a jump is exponential with rate $v_i \equiv q_{i,j_1} + \ldots + q_{i,j_{n_i}}$.

(II) When the process jumps from state i, it jumps to state j_ℓ with probability $q_{i,j_\ell}/(q_{i,j_1} + \ldots + q_{i,j_{n_i}}) = q_{i,j_\ell}/v_i$, for $\ell = 1, \ldots, n_i$.

(III) When the process makes a jump from state i we can start up a whole new set of clocks corresponding to the state we jumped to.

The above description of a continuous-time stochastic process corresponds to a continuous-time Markov chain. This is not how a continuous-time Markov chain is defined in the text (which we will also look at), but the above description is equivalent to saying the process is a time-homogeneous, continuous-time Markov chain, and it is a more revealing and useful way to think about such a process than the formal definition given in the text.

Example: *The Poisson Process.* The Poisson process is a continuous-time Markov chain with state space $S = \{0, 1, 2, \ldots\}$. If at any time we are in state i we can only possibly jump to state $i + 1$ when we leave state i and there is a single exponential clock running that has rate $q_{i,i+1} = \lambda$. The time until we leave state i is exponential with rate $v_i = q_{i,i+1} = \lambda$. When the process leaves state i, it jumps to state $i + 1$ with probability $q_{i,i+1}/v_i = v_i/v_i = 1$. □

Example: *Pure Birth Processes.* We can generalize the Poisson process by replacing $q_{i,i+1} = \lambda$ with $q_{i,i+1} = \lambda_i$. Such a process is called a *pure birth process*, or just *birth process*. The state space is the same as that of a Poisson process, $S = \{0, 1, 2, \ldots\}$. If at any time the birth process is in state i there is a single exponential clock running with rate λ_i, and so $v_i = \lambda_i$. We see that the only difference between a Poisson process and a pure birth process is that in the pure birth process the rate of leaving a state can depend on the state. □

Example: *Birth/Death Processes.* A birth/death process generalizes the pure birth process by allowing jumps from state i to state $i-1$ in addition to jumps from state i to state $i+1$. The state space is typically the set of all integers or a subset of the integers, but varies depending on the particular modeling scenario. We can make the state space a proper subset of the integers by making the rates of any jumps that go out of the subset equal to 0. Whenever a birth/death process is in state i there are *two* independent exponential clocks running, one that will take us to state $i + 1$ if it goes off first and the other which will take us to state $i - 1$ if it goes off first. Following the text, we denote the rates of these clocks by $q_{i,i+1} = \lambda_i$ (the *birth rates*) and $q_{i,i-1} = \mu_i$ (the *death rates*), and $v_i = \lambda_i + \mu_i$. This important class of processes is the subject of all of Section 6.3 of the text. □

Example: *The n-Server Parallel Queueing System.* We can see that the n-server parallel queueing system described in our opening example is a birth/death process. The state space is $S = \{0, 1, \ldots, n\}$. When in state i, for $i = 0, \ldots, n-1$, the birth rate is λ and when in state n the birth rate is 0. When in state i, for $i = 0, \ldots, n$, the death rate is $i\mu$. That is, this process is a birth/death process with

$$\begin{aligned} \lambda_i &= \lambda \quad \text{for } i = 0, 1, \ldots, n-1, \\ \lambda_n &= 0, \\ \mu_i &= i\mu \quad \text{for } i = 0, 1, \ldots, n. \end{aligned}$$

The main thing I would like you to focus on in this lecture is the description of a continuous-time stochastic process with countable state space S given in (I), (II) and (III). Imagine a particle jumping around the state space as time moves forward according to the mechanisms described there.

Next we will formally define a continuous-time Markov chain in terms of the Markov property for continuous-time processes and see how this corresponds to the description given in (I), (II) and (III).

26

Continuous-Time Markov Chains - Introduction II

Our starting point for today is the description of a continuous-time stochastic process discussed in the previously. Specifically, the process can have any countable state space S. With each state $i \in S$ there is associated as set of n_i independent, exponential alarm clocks with rates $q_{i,j_1}, \ldots, q_{i,j_{n_i}}$, where $j_1, \ldots, j_{n_i}$ is the set of possible states the process may jump to when it leaves state i. We have seen that when the process enters state i, the amount of time it spends in state i is Exponentially distributed with rate $v_i = q_{i,j_1} + \ldots + q_{i,j_{n_i}}$ and when it leaves state i it will go to state j_ℓ with probability $q_{i,j_\ell}/v_i$ for $\ell = 1, \ldots, n_i$.

We also stated previously that any process described by the above probabilistic mechanisms corresponds to a continuous-time Markov chain. We will now elaborate on this statement in more detail. We start by defining the Markov property for a continuous-time process, which leads to the formal definition of what it means for a stochastic process to be a continuous-time Markov chain.

The Markov Property for Continuous-Time Processes:

You should be familiar and comfortable with what the Markov property means for discrete-time stochastic processes. The natural extension of this property to continuous-time processes can be stated as follows. For a continuous-time stochastic process $\{X(t) : t \geq 0\}$ with state space S, we say it has the *Markov property* if

$$\begin{aligned} &P(X(t) = j | X(s) = i, X(t_{n-1}) = i_{n-1}, \ldots, X(t_1) = i_1) \\ &= P(X(t) = j | X(s) = i), \end{aligned}$$

where $0 \leq t_1 \leq t_2 \leq \ldots \leq t_{n-1} \leq s \leq t$ is any nondecreasing sequence of $n+1$ times and $i_1, i_2, \ldots, i_{n-1}, i, j \in S$ are any $n+1$ states in the state space, for any integer $n \geq 1$. That is, given the state of the process at any set of times prior to time t, the distribution of the process at time t depends only on the process at the most recent time prior to time t. An equivalent way to say this is to say that given the state of the process at time s, the distribution of the process at any time after s is independent of the entire past of the process before time s. This notion is exactly analogous to the Markov property for a discrete-time process.

Definition: A continuous-time stochastic process $\{X(t) : t \geq 0\}$ is called a *continuous-time Markov chain* if it has the Markov property.

The Markov property is a "forgetting" property, suggesting memorylessness in the distribution of the time a continuous-time Markov chain spends in any state. This is indeed the case if the process is also *time homogeneous*.

Time Homogeneity: We say that a continuous-time Markov chain is time homogeneous if for any $s \leq t$ and any states $i, j \in S$,

$$P(X(t) = j | X(s) = i) = P(X(t-s) = j | X(0) = i).$$

As with discrete-time Markov chains, a continuous-time Markov chain need not be time homogeneous, but in this course we will consider only time homogeneous Markov chains.

By time homogeneity, whenever the process enters state i, the way it evolves probabilistically from that point is the same as if the process started in state i at time 0. When the process enters state i, the time it spends there before it leaves state i is called the *holding time* in state i. By time homogeneity, we can speak of *the* holding time distribution because it is the same every time the process enters state i. Let T_i denote the holding time in state i. Then we have the following Proposition.

Proposition: T_i is exponentially distributed.

Proof. By time homogeneity, we assume that the process starts out in state i. For $s \geq 0$ the event $\{T_i > s\}$ is equivalent to the event $\{X(u) = i \text{ for } 0 \leq u \leq s\}$. Similarly, for $s, t \geq 0$ the event $\{T_i > s+t\}$ is equivalent to the event $\{X(u) = i \text{ for } 0 \leq u \leq s+t\}$. Therefore,

$$\begin{aligned}
&P(T_i > s+t | T_i > s) \\
&= P(X(u) = i \text{ for } 0 \leq u \leq s+t | X(u) = i \text{ for } 0 \leq u \leq s) \\
&= P(X(u) = i \text{ for } s < u \leq s+t | X(u) = i \text{ for } 0 \leq u \leq s) \\
&= P(X(u) = i \text{ for } s < u \leq s+t | X(s) = i) \\
&= P(X(u) = i \text{ for } 0 < u \leq t | X(0) = i) \\
&= P(T_i > t),
\end{aligned}$$

where

- the second equality follows from the simple fact that $P(A \bigcap B|A) = P(B|A)$, where we let $A = \{X(u) = i \text{ for } 0 \leq u \leq s\}$ and $B = \{X(u) = i \text{ for } s < u \leq s + t\}$.
- the third equality follows from the Markov property.
- the fourth equality follows from time homogeneity.

Therefore, the distribution of T_i has the memoryless property, which implies that it is exponential. □

By time homogeneity, every time our continuous-time Markov chain leaves state i,

- the number of states it could possibly jump to must stay the same, and we can let n_i denote this number.
- the set of states it could possibly jump to must stay the same, and we can let $\{j_1, \ldots, j_{n_i}\}$ denote this set of states.
- the probability of going to state j_ℓ must stay the same, and we can let p_{i,j_ℓ} denote this probability, for $\ell = 1, \ldots, n_i$.

Essentially, starting with the Markov property and time homogeneity, we have rebuilt our original description of a continuous-time Markov chain that was in terms of exponential alarm clocks. It may not be immediately obvious that we have done so because our current description uses the probabilities p_{i,j_ℓ} while our original description used the *rates* q_{i,j_ℓ}. But the two descriptions are the same, with the following correspondence between the p_{i,j_ℓ} and the q_{i,j_ℓ}:

$$p_{i,j_\ell} = q_{i,j_\ell}/v_i \quad \text{or} \quad q_{i,j_\ell} = v_i p_{i,j_\ell}.$$

Let us stop using the notation j_ℓ to denote a state that we can get to from state i, and just use the simpler notation j (or something similar like k), with the understanding that j is just a label. In this simpler notation, we have

$$p_{ij} = q_{ij}/v_i \quad \text{or} \quad q_{ij} = v_i p_{ij}.$$

We make the following remarks regarding p_{ij} and q_{ij}.

Remark Concerning p_{ij} (Embedded Markov Chains): The probability p_{ij} is the probability of going to state j at the next jump given that we are currently in state i. The matrix $\mathbf{P}$ whose (i,j)th entry is p_{ij} is a stochastic matrix and so is the one-step transition probability matrix of a (discrete-time) Markov chain. We call this discrete-time chain the *embedded Markov chain*. Every continuous-time Markov chain has an associated embedded discrete-time Markov chain. While the transition matrix $\mathbf{P}$ completely determines the probabilistic behaviour of the embedded discrete-time Markov chain, it does not fully capture the behaviour of the continuous-time process because it does not specify the *rates* at which transitions occur.

Remark Concerning q_{ij}: Recall that q_{ij} is the rate of the exponential alarm clock corresponding to state j that starts up whenever we enter state i. We say that q_{ij} is the rate of going from state i to state j. Note that $q_{ii} = 0$ for any i. The rates q_{ij} taken all together contain more information about the process than the probabilities p_{ij} taken all together. This is because if we know all the q_{ij} we can calculate all the v_i and then all the p_{ij}. But if we know all the p_{ij} we can't recover the q_{ij}. In many ways the q_{ij} are to continuous-time Markov chains what the p_{ij} are to discrete-time Markov chains.

However, there is an important difference between the q_{ij} in a continuous-time Markov chain and the p_{ij} in a discrete-time Markov chain to keep in mind. Namely, *the q_{ij} are rates, not probabilities and, as such, while they must be nonnegative, they are not bounded by 1*.

The Transition Probability Function

Just as the rates q_{ij} in a continuous-time Markov chain are the counterpart of the transition probabilities p_{ij} in a discrete-time Markov chain, there is a counterpart to the n-step transition probabilities $p_{ij}(n)$ of a discrete-time Markov chain. The *transition probability function*, $P_{ij}(t)$, for a time homogeneous, continuous-time Markov chain is defined as

$$P_{ij}(t) = P(X(t) = j | X(0) = i).$$

Note that there is no time "step" in a continuous-time Markov chain. For each pair of states $i, j \in S$, the transition probability function $P_{ij}(t)$ is in fact a continuous function of t. In the next lecture we will explore the relationship, which is fundamental, between the transition probability functions $P_{ij}(t)$ and the exponential rates q_{ij}. In general, one cannot determine the transition probability function $P_{ij}(t)$ in a nice closed form. In simple cases we can. For example, in the Poisson process we have seen that for $i \leq j$,

$$\begin{aligned} P_{ij}(t) &= P(\text{there are } j-i \text{ events in an interval of length } t) \\ &= \frac{(\lambda t)^{j-i}}{(j-i)!} e^{-\lambda t}. \end{aligned}$$

In Proposition 6.1, the text shows how one can explicitly compute $P_{ij}(t)$ for a pure birth process, which was described last time, in which the birth rates λ_i are all different (that is, $\lambda_i \neq \lambda_j$ for $i \neq j$). Please read this example in the text.

We can say some important general things about $P_{ij}(t)$, however. Since these functions are the counterpart of the n-step transition probabilities, one might guess that there is a counterpart to the Chapman-Kolmogorov equations for these functions. There is, and we will end today's lecture with this result, whose proof is essentially identical to the proof in the discrete case.

Lemma. *(Lemma 6.3 in text, Chapman-Kolmogorov Equations).* Let $\{X(t) : t \geq 0\}$ be a continuous-time Markov chain with state space S, rates $(q_{ij})_{i,j \in S}$ and transition probability functions $(P_{ij}(t))_{i,j \in S}$. Then for any $s, t \geq 0$,

$$P_{ij}(t+s) = \sum_{k \in S} P_{ik}(t) P_{kj}(s).$$

Proof. By conditioning on $X(t)$, we have

$$\begin{aligned}
P_{ij}(t+s) &= P(X(t+s) = j | X(0) = i) \\
&= \sum_{k \in S} P(X(t+s) = j | X(t) = k, X(0) = i) \\
&\qquad \times P(X(t) = k | X(0) = i) \\
&= \sum_{k \in S} P(X(t+s) = j | X(t) = k) P(X(t) = k | X(0) = i) \\
&= \sum_{k \in S} P(X(s) = j | X(0) = k) P(X(t) = k | X(0) = i) \\
&= \sum_{k \in S} P_{kj}(s) P_{ik}(t),
\end{aligned}$$

as desired. □

For a given t, if we form the probabilities $P_{ij}(t)$ into an $|S| \times |S|$ matrix $\mathbf{P}(t)$ whose (i,j)th entry is $P_{ij}(t)$, then the Chapman-Kolmogorov equation

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

says that the (i,j)th entry of $\mathbf{P}(t+s)$ is the dot product of the ith row of $\mathbf{P}(t)$ and the jth column of $\mathbf{P}(s)$. But that is the same thing as the (i,j)th entry in the matrix product of $\mathbf{P}(t)$ and $\mathbf{P}(s)$. That is,

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s).$$

This is the direct analogue of the discrete-time result. Just a note on terminology: in the discrete-time case we called the matrix $\mathbf{P}(n)$ the n-step transition probability matrix. Because there is no notion of a time step in continuous time, we just simply call $\mathbf{P}(t)$ the matrix transition probability function. Note that it is a *matrix-valued* function of the continuous variable t.

27

Key Properties of Continuous-Time Markov Chains

The key quantities that specify a discrete-time Markov chains are the transition *probabilities* p_{ij}. In continuous time, the corresponding key quantities are the transition *rates* q_{ij}. Recall that we may think of q_{ij} as the rate of an exponentially distributed alarm clock that starts as soon as we enter state i, where there is one alarm clock that starts for each state that we could possibly go to when we leave state i. We leave state i as soon as an alarm clock goes off and we go to state j if it was the clock corresponding to state j that went off first. The time until the first alarm clock goes off is exponentially distributed with rate $v_i = \sum_{j \in S} q_{ij}$, where we let $q_{ij} = 0$ if we cannot go to state j from state i. When we leave state i we go to state j with probability q_{ij}/v_i, which we also denote by p_{ij}. The p_{ij} are the transition probabilities of the *embedded* discrete-time Markov chain, also called the *jump* chain.

To summarize, the quantities q_{ij}, v_i and p_{ij} are related by the equalities

$$\begin{aligned} v_i &= \sum_{j \in S} q_{ij} \\ q_{ij} &= v_i p_{ij} \\ p_{ij} &= q_{ij}/v_i. \end{aligned}$$

To avoid technicalities, we will assume that $v_i < \infty$ for all i for this course. It is possible for v_i to equal $+\infty$ since the rates q_{ij} need not form a convergent sum when we sum over j. If $v_i = \infty$ then the process will leave state i immediately after it enters state i. This behaviour is not typical of the models we will consider in this course (though it can be typical for some kinds of systems, such as configurations on an infinite lattice for example). We also will assume that $v_i > 0$ for all i. If $v_i = 0$ then when we enter state i we will stay there forever so $v_i = 0$ would correspond to state i being an absorbing state. This does not present any real technical difficulties (we have already considered this possibility in the discrete time setting). However, we will not consider any absorbing states in the continuous time models we will look at.

Since the time spent in state i is exponentially distributed with rate v_i (where $0 < v_i < \infty$), we may expect from what we know about the Poisson process that the probability of 2 or more transitions in a time interval of length h should be $o(h)$. This is indeed the case. If T_i denotes the holding time in state i, then T_i is Exponential(v_i), and

$$P(T_i > h) = e^{-v_i h}.$$

Expanding the exponential function in a Taylor series, we have

$$\begin{aligned} P(T_i > h) &= e^{-v_i h} \\ &= 1 - v_i h + \frac{(v_i h)^2}{2!} - \frac{(v_i h)^3}{3!} + \dots \\ &= 1 - v_i h + o(h). \end{aligned}$$

This also implies that

$$P(T_i \leq h) = v_i h + o(h).$$

Furthermore, if T_j denotes the holding time in state j, for $j \neq i$, then T_j is Exponentially distributed with rate v_j and T_j is independent of T_i. Since the event $\{T_i + T_j \leq h\}$ implies the event $\{T_i \leq h, T_j \leq h\}$ we have that

$$\{T_i + T_j \leq h\} \subset \{T_i \leq h, T_j \leq h\},$$

so that

$$\begin{aligned} P(T_i + T_j \leq h) &\leq P(T_i \leq h, T_j \leq h) \\ &= P(T_i \leq h)P(T_j \leq h) \\ &= (v_i h + o(h))(v_j h + o(h)) \\ &= v_i v_j h^2 + o(h) = o(h), \end{aligned}$$

which implies that $P(T_i + T_j \leq h) = o(h)$. Thus, starting in state i, if we compute the probability of 2 or more transitions by time h by conditioning on the first transition, we obtain

$$\begin{aligned} &P(\text{2 or more transitions by time } h | X(0) = i) \\ &= \sum_{j \neq i} P(\text{2 or more transitions by } h | X(0) = i, \text{1st transition to } j) p_{ij} \\ &= \sum_{j \neq i} P(T_i + T_j \leq h) p_{ij} = \sum_{j \neq i} o(h) p_{ij} = o(h). \end{aligned}$$

Since

$$\begin{aligned} P(\text{0 transitions by time } h | X(0) = i) &= P(T_i > h) \\ &= 1 - v_i h + o(h), \end{aligned}$$

we also have

$$\begin{aligned} &P(\text{exactly 1 transition by time } h | X(0) = i) \\ &= 1 - (1 - v_i h + o(h)) - o(h) \\ &= v_i h + o(h). \end{aligned}$$

To summarize, we have

$$\begin{aligned} P(\text{0 transitions by time } h | X(0) = i) &= 1 - v_i h + o(h) \\ P(\text{exactly 1 transition by time } h | X(0) = i) &= v_i h + o(h) \\ P(\text{2 or more transitions by time } h | X(0) = i) &= o(h). \end{aligned}$$

Now, for $j \neq i$, consider the conditional probability

$$P(X(h) = j | X(0) = i).$$

Given that $X(0) = i$, one way for the event $\{X(h) = j\}$ to occur is for there to be exactly one transition in the interval $[0, h]$ and for that transition to be to state j. The probability of this is $(v_i h + o(h))p_{ij} = v_i p_{ij} h + o(h)$. Moreover, the event consisting of the union of *every other* way to be in state j at time h starting in state i implies the event that there were 2 or more transitions in the interval $[0, h]$. So the probability of this second event is $o(h)$. Summarizing, we have

$$\begin{aligned} P(X(h) = j | X(0) = i) &= v_i p_{ij} h + o(h) + o(h) \\ &= v_i p_{ij} h + o(h). \end{aligned}$$

Similarly, if we consider the conditional probability

$$P(X(h) = i | X(0) = i),$$

the only way for the event $\{X(h) = i\}$ to occur given that $X(0) = i$ that does not involve at least 2 transitions in the interval $[0, h]$ is for there to be 0 transitions in the interval $[0, h]$. Thus,

$$\begin{aligned} P(X(h) = i | X(0) = i) &= P(\text{0 transitions in } [0, h] | X(0) = i) + o(h) \\ &= 1 - v_i h + o(h) + o(h) \\ &= 1 - v_i h + o(h). \end{aligned}$$

Now we are in a position to derive a set of differential equations, called Kolmogorov's Equations, for the probability functions $p_{ij}(t)$. We proceed in a familiar way, by deriving a system of equations by conditioning. There are actually 2 sets of equations we can derive for the $p_{ij}(t)$ — Kolmogorov's *Backward* Equations and Kolmogorov's *Forward* Equations. We will now derive the Backward Equations. To do so we will evaluate $p_{ij}(t+h)$ by conditioning on $X(h)$ (here h is some small positive amount). We obtain

$$\begin{aligned}
&p_{ij}(t+h) \\
&\quad = P(X(t+h)=j|X(0)=i) \\
&\quad = \sum_{k\in S} P(X(t+h)=j|X(h)=k,X(0)=i)P(X(h)=k|X(0)=i) \\
&\quad = \sum_{k\in S} P(X(t+h)=j|X(h)=k)P(X(h)=k|X(0)=i) \\
&\quad = \sum_{k\in S} P(X(t)=j|X(0)=k)P(X(h)=k|X(0)=i) \\
&\quad = \sum_{k\in S} p_{kj}(t)P(X(h)=k|X(0)=i),
\end{aligned}$$

where the third equality follows from the Markov property and the fourth equality follows from time-homogeneity. Now we separate out the term with $k=i$ and use our results from the previous page to obtain

$$p_{ij}(t+h) = p_{ij}(t)(1-v_ih+o(h)) + \sum_{k\neq i} p_{kj}(t)(v_ip_{ik}h+o(h)),$$

which is equivalent to

$$p_{ij}(t+h) - p_{ij}(t) = -v_ip_{ij}(t)h + \sum_{k\neq i} p_{kj}(t)v_ip_{ik}h + o(h).$$

Upon dividing by h, and using the fact that $v_i p_{ik} = q_{ik}$, we get

$$\frac{p_{ij}(t+h) - p_{ij}(t)}{h} = \sum_{k \neq i} q_{ik} p_{kj}(t) - v_i p_{ij}(t) + \frac{o(h)}{h},$$

As we let $h \to 0$, the left hand side above approaches $p'_{ij}(t)$, which shows that $p_{ij}(t)$ is differentiable, and given by

$$p'_{ij}(t) = \sum_{k \neq i} q_{ik} p_{kj}(t) - v_i p_{ij}(t).$$

The above differential equations, for $i, j \in S$, are called Kolmogorov's Backward Equations. We may write down the entire set of equations more succinctly in matrix form. Let $\mathbf{P}(t)$ be the $|S| \times |S|$ matrix with (i,j)th entry $p_{ij}(t)$ and $\mathbf{P}'(t)$ the $|S| \times |S|$ matrix with (i,j)th entry $p'_{ij}(t)$. We call $\mathbf{P}(t)$ the matrix transition probability function, which is a (matrix-valued) differentiable function of t. If we form a matrix, which we will call $\mathbf{G}$, whose ith row has $-v_i$ in the ith column and q_{ik} in the kth column, then we see that the right hand side of Kolmogorov's Backward Equation for $p_{ij}(t)$ is just the dot product of the ith row of $\mathbf{G}$ with the jth column of $\mathbf{P}(t)$. That is, the differential equation above is the same as

$$[\mathbf{P}'(t)]_{ij} = [\mathbf{G}\mathbf{P}(t)]_{ij},$$

so that in matrix form, Kolmogorov's Backward Equations can be written as

$$\mathbf{P}'(t) = \mathbf{G}\mathbf{P}(t).$$

The Infinitesimal Generator: The matrix $\mathbf{G}$ is a fundamental quantity associated with the continuous-time Markov chain $\{X(t) : t \geq 0\}$. It is called the *infinitesimal generator*, or simply *generator*, of the chain. If we let g_{ij} denote the (i, j)th entry of $\mathbf{G}$, then

$$\begin{aligned} g_{ij} &= q_{ij} \quad \text{for } i \neq j, \text{ and} \\ g_{ii} &= -v_i. \end{aligned}$$

The generator matrix $\mathbf{G}$ contains all the rate information for the chain and, even though it's entries are not probabilities, it is the counterpart of the one-step transition probability matrix $\mathbf{P}$ for discrete-time Markov chains. In deriving Kolmogorov's Backward Equations, if we had conditioned on $X(t)$ instead of $X(h)$ we would have derived another set of differential equations called Kolmogorov's *Forward* Equations, which in matrix form are given by

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{G}.$$

For both the backward and the forward equations, we have the boundary condition

$$\mathbf{P}(0) = \mathbf{I},$$

where $\mathbf{I}$ is the $|S| \times |S|$ identity matrix. The boundary condition follows since

$$p_{ii}(0) = P(X(0) = i | X(0) = i) = 1$$

and, for $i \neq j$,

$$p_{ij}(0) = P(X(0) = j | X(0) = i) = 0.$$

Though the backward and forward equations are two different sets of differential equations, with the above boundary condition they have the same solution, given by

$$\begin{aligned}\mathbf{P}(t) &= e^{t\mathbf{G}} \equiv \sum_{n=0}^{\infty} \frac{(t\mathbf{G})^n}{n!} \\ &= \mathbf{I} + t\mathbf{G} + \frac{(t\mathbf{G})^2}{2!} + \frac{(t\mathbf{G})^3}{3!} + \ldots\end{aligned}$$

Keep in mind that the notation $e^{t\mathbf{G}}$ is meaningless except as shorthand notation for the infinite sum above. To see that the above satisfies the backward equations we may simply plug it into the differential equations and check that it solves them. Differentiating with respect to t, we get

$$\begin{aligned}\mathbf{P}'(t) &= \mathbf{G} + t\mathbf{G}^2 + \frac{t^2}{2!}\mathbf{G}^3 + \frac{t^3}{3!}\mathbf{G}^4 + \ldots \\ &= \mathbf{G}\left[\mathbf{I} + t\mathbf{G} + \frac{(t\mathbf{G})^2}{2!} + \frac{(t\mathbf{G})^3}{3!} + \ldots\right] \\ &= \mathbf{G}\mathbf{P}(t).\end{aligned}$$

Also, $\mathbf{P}(0) = \mathbf{I}$ is clearly satisfied. Moreover, we could also have written

$$\mathbf{P}'(t) = \left[\mathbf{I} + t\mathbf{G} + \frac{(t\mathbf{G})^2}{2!} + \frac{(t\mathbf{G})^3}{3!} + \ldots\right]\mathbf{G} = \mathbf{P}(t)\mathbf{G},$$

showing that $\mathbf{P}(t) = e^{t\mathbf{G}}$ satisfies the forward equations as well.

Thus, even though we cannot normally obtain $\mathbf{P}(t)$ in a simple and explicit closed form, the infinite sum representation $e^{t\mathbf{G}}$ is general, and can be used to obtain numerical approximations to $\mathbf{P}(t)$ if $|S|$ is finite, by truncating the infinite sum to a finite sum (see Section 6.8).

Remark: The text uses the notation $\mathbf{R}$ for the generator matrix, presumably to stand for the Rate matrix. The notation $\mathbf{G}$ is more common and will be adopted here, and the terminology *generator matrix* or *infinitesimal generator matrix* is standard.

The solution $\mathbf{P}(t) = e^{t\mathbf{G}}$ shows how basic the generator matrix $\mathbf{G}$ is to the properties of a continuous-time Markov chain. We will now show that the generator $\mathbf{G}$ is also the key quantity for determining the stationary distribution of the chain. First, we define what we mean by a stationary distribution for a continuous-time Markov chain.

Stationary Distributions:

Definition: Let $\{X(t) : t \geq 0\}$ be a continuous-time Markov chain with state space S, generator $\mathbf{G}$, and matrix transition probability function $\mathbf{P}(t)$. An $|S|$-dimensional (row) vector $\boldsymbol{\pi} = (\pi_i)_{i \in S}$ with $\pi_i \geq 0$ for all i and $\sum_{i \in S} \pi_i = 1$, is said to be a *stationary distribution* if $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t)$, for all $t \geq 0$.

A vector $\boldsymbol{\pi}$ which satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t)$ for all $t \geq 0$ is called a stationary distribution for exactly the same reason that the stationary distribution of a discrete-time Markov chain is called the stationary distribution. It makes the process stationary. That is, if we set the initial distribution of $X(0)$ to be such a $\boldsymbol{\pi}$, then the distribution of $X(t)$ will also be $\boldsymbol{\pi}$ for all $t > 0$ (i.e. $P(X(t) = j) = \pi_j$ for all $j \in S$ and all $t > 0$). To see this, set the initial distribution of $X(0)$ to be $\boldsymbol{\pi}$ and compute $P(X(t) = j)$ by conditioning on $X(0)$. This gives

$$\begin{aligned} P(X(t)=j) &= \sum_{i\in S} P(X(t)=j|X(0)=i)P(X(0)=i) \\ &= \sum_{i\in S} p_{ij}(t)\pi_i = [\boldsymbol{\pi}\mathbf{P}(t)]_j = \pi_j, \end{aligned}$$

as claimed.

To see how the generator $\mathbf{G}$ relates to the definition of a stationary distribution, we can replace $\mathbf{P}(t)$ in the definition of $\boldsymbol{\pi}$ with $e^{t\mathbf{G}}$. Doing so, we obtain the following equivalences:

$$\begin{aligned} \boldsymbol{\pi} \text{ is a stationary distribution } &\Leftrightarrow \boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t) \quad \text{for all } t \geq 0 \\ &\Leftrightarrow \boldsymbol{\pi} = \boldsymbol{\pi}\sum_{n=0}^{\infty} \frac{(t\mathbf{G})^n}{n!} \quad \text{for all } t \geq 0 \\ &\Leftrightarrow \mathbf{0} = \sum_{n=1}^{\infty} \frac{t^n}{n!}\boldsymbol{\pi}\mathbf{G}^n \quad \text{for all } t \geq 0 \\ &\Leftrightarrow \mathbf{0} = \boldsymbol{\pi}\mathbf{G}^n \quad \text{for all } n \geq 1 \\ &\Leftrightarrow \mathbf{0} = \boldsymbol{\pi}\mathbf{G}. \end{aligned}$$

You should convince yourself that the implications are true in both directions in each of the lines above.

Thus, we see that the condition $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t)$ for all $t \geq 0$, which would be quite difficult to check, reduces to the much simpler condition $\mathbf{0} = \boldsymbol{\pi}\mathbf{G}$, in terms of the generator matrix $\mathbf{G}$. The equations $\mathbf{0} = \boldsymbol{\pi}\mathbf{G}$ are a set of $|S|$ linear equations which, together with the normalization constraint $\sum_{i\in S} \pi_i = 1$, determines the stationary distribution $\boldsymbol{\pi}$ if one exists.

The jth equation in $\mathbf{0} = \boldsymbol{\pi}\mathbf{G}$ is given by

$$0 = -v_j\pi_j + \sum_{i \neq j} q_{ij}\pi_i,$$

which is equivalent to

$$\pi_j v_j = \sum_{i \neq j} \pi_i q_{ij}.$$

This equation has the following interpretation. On the left hand side, π_j is the long run proportion of time that the process is in state j, while v_j is that rate of leaving state j when the process is in state j. Thus, the product $\pi_j v_j$ is interpreted as the long run rate of leaving state j. On the right hand side, q_{ij} is the rate of going to state j when the process is in state i, so the product $\pi_i q_{ij}$ is interpreted as the long run rate of going from state i to state j. Summing over all $i \neq j$ then gives the long run rate of going to state j. That is, the equation

$$\pi_j v_j = \sum_{i \neq j} \pi_i q_{ij}$$

is interpreted as

"the long run rate out of state j" $=$ "the long run rate into state j",

and for this reason the equations $\mathbf{0} = \boldsymbol{\pi}\mathbf{G}$ are called the *Global Balance Equations*, or just *Balance Equations*, because they express the fact that when the process is made stationary, there must be equality, or balance, in the long run rates into and out of any state.

28

Limiting Probabilities

We now consider the limiting probabilities

$$\lim_{t\to\infty} p_{ij}(t),$$

for a continuous-time Markov chain $\{X(t) : t \geq 0\}$, where

$$p_{ij}(t) = P(X(t) = j | X(0) = i)$$

is the transition probability function for the states i and j.

Last time we considered the stationary distribution $\boldsymbol{\pi}$ of a continuous-time Markov chain, and saw that $\boldsymbol{\pi}$ is the distribution of $X(t)$ for all t when the process is stationary. We also interpret π_j as the long run proportion of time that the process is in state j. Based on what we know about discrete-time Markov chains, we may expect that the limiting probability $\lim_{t\to\infty} p_{ij}(t)$ is equal to the stationary probability π_j for all $i \in S$. That is, no matter what state i we start in at time 0, the probability that we are in state j at time t approaches π_j as t gets larger and larger. This is indeed the case, assuming the stationary distribution $\boldsymbol{\pi}$ exists, although it may not exist. However, in this course the only continuous-time Markov chains we will consider will be those for which the stationary distribution exists.

Actually, the fact that the limiting probabilities and the stationary probabilities are the same is even more true in continuous-time than in discrete time. In discrete time, we saw that even though the stationary distribution may exist, the limiting probabilities still may not exist if the discrete-time chain is not aperiodic. However, in continuous time we don't run into such difficulties because continuous-time Markov chains don't have a period! There is no "step" in continuous time, so there is no definition of "period" for a continuous-time Markov chain. In fact, it can be shown (though we won't prove it) that for any two states i and j in a continuous-time Markov chain, exactly one of the following two statements must be true:

1. $p_{ij}(t) = 0$ for all $t > 0$, or
2. $p_{ij}(t) > 0$ for all $t > 0$.

This is called the *Levy Dichotomy*, and it shows that if a continuous-time Markov chain is irreducible, in the sense that the embedded jump chain is irreducible, then starting in state i we could possibly be in state j at any positive time, for any state j, including the starting state i.

We may state the following theorem which summarizes the basic result we would like to have concerning the limiting probabilities.

Theorem: In a continuous-time Markov chain, if a stationary distribution $\boldsymbol{\pi}$ exists, then it is unique and

$$\lim_{t\to\infty} p_{ij}(t) = \pi_j,$$

for all i.

We will not prove the preceding theorem completely. We will say something about the uniqueness of $\boldsymbol{\pi}$ (if it exists) in tomorrow's lecture. For now let us focus on the second statement in the theorem concerning the limiting probabilities. Using the Kolmogorov Forward Equations, we can easily prove something slightly weaker, which is that *assuming the limiting probabilities exist and are independent of the starting state*, then $\lim_{t\to\infty} p_{ij}(t) = \pi_j$. This is the extent of what is shown in the text in Section 6.5, and we'll content ourselves with that. However, you should be aware that we are not completely proving the statement in the theorem (not because it is too difficut to prove, but just in the interest of time).

Thus, assuming $\lim_{t\to\infty} p_{ij}(t)$ exists and is independent of i, let $\nu_j = \lim_{t\to\infty} p_{ij}(t)$ and let $\boldsymbol{\nu} = (\nu_j)_{j\in S}$ be the $|S|$-dimensional row vector whose jth component is ν_j. In matrix form the assumption that $\lim_{t\to\infty} p_{ij}(t) = \nu_j$ for all i and j is

$$\lim_{t\to\infty} \mathbf{P}(t) = \mathbf{V},$$

where $\mathbf{P}(t)$ is the matrix transition probability function with (i,j)th entry $p_{ij}(t)$ introduced last time, and $\mathbf{V}$ is an $|S| \times |S|$ matrix in which each row is equal to $\boldsymbol{\nu}$.

Now, if $p_{ij}(t) \to \nu_j$ as $t \to \infty$, then we must have $p'_{ij}(t) \to 0$ as $t \to \infty$ because $p_{ij}(t)$ is becoming more and more a constant as t gets larger and larger. In matrix form we may write this as

$$\lim_{t\to\infty} \mathbf{P}'(t) = \mathbf{0},$$

where $\mathbf{P}'(t)$ is the $|S| \times |S|$ matrix with (i,j)th entry $p'_{ij}(t)$ and $\mathbf{0}$ is just the $|S| \times |S|$ matrix of zeros.

Now, recall that Kolmogorov's Forward Equations state that

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{G},$$

where $\mathbf{G}$ is the infinitesimal generator of the chain. Thus, letting $t \to \infty$, we obtain

$$\mathbf{0} = \mathbf{V}\mathbf{G}.$$

But since each row of $\mathbf{V}$ is equal to the limiting probability vector $\boldsymbol{\nu}$, this implies that

$$\mathbf{0} = \boldsymbol{\nu}\mathbf{G},$$

where now (slightly abusing notation), $\mathbf{0}$ denotes the $|S|$-dimensional row vector of zeros.

Thus, we see that $\boldsymbol{\nu}$ satisfies the global balance equations, which are the equations which determine the stationary distribution $\boldsymbol{\pi}$. Assuming the stationary distribution is unique this implies that $\boldsymbol{\nu} = \boldsymbol{\pi}$. So the global balance equations yield both the stationary distribution $\boldsymbol{\pi}$ and the limiting probability vector $\boldsymbol{\pi}$. As in the discrete-time setting, this is an important and useful result because if it were not true then we would need to do different calculations depending on what questions we were asking about our system being modeled, and it is not clear that finding limiting probabilities would be very easy or even possible.

29

Local Balance Equations

We have seen that for a continuous-time Markov chain $\boldsymbol{X} = \{X(t) : t \geq 0\}$, the stationary distribution $\boldsymbol{\pi}$, if it exists, must satisfy the global balance equations $\mathbf{0} = \boldsymbol{\pi}\mathbf{G}$, where $\mathbf{G}$ is the infinitesimal generator of the chain. As for discrete-time Markov chains, there is also a set of equations called the *local balance equations* that the stationary distribution of $\boldsymbol{X}$ may or may not satisfy. Today we will discuss the local balance equations for a continuous-time Markov chain, and give some examples.

First, however, we will sidetrack from this discussion to note the important relationship between the stationary distribution of $\boldsymbol{X}$ and the stationary distribution of the embedded discrete-time jump chain of $\boldsymbol{X}$. These two distributions are in fact *not* the same. We will also discuss some consequences of this relationship.

Relationship Between the Stationary Distribution $\boldsymbol{\pi}$ of a Continuous-Time Markov Chain and the Stationary Distribution $\boldsymbol{\psi}$ of its Corresponding Embedded Discrete-Time Jump Chain:

Let $\boldsymbol{X} = \{X(t) : t \geq 0\}$ be a continuous-time Markov chain with state space S and transition rates q_{ij} and, as usual, we let

$$v_i = \sum_{j \neq i} q_{ij}$$

be the rate out of state i, and

$$p_{ij} = \frac{q_{ij}}{v_i}$$

be the one-step transition probabilities of the embedded discrete-time jump chain. Note also that $q_{ii} = 0$ for all $i \in S$ so that we may also write

$$v_i = \sum_{j \in S} q_{ij}.$$

We let $\mathbf{G}$ denote the infinitesimal generator of the continuous-time Markov chain (with entries $g_{ij} = q_{ij}$ for $i \neq j$ and $g_{ii} = -v_i$) and let $\mathbf{P}$ denote the one-step transition matrix for the embedded jump chain (with entries p_{ij}). If $\boldsymbol{\pi}$ is the stationary distribution of the continuous-time chain and $\boldsymbol{\psi}$ is the stationary distribution of the embedded jump chain, then $\boldsymbol{\pi}$ and $\boldsymbol{\psi}$ must satisfy, respectively, the two sets of global balance equations $\mathbf{0} = \boldsymbol{\pi}\mathbf{G}$ and $\boldsymbol{\psi} = \boldsymbol{\psi}\mathbf{P}$. Writing out the jth equation in each of these two sets of equations, we have

$$\pi_j v_j = \sum_{i \neq j} \pi_i q_{ij},$$

and

$$\psi_j = \sum_{i \in S} \psi_i p_{ij}.$$

These two sets of global balance equations give us a relationship between $\boldsymbol{\pi}$ and $\boldsymbol{\psi}$, as follows. Since $q_{ij} = v_i p_{ij}$ and $q_{jj} = 0$, we may rewrite the jth equation in $\mathbf{0} = \boldsymbol{\pi}\mathbf{G}$ as

$$\pi_j v_j = \sum_{i \in S} \pi_i v_i p_{ij}.$$

Assuming $\boldsymbol{\pi}$ satisfies $\mathbf{0} = \boldsymbol{\pi}\mathbf{G}$, we see that the vector $(\pi_j v_j)_{j \in S}$, with jth entry $\pi_j v_j$, satisfies the global balance equations for the stationary distribution of the embedded jump chain. Furthermore, since we know that the stationary distribution of the embedded jump chain is unique from our theory for discrete-time Markov chains, we may conclude that

$$\psi_j = C \pi_j v_j,$$

where C is an appropriate normalizing constant. This also gives that

$$\pi_j = \frac{1}{C} \times \frac{\psi_j}{v_j}.$$

Indeed, we have that

$$\psi_j = \frac{\pi_j v_j}{\sum_{i \in S} \pi_i v_i}$$

and

$$\pi_j = \frac{\psi_j / v_j}{\sum_{i \in S} \psi_i / v_i}.$$

The above relationship between ψ_j and π_j is intuitively correct. We may interpret ψ_j as the long run proportion of transitions that the continuous-time chain makes into state j. Also, over all the times that we make a transition into state j, we stay in state j for an average of $1/v_j$ time units. Therefore, the product ψ_j / v_j should be proportional to the long run proportion of time that the continuous-time chain spends in state j, and this is how we interpret π_j.

We make several remarks associated with the relationship between $\boldsymbol{\pi}$ and $\boldsymbol{\psi}$.

Remark 1: If $\boldsymbol{\psi}$ and $\boldsymbol{\pi}$ both exist and the embedded jump chain is irreducible, then the uniqueness of $\boldsymbol{\psi}$, which we proved in our theory for discrete-time Markov chains, and the fact that $\boldsymbol{\pi}$ is determined from $\boldsymbol{\psi}$ through the above relationship, implies that $\boldsymbol{\pi}$ is also unique.

Remark 2: We have assumed that $\boldsymbol{\pi}$ and $\boldsymbol{\psi}$ both exist. However, it is possible for $\boldsymbol{\psi}$ to exist but for $\boldsymbol{\pi}$ not to exist. From any discrete-time Markov chain with transition probabilities p_{ij}, we may construct a continuous-time Markov chain by specifying the rates v_i. However, if in addition the discrete-time chain has a unique stationary distribution $\boldsymbol{\psi}$, it is not always true that the corresponding continuous-time chain will have a stationary distribution $\boldsymbol{\pi}$. This is because there is nothing forcing the normalizing constant $\sum_{i \in S} \psi_i v_i$ to be finite if the state space S is infinite and we are free to choose our rates v_i as we please. In particular, one can show that this sum is not finite if the state space S is countably infinite and we choose $v_i = \psi_i$ for all $i \in S$.

Remark 3: The fact that $\boldsymbol{\pi}$ can be obtained from $\boldsymbol{\psi}$, assuming both exist, has practical value, especially when the state space is large but finite, and the transition matrix $\mathbf{P}$ of the embedded jump chain is sparse in the sense of having many zero entries (this occurs if, even though the state space may be large, the number of possible states that can be reached from state i, for any i, in one step remains small). Such models turn out to be quite common for many practical systems. The global balance equation $\boldsymbol{\psi} = \boldsymbol{\psi}\mathbf{P}$ is what is called a *fixed point equation*, which means that the stationary vector $\boldsymbol{\psi}$ is a fixed point of the mapping which takes a row vector $\boldsymbol{x}$ to the row vec-

tor $\boldsymbol{x}\mathbf{P}$. A common, and simple, numerical procedure for solving a fixed point equation is something called *successive substitution*. In this procedure, we simply start with a convenient initial probability vector $\boldsymbol{\psi}^{(0)}$ (such as $\psi_j^{(0)} = 1/|S|$ for all $i \in S$) and then obtain $\boldsymbol{\psi}^{(1)} = \boldsymbol{\psi}^{(0)}\mathbf{P}$. We continue iterating, obtaining $\boldsymbol{\psi}^{(n+1)} = \boldsymbol{\psi}^{(n)}\mathbf{P}$ from $\boldsymbol{\psi}^{(n)}$, for $n \geq 1$. Then, under certain conditions, the sequence of vectors $\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots$ will converge to a fixed point $\boldsymbol{\psi}$. Note that

$$\boldsymbol{\psi}^{(n)} = \boldsymbol{\psi}^{(n-1)}\mathbf{P} = \boldsymbol{\psi}^{(n-2)}\mathbf{P}^2 = \cdots = \boldsymbol{\psi}^{(0)}\mathbf{P}^n,$$

for all $n \geq 1$. If the embedded jump chain is irreducible (which it must be for a unique stationary distribution $\boldsymbol{\psi}$ to exist) and aperiodic, then from our theory on the limiting probabilities of a discrete-time Markov chain, we know that $\mathbf{P}^n$ converges, as $n \to \infty$, to an $|S| \times |S|$ matrix in which each row is equal to the stationary distribution $\boldsymbol{\psi}$ (since we know $p_{ij}(n) \to \psi_j$ as $n \to \infty$, for all $i, j \in S$). This implies that $\boldsymbol{\psi}^{(0)}\mathbf{P}^n$, and so $\boldsymbol{\psi}^{(n)}$, converges to $\boldsymbol{\psi}$. The numerical efficiency one can obtain by computing $\boldsymbol{\psi}$, and then $\boldsymbol{\pi}$ through the relationship between $\boldsymbol{\psi}$ and $\boldsymbol{\pi}$, in this way can be an order of magnitude. A direct, numerical solution of the system of linear equations $\boldsymbol{\psi} = \boldsymbol{\psi}\mathbf{P}$ has numerical complexity $O(|S|^2)$. On the other hand, each iteration in the successive substitution procedure requires a vector – matrix multiplication, which is in general also an $O(|S|^2)$ operation. However, assuming each column of $\mathbf{P}$ has only a very small number, relative to $|S|$, postive entries, one may cleverly compute $\boldsymbol{\psi}^{(n)}\mathbf{P}$ with only $K|S|$ multiplications and additions, where K is much smaller than $|S|$. In other words, the complexity of this operation can be reduced to $O(|S|)$. Moreover, in practice it takes only a few iterations for the sequence $\{\boldsymbol{\psi}^{(n)}\}$ to converge to $\boldsymbol{\psi}$ to within a reasonable tolerance (say 10^{-8}).

Local Balance Equations: As for discrete-time Markov chains, the stationary distribution of a continuous-time Markov chain *must* satisfy the global balance equations, but may also satisfy the *local balance equations*. For a continuous-time Markov chain the local balance equations are given by

$$\pi_i q_{ij} = \pi_j q_{ji},$$

for all $i, j \in S$ such that $i \neq j$. The local balance equations express a balance of flow between any pair states. We interpret $\pi_i q_{ij}$ as the rate from state i to state j and $\pi_j q_{ji}$ as the rate from state j to state i. There are actually $\binom{|S|}{2}$ equations in the set of local balance equations (the same as in the local balance equations for a discrete-time Markov chain), but typically most of the equations are trivially satisfied because $q_{ij} = q_{ji} = 0$. Note that one way to quickly check if the local balance equations *cannot* be satisified by the stationary distribution $\boldsymbol{\pi}$ is to check if there are any rates q_{ij} and q_{ji} such that $q_{ij} > 0$ and $q_{ji} = 0$ or $q_{ij} = 0$ and $q_{ji} > 0$.

Not every continuous-time Markov chain that has a stationary distribution has a stationary distribution that satisfies the local balance equations. On the other hand, if we can find a probability vector that does satisfy the local balance equations, then this probability vector will be the stationary distribution of the Markov chain. We have seen this with discrete-time Markov chains, and we can easily show it again here.

Suppose that $\boldsymbol{\pi}$ is a probability vector that satisfies the local balance equations. That is,

$$\pi_i q_{ij} = \pi_j q_{ji},$$

for all $i, j \in S$ such that $i \neq j$. Then, since $q_{jj} = 0$ for any $j \in S$, we may sum both sides of the above equality over all $i \in S$ to obtain

$$\sum_{i \in S} \pi_i q_{ij} = \pi_j \sum_{i \in S} q_{ji} = \pi_j v_j,$$

for all $j \in S$. But these are just the global balance equations. That is, the probability vector $\boldsymbol{\pi}$ also satisfies the global balance equations, and this implies that $\boldsymbol{\pi}$ is the stationary distribution.

If there is a probability vector $\boldsymbol{\pi}$ that satisfies the local balance equations, then using the local balance equations to find $\boldsymbol{\pi}$ is typically much easier than using the global balance equations because each equation in the local balance equations involves only two unknowns, while at least some of the equations in the global balance equations will usually involve more than two unknowns.

We will now give two examples of continuous-time Markov chains whose stationary distributions do satisfy the local balance equations, in part to illustrate the utility of using the local balance equations to find the stationary distributions.

Example: *Birth/Death Processes:* We introduced birth/death processes previously. The state space of a birth/death process is a subset (possibly infinite) of the integers, and from any state i the process can only jump up to state $i + 1$ or down to state $i - 1$. The transition rates $q_{i,i+1}$, usually denoted by λ_i, are called the *birth rates* of the process and the transition rates $q_{i,i-1}$, usually denoted by μ_i, are

called the *death rates* of the process. In this example we will consider a birth/death process on $S = \{0, 1, 2, \ldots\}$, the nonnegative integers, but with general birth rates λ_i, for $i \geq 0$ and general death rates μ_i, for $i \geq 1$. Since whenever the process goes from state i to state $i+1$ it must make the transition from state $i+1$ to i before it can make the transition from state i to state $i+1$ again, we may expect that for any state i, the rate of flow from state i to state $i+1$ is equal to the rate of flow from state $i+1$ to i, when the process is stationary. The local balance equations are given by

$$\pi_i \lambda_i = \pi_{i+1} \mu_{i+1},$$

for $i \geq 0$ (all the other local balance equations are trivially satisfied since $q_{ij} = q_{ji} = 0$ if $j \neq i-1, i+1$). Thus, we have that $\pi_{i+1} = (\lambda_i / \mu_{i+1}) \pi_i$. Solving recursively, we obtain

$$\begin{aligned} \pi_{i+1} &= \frac{\lambda_i}{\mu_{i+1}} \pi_i \\ &= \frac{\lambda_i \lambda_{i-1}}{\mu_{i+1} \mu_i} \pi_{i-1} \\ &\vdots \\ &= \frac{\lambda_i \ldots \lambda_0}{\mu_{i+1} \ldots \mu_1} \pi_0. \end{aligned}$$

The stationary distribution $\boldsymbol{\pi}$ will exist if and only if we can normalize this solution to the local balance equations, which will be possible if and only if

$$1 + \sum_{i=1}^{\infty} \frac{\lambda_{i-1} \ldots \lambda_0}{\mu_i \ldots \mu_1} < \infty.$$

Assuming the above sum is finite, then the normalization constraint $\sum_{i=0}^{\infty} \pi_i = 1$ is equivalent to

$$\pi_0 \left[1 + \sum_{i=1}^{\infty} \frac{\lambda_{i-1} \dots \lambda_0}{\mu_i \dots \mu_1} \right] = 1,$$

which implies that

$$\pi_0 = \left[1 + \sum_{i=1}^{\infty} \frac{\lambda_{i-1} \dots \lambda_0}{\mu_i \dots \mu_1} \right]^{-1}.$$

Then we obtain

$$\pi_i = \frac{\lambda_{i-1} \dots \lambda_0}{\mu_i \dots \mu_1} \left[1 + \sum_{i=1}^{\infty} \frac{\lambda_{i-1} \dots \lambda_0}{\mu_i \dots \mu_1} \right]^{-1},$$

for $i \geq 1$. □

Example: $M/M/1$ *Queue:* As our final example for today, we will consider the $M/M/1$ queue, which is one of the most basic queueing models in queueing theory (which we will cover in more detail next week). This is a model for a single server system to which customers arrive, are served in a first-come first-served fashion by the server, and then depart the system upon finishing service. Customers that arrive to a nonempty system will wait in a queue for service. The canonical example is a single teller bank queue. The notation "$M/M/1$" is an example of something called *Kendall's notation*, which is a shorthand for describing most queueing models. The first entry (the first "M") is a letter which denotes the arrival process to the queue. The "M" stands for "Markov" and it denotes a Poisson arrival process to the system. That is, customers arrive to the system according to a Poisson process with some rate $\lambda > 0$. The second entry (the second "M")

is a letter which denotes the service time distribution. The "M" here, which also stands for "Markov", denotes exponentially distributed service times. As well, unless explicitly stated, the implicit assumption is that service times are independent and identically distributed. Thus, the second M signifies that all service times are independent and identically distributed Exponential random variables with some rate $\mu > 0$. It is also implicitly assumed that the service times are independent of the arrival process. Finally, the third entry (the "1") is a number which denotes the number of servers in the system.

If $X(t)$ denotes the number of customers in the system at time t, then since the customer interarrival times and the service times are all independent, exponentially distributed random variables, the process $\{X(t) : t \geq 0\}$ is a continuous-time Markov chain. The state space is $S = \{0, 1, 2, \ldots\}$. Indeed, it is not hard to see that $\{X(t) : t \geq 0\}$ is a birth/death process with birth rates $\lambda_i = \lambda$, for $i \geq 0$, and death rates $\mu_i = \mu$, for $i \geq 1$. Thus, we may simply plug in these birth and death rates into our previous example. The condition for the stationary distribution to exist becomes

$$1 + \sum_{i=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^i < \infty,$$

or

$$\sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i < \infty.$$

The sum on the left hand side is just a Geometric series, and so it converges if and only if $\lambda/\mu < 1$. This condition for the stationary distribution to exist is equivalent to $\lambda < \mu$, and has the intuitive interpretation that the arrival rate to the system, λ, must be less than

the service rate of the server, μ. If $\lambda > \mu$ then customers are arriving to the system at a faster rate than the server can serve them, and the number in the system eventually blows up to ∞. In the language of queueing theory, a queueing system in which the number of customers in the system blows up to ∞ is called *unstable*, and the condition $\lambda < \mu$ is called a *stability* condition. Note that when $\lambda = \mu$ the system is also unstable, in the sense that no stationary distribution exists.

If the condition $\lambda < \mu$ is satisfied, then we obtain from the general solution in the previous example that

$$\pi_0 = \left[\sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i\right]^{-1} = \left[\frac{1}{1-\lambda/\mu}\right]^{-1} = 1 - \frac{\lambda}{\mu},$$

and

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \left(1 - \frac{\lambda}{\mu}\right),$$

for $i \geq 1$. So we see that the stationary distribution of the number in the system for an $M/M/1$ queue is a *Geometric* distribution with parameter $1 - \lambda/\mu$ (though it has the version of the Geometric distribution which is usually interpreted as the number of *failures* before the first success rather than the number of *trials* until the first success, so that it is a distribution on $\{0, 1, 2, \ldots\}$ rather than on $\{1, 2, \ldots\}$).

Next we will reconsider the notion of *time reversibility*, which we first encountered with discrete-time Markov chains, and see that there is a strong connection between time reversibility and the local balance equations, much as there was for discrete-time Markov chains.

30

Time Reversibility

The notion of *time reversibility* for continuous-time Markov chains is essentially the same as for discrete-time Markov chains. The concept of running a stochastic process backwards in time applies equally well in continuous time as in discrete time. However, even though we can imagine running any stochastic process backwards in time, the notion of time reversibility applies only to stationary processes. Therefore, we begin by assuming that $\{X(t) : -\infty < t < \infty\}$ is a stationary, continuous-time Markov chain, where we extend the time index back to $-\infty$ to accommodate running the process backwards in time. The *reversed* process is $Y(t) = X(-t)$. The first thing we need to see is that the reversed process $\boldsymbol{Y}$ is also a continuous-time Markov chain.

But this fact follows almost directly from the fact that a reversed discrete-time Markov chain is also a Markov chain. In the continuous-time chain, it is clear that whether we run the chain forwards in time or backwards in time, the amount of time we spend in any state i when we enter it has the same distribution, namely Exponential(v_i). The times we enter state i in the forward chain are the times we leave state i in the reversed chain, and vice-versa, but the times spent in state i are still distributed the same.

Therefore, the reversed chain $\boldsymbol{Y}$ will be a continuous-time Markov chain if the embedded jump process of the reversed process is a discrete-time Markov chain. But this embedded discrete-time process is just the reversed process of the forward embedded jump chain, and since the forward embedded jump chain is a discrete-time Markov chain, the embedded jump process in the reversed process is also a discrete-time Markov chain. We know this from our discussions of time reversibility in the discrete-time setting from Chapter 4 (see Lecture 18). So we may conclude that the reversed continuous-time process $\boldsymbol{Y}$ is indeed a continuous-time Markov chain.

Definition: A continuous-time Markov chain $\boldsymbol{X} = \{X(t) : t \geq 0\}$ is *time reversible* if $\boldsymbol{X}$ has a stationary distribution $\boldsymbol{\pi}$ (and so can be made stationary by setting the initial distribution of $X(0)$ to be $\boldsymbol{\pi}$), and when the stationary process is extended to the whole real line to obtain the stationary process $\{X(t) : -\infty < t < \infty\}$, the reversed process $\boldsymbol{Y} = \{Y(t) = X(-t) : -\infty < t < \infty\}$ is probabilistically the same continuous-time Markov chain as $\boldsymbol{X}$ (i.e. has the same transition rates). Equivalently, from our discussion above, $\boldsymbol{X}$ is time reversible if the embedded discrete-time jump chain of $\boldsymbol{X}$ is time reversible (i.e. the embedded jump chain of the reversed process $\boldsymbol{Y}$ has the same one-step transition probability matrix as that of the embedded jump chain of the forward process $\boldsymbol{X}$).

Let $\mathbf{P} = ((p_{ij}))_{i,j \in S}$ denote the transition probability matrix of the embedded jump chain of the forward process $\boldsymbol{X}$, and let $\boldsymbol{\psi} = (\psi_i)_{i \in S}$ denote the stationary distribution of this embedded jump chain (so that $\boldsymbol{\psi}$ satisfies the global balance equations $\boldsymbol{\psi} = \boldsymbol{\psi}\mathbf{P}$).

We saw previously that the relationship between $\boldsymbol{\psi}$ and the stationary distribution of $\boldsymbol{X}$, denoted by $\boldsymbol{\pi}$, is given by

$$\psi_i = C\pi_i v_i$$

for all $i \in S$, where C is an appropriate normalizing constant and v_i is the rate out of state i. We also have our basic relationship that

$$p_{ij} = \frac{q_{ij}}{v_i}$$

for all $i, j \in S$, where q_{ij} is the transition rate from state i to state j. From our discussions on time reversibility for discrete-time Markov chains (see Lecture 18), we know that the embedded jump chain of the forward chain $\boldsymbol{X}$ will be time reversible if and only if the stationary distribution $\boldsymbol{\psi}$ of this jump chain satisfies the local balance equations

$$\psi_i p_{ij} = \psi_j p_{ji},$$

for all $i, j \in S$. But from the relationships above between ψ_i and π_i and between p_{ij} and q_{ij}, these local balance equations for $\boldsymbol{\psi}$ are equivalent to

$$(C\pi_i v_i)\left(\frac{q_{ij}}{v_i}\right) = (C\pi_j v_j)\left(\frac{q_{ji}}{v_j}\right),$$

for all $i, j \in S$. Now canceling out C, v_i and v_j gives the equivalent equations

$$\pi_i q_{ij} = \pi_j q_{ji},$$

for all $i, j \in S$. But note that these are exactly the local balance equations for the stationary distribution $\boldsymbol{\pi}$.

In other words, we conclude from the preceding discussion that a continuous-time Markov chain $\boldsymbol{X}$ is time reversible if and only if it has a stationary distribution $\boldsymbol{\pi}$ which satisfies the local balance equations

$$\pi_i q_{ij} = \pi_j q_{ji},$$

discussed in the last lecture. Note that this gives us a way to show that a given continuous-time Markov chain $\boldsymbol{X}$ is time reversible. If we can solve the local balance equations to find the stationary distribution $\boldsymbol{\pi}$, then this not only gives us a more convenient way to determine $\boldsymbol{\pi}$. It also shows that $\boldsymbol{X}$ is time reversible, almost as a side-effect. One part of Problem 8 on Assignment #6 (Ross, p.359 #36 in the 7th Edition and p.348 #36 in the 6th Edition) asks you to show that a given continuous-time Markov chain is time reversible, and this is how you may show it. The continuous-time Markov chain in this problem is also multi-dimensional. As you might imagine, if this process were not time reversible, so that one had to solve the global balance equations to find the stationary distribution $\boldsymbol{\pi}$, finding the stationary distribution might prove quite daunting. Lucky for us, the multi-dimensional process in this problem is time reversible and so may be obtained by solving the local balance equations, which are typcially simpler than the global balance equations as discussed in the previous lecture. Despite this, even the local balance equations may not look all that trivial to you to solve, especially when the states i and j represent the vector-valued states of a multi-dimensional process. In practice (at least for this course and, to a significant extent, for a great variety of modeling situations in the "real world"), the local balance equations when the underlying process is multi-dimensional can often be solved by inspection, by determining that the stationary probabilities must have a certain form, by trial and error. The only way to develop your

sense of what the form of the stationary distribution should be in these situations is to do problems and see examples (i.e. experience), so let's do one such example now.

Example: (Ross, #31 in Chapter 6): Consider a system with r servers, where the ith service times for the ith server are independent and identically distributed Exponential(μ_i) random variables, for $i = 1, \ldots, r$ (and the service times at different servers are also independent). A total of N customers move about among these servers as follows. Whenever a customer finishes service at a server, it moves to a different server at random. That is, if a customer has just finished service at server i then it will next move to server j, where $j \neq i$, with probability $1/(r-1)$. Each server also has a queue for waiting customers and the service discipline is first-come, first-served. Let $X(t) = (n_1(t), \ldots, n_r(t))$ denote the number of customers at each server at time t (that is, $n_i(t)$ is the number of customers at server i at time t). Then $\{X(t) : t \geq 0\}$ is a continuous-time Markov chain. Show that this chain is time reversible and find the stationary distribution.

Solution: Basically, we need to set up the local balance equations and solve them, which does two things: i) it shows that the local balance equations have a solution and thus that the process is time reversible and ii) it gives us the stationary distribution. Firstly, let us note that the state space of the process is

$$S = \{(n_1, \ldots, n_r) : \text{the } n_i \text{ are nonnegative integers and } \textstyle\sum_{i=1}^{r} n_i = N\}.$$

Now let us consider the transition rates for $\{X(t) : t \geq 0\}$. To ease the writing we first define some convenient notation. Let $\boldsymbol{n}$ denote an arbitrary vector $(n_1, \ldots, n_r) \in S$ and let $\boldsymbol{e}_i$ denote the

r-dimensional vector which has a 1 for the ith component and a 0 for every other component. Now suppose that we are currently in state $\boldsymbol{n}$. We will jump to a new state as soon as a customer finishes service at some server and moves to a different server. Thus, from state $\boldsymbol{n}$ we will next jump to a state which is of the form $\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j$, where $i \neq j$ and $n_i > 0$. As $\boldsymbol{n}$ ranges over S this accounts for all the possible transitions that can occur. The transition from state $\boldsymbol{n}$ to state $\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j$ occurs when the customer at server i finishes service and then moves to server j, and this occurs at rate $\mu_i/(r-1)$. That is, the transition rate $q_{\boldsymbol{n},\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j}$ from state $\boldsymbol{n}$ to state $\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j$ is given by

$$q_{\boldsymbol{n},\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j} = \frac{\mu_i}{r-1},$$

for $i \neq j$ and for all $\boldsymbol{n}$ such that $n_i > 0$. Similarly, the transition from state $\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j$ to state $\boldsymbol{n}$ occurs only when the customer at server j finishes service and then moves to server i, and this occurs at rate $\mu_j/(r-1)$. That is,

$$q_{\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j,\boldsymbol{n}} = \frac{\mu_j}{r-1},$$

Thus, our local balance equations

$$\pi_{\boldsymbol{n}} q_{\boldsymbol{n},\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j} = \pi_{\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j,\boldsymbol{n}} q_{\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j,\boldsymbol{n}},$$

where $\pi_{\boldsymbol{n}}$ is the stationary probability of state $\boldsymbol{n}$, are given by

$$\pi_{\boldsymbol{n}} \frac{\mu_i}{r-1} = \pi_{\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j,\boldsymbol{n}} \frac{\mu_j}{r-1},$$

for $i \neq j$ and all $\boldsymbol{n} \in S$ such that $n_i > 0$. We may cancel out the $r-1$ from both sides of the above equations to obtain

$$\pi_{\boldsymbol{n}} \mu_i = \pi_{\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j,\boldsymbol{n}} \mu_j,$$

for $i \neq j$ and all $\boldsymbol{n} \in S$ such that $n_i > 0$.

As mentioned, it may not seem obvious what form $\pi_{\boldsymbol{n}}$ should have in order to satisfy these equations. On the other hand, the equations certainly look simple enough that one should believe that $\pi_{\boldsymbol{n}}$ might have some simple, regular form. In words, $\pi_{\boldsymbol{n}}$ should be some function of $\boldsymbol{n}$ such that when we multiply it by μ_i, that is the same thing as taking this function evaluated at $\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j$ and multiplying that by μ_j. A little inspection, and perhaps some trial and error, will lead to $\pi_{\boldsymbol{n}}$ of the form

$$\pi_{\boldsymbol{n}} = \frac{C}{\mu_1^{n_1} \dots \mu_r^{n_r}},$$

for all $\boldsymbol{n} \in S$, where C is the appropriate normalizing constant. We can verify that this claimed form for $\pi_{\boldsymbol{n}}$ is correct by plugging it into the local balance equations. On the left hand side we obtain

$$\text{LHS} = \frac{C\mu_i}{\mu_1^{n_1} \dots \mu_r^{n_r}},$$

while on the right hand side we obtain

$$\text{RHS} = \frac{C\mu_j}{\mu_1^{n_1} \dots \mu_r^{n_r}} \times \frac{\mu_i}{\mu_j},$$

where the factor μ_i/μ_j is needed to account for the fact that there is one less customer at server i and one more customer at server j in the state $\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j$ relative to state $\boldsymbol{n}$. Clearly both the LHS and the RHS are equal for all $i \neq j$ and all $\boldsymbol{n} \in S$ such that $n_i > 0$. Since the state space S is finite the normalizing constant C is strictly positive and given by

$$C = \left[\sum_{\boldsymbol{m} \in S} \frac{1}{\mu_1^{m_1} \dots \mu_r^{m_r}} \right]^{-1},$$

where $\boldsymbol{m} = (m_1, \dots, m_r)$ ranges over all states in S.

Thus, the stationary distribution $\boldsymbol{\pi} = (\pi_{\boldsymbol{n}})_{\boldsymbol{n} \in S}$ is given by

$$\pi_{\boldsymbol{n}} = \frac{1}{\mu_1^{n_1} \dots \mu_r^{n_r}} \left[\sum_{\boldsymbol{m} \in S} \frac{1}{\mu_1^{m_1} \dots \mu_r^{m_r}} \right]^{-1},$$

for all $\boldsymbol{n} \in S$. So we have found the stationary distribution $\boldsymbol{\pi}$ and, since we have also shown that $\boldsymbol{\pi}$ satisfies the local balance equations (since that is how we found $\boldsymbol{\pi}$), we have also shown that the continuous-time Markov chain $\{X(t) : t \geq 0\}$ is time reversible. $\square$