1 Continuous-Time Markov Chains

A Markov chain in discrete time, $\{X_n : n \geq 0\}$, remains in any state for exactly one unit of time before making a transition (change of state). We proceed now to relax this restriction by allowing a chain to spend a continuous amount of time in any state, but in such a way as to retain the Markov property. As motivation, suppose we consider the rat in the maze Markov chain. Clearly it is more realistic to be able to keep track of where the rat is at any continuous-time $t \geq 0$ as oppposed to only where the rat is after n "steps".

Assume throughout that our state space is $\mathcal{S} = \mathbb{Z} = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$ (or some subset thereof). Suppose now that whenever a chain enters state $i \in \mathcal{S}$, independent of the past, the length of time spent in state i is a continuous, strictly positive (and proper) random variable H_i called the *holding time* in state i. When the holding time ends, the process then makes a transition into state j according to transition probability P_{ij}, independent of the past, and so on.[1] Letting $X(t)$ denote the state at time t, we end up with a continuous-time stochastic process $\{X(t) : t \geq 0\}$ with state space $\mathcal{S}$.

Our objective is to place conditions on the holding times to ensure that the continuous-time process satisfies the Markov property: *The future, $\{X(s+t) : t \geq 0\}$, given the present state, $X(s)$, is independent of the past, $\{X(u) : 0 \leq u < s\}$.* Such a process will be called a continuous-time Markvov chain (CTMC), and as we will conclude shortly, the holding times will have to be exponentially distributed.

The formal definition is given by

Definition 1.1 *A stochastic process $\{X(t) : t \geq 0\}$ with discrete state space $\mathcal{S}$ is called a continuous-time Markvov chain (CTMC) if for all $t \geq 0,\ s \geq 0,\ i \in \mathcal{S},\ j \in \mathcal{S}$,*

$$P(X(s+t) = j | X(s) = i, \{X(u) : 0 \leq u < s\}) = P(X(s+t) = j | X(s) = i) = P_{ij}(t).$$

$P_{ij}(t)$ is the probability that the chain will be in state j, t time units from now, given it is in state i now.

For each $t \geq 0$ there is a transition matrix

$$P(t) = (P_{ij}(t)),\ i, j \in \mathcal{S},$$

and $P(0) = I$, the identity matrix.

As for discrete-time Markov chains, we are assuming here that the distribution of the future, given the present state $X(s)$, does not depend on the present time s, but only on

[1] $P_{ii} > 0$ is allowed, meaning that a transition back into state i from state i can ocurr. Each time this happens though, a new H_i, independent of the past, determines the new length of time spent in state i. See Section 1.14 for details.

the present state $X(s) = i$, whatever it is, and the amount of time that has elapsed, t, since time s. In particular, $P_{ij}(t) = P(X(t) = j|X(0) = i)$.

But unlike the discrete-time case, there is no smallest "next time" until the next transition, there is a continuum of such possible times t. For each fixed i and j, $P_{ij}(t),\ t \geq 0$ defines a function which in principle can be studied by use of calculus and differential equations. Although this makes the analysis of CTMC's more difficult/technical than for discrete-time chains, we will, non-the-less, find that many similarities with discrete-time chains follow, and many useful results can be obtained.

A little thought reveals that the holding times must have the memoryless property and thus are exponentially distributed. To see this, suppose that $X(t) = i$. Time t lies somewhere in the middle of the holding time H_i for state i. The future after time t tells us, in particular, the remaining holding time in state i, whereas the past before time t, tells us, in particular, the age of the holding time (how long the process has been in state i). In order for the future to be independent of the past given $X(t) = i$, we deduce that the remaining holding time must only depend (in distribution) on i and be independent of its age; the memoryless property follows. Since an exponential distribution is completely determined by its rate we conclude that for each $i \in \mathcal{S}$, there exists a constant (rate) $a_i > 0$, such that the chain, when entering state i, remains there, independent of the past, for an amount of time $H_i \sim exp(a_i)$:

> *A CTMC makes transitions from state to state, independent of the past, according to a discrete-time Markov chain, but once entering a state remains in that state, independent of the past, for an exponentially distributed amount of time before changing state again.*

Thus a CTMC can simply be described by a transition matrix $P = (P_{ij})$, describing how the chain changes state step-by-step at transition epochs, together with a set of rates $\{a_i : i \in \mathcal{S}\}$, the holding time rates. Each time state i is visited, the chain spends, on average, $E(H_i) = 1/a_i$ units of time there before moving on.

1.1 The embedded discrete-time Markov chain

Letting τ_n denote the time at which the n^{th} change of state (transition) occurs, we see that $X_n = X(\tau_n+)$, the state *right after* the n^{th} transition, defines the underlying discrete-time Markov chain, called the *embedded Markov chain*. $\{X_n\}$ keeps track, consecutively, of the states visited right after each transition, and moves from state to state according to the one-step transition probabilities $P_{ij} = P(X_{n+1} = j|X_n = i)$. This transition matrix (P_{ij}), together with the holding-time rates $\{a_i\}$, completely determines the CTMC.

1.2 Chapman-Kolmogorov equations

The Chapman-Kolmogorov equations for discrete-time Markov chains generalizes to

Lemma 1.1 (Chapman-Kolmogorov equations for CTMC's) *For all $t \geq 0,\ s \geq 0$,*

$$P(t+s) = P(s)P(t),$$

that is, for all $t \geq 0,\ s \geq 0,\ i \in \mathcal{S},\ j \in \mathcal{S}$

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

As for discrete-time chains, the (easy) proof involves first conditioning on what state k the chain is in at time s given that $X(0) = i$, yielding $P_{ik}(s)$, and then using the Markov property to conclude that the probability that the chain, now in state k, would then be in state j after an additional t time units is, independent of the past, $P_{kj}(t)$.

1.3 Examples of CTMC's

1. *Poisson counting process:* Let $\{N(t) : t \geq 0\}$ be the counting process for a Poisson process $\psi = \{t_n\}$ at rate λ. Then $\{N(t)\}$ forms a CTMC with $\mathcal{S} = \{0, 1, 2, \ldots\}$, $P_{i,i+1} = 1,\ a_i = \lambda,\ i \geq 0$: If $N(t) = i$ then, by the memoryless property, the next arrival, arrival $i + 1$, will, independent of the past, occur after an exponentially distributed amount of time at rate λ. The holding time in state i is simply the interarrival time, $t_{i+1} - t_i$, and $\tau_n = t_n$ since $N(t)$ only changes state at an arrival time. Assuming that $N(0) = 0$ we conclude that $X_n = N(t_n+) = n,\ n \geq 0$; the embedded chain is deterministic. This is a very special kind of CTMC for several reasons. (1) all holding times H_i have the same rate $a_i = \lambda$, and (2) $N(t)$ is a non-decreasing process; it increases by one at each arrival time, and remains constant otherwise. As $t \to \infty$, $N(t) \uparrow \infty$ step by step.

2. Consider the rat in the closed maze, in which at each transition, the rat is equally likely to move to one of the neighboring two cells, but where now we assume that the holding time, H_i, in cell i is exponential at rate $a_i = i,\ i = 1, 2, 3, 4$. Time is in minutes (say). Let $X(t)$ denote the cell that the rat is in at time t. Given the rat is now in cell 2 (say), he will remain there, independent of the past, for an exponential amount of time with mean 1/2, and then move, independent of the past, to either cell 1 or 4 w.p.=1/2. The other transitions are similarly explained. $\{X(t)\}$ forms a CTMC. Note how cell 4 has the shortest holding time (mean 1/4 minutes), and cell 1 has the longest (mean 1 minute). Of intrinisic interest is to calculate the long-run proportion of time (continuous time now) that the rat spends in each cell;

$$P_i \overset{\text{def}}{=} \lim_{t \to \infty} \frac{1}{t} \int_0^t I\{X(s) = i\}ds,\ i = 1, 2, 3, 4.$$

We will learn how to compute these later; they serve as the continuous-time analog to the discrete-time stationary probabilities π_i for discrete-time Markov chains.

$\vec{P} = (P_1, P_2, P_3, P_4)$ is called the limiting (stationary) distribution for the CTMC. The intuitive interpretation: If way out in the future we were to observe the maze, then P_i is the probability that we would find the rat in cell i.

3. *FIFO M/M/1 queue:* Arrivals to a single-server queue are Poisson at rate λ. There is one line (queue) to wait in, and customers independently (and independent of the Poisson arrival process) have service times $\{S_n\}$ that are exponentially distributed at rate μ. We assume that customers join the tail of the queue, and hence begin service in the order that they arrive *first-in-queue-first-out-of-queue (FIFO)*. Let $X(t)$ denote the number of customers in the system at time t, where "system" means the line plus the service area. So (for example), $X(t) = 2$ means that there is one customer in service and one waiting in line. Note that a transition can only occur at customer arrival or departure times, and that departures occur whenever a service completion occurs. At an arrival time $X(t)$ jumps up by the amount 1, whereas at a departure time $X(t)$ jumps down by the amount 1.

   Determining the rates a_i: If $X(t) = 0$ then only an arrival can occur next, so the holding time is given by $H_0 \sim exp(\lambda)$ the time until the next arrival; $a_0 = \lambda$, the arrival rate. If $X(t) = i \geq 1$, then the holding time is given by $H_i = \min\{S_r, X\}$ where S_r is the remaining service time of the customer in service, and X is the time until the next arrival. The memoryless property for both service times and interarrival times implies that $S_r \sim exp(\mu)$ and $X \sim exp(\lambda)$ independent of the past. Also, they are independent r.v.s. because the service times are assumed independent of the Poisson arrival process. Thus $H_i \sim exp(\lambda + \mu)$ and $a_i = \lambda + \mu,\ i \geq 1$. The point here is that each of the two independent events "service completion will ocurr", "new arrival will ocurr" is competing to be the next event so as to end the holding time.

   The transition probabilities P_{ij} for the embedded discrete-time chain are derived as follows: X_n denotes the number of customers in the system right after the n^{th} transition. Transitions are caused only by arrivals and departures.

   If $X_n = 0$, then the system is empty and we are waiting for the next arrival; $P(X_{n+1} = 1 | X_n = 0) = 1$. But if $X_n = i \geq 1$, then $X_{n+1} = i + 1$ w.p. $P(X < S_r) = \lambda/(\lambda + \mu)$, and $X_{n+1} = i - 1$ w.p. $P(S_r < X) = \mu/(\lambda + \mu)$, depending on whether an arrival or a departure is the first event to occur next. So, $P_{0,1} = 1$, and for $i \geq 1$, $P_{i,i+1} = p = \lambda/(\lambda + \mu)$, and $P_{i,i-1} = 1 - p = \mu/(\lambda + \mu)$. We conclude that

   > *The embedded Markov chain for a FIFO M/M/1 queue is a simple random walk ("up" probability $p = \lambda/(\lambda + \mu)$, "down" probability $1 - p = \mu/(\lambda + \mu)$) that is restricted to be non-negative ($P_{0,1} = 1$).*

4. *M/M/c multi-server queue:* This is the same as the FIFO M/M/1 queue except there are now c servers working in parallel. As in a USA postoffice, arrivals wait in one FIFO line (queue) and enter service at the first available free server. $X(t)$

denotes the number of customers in the system at time t. For illustration, let's assume $c = 2$. Then, for example, $X(t) = 4$ means that two customers are in service (each with their own server) and two others are waiting in line. When $X(t) = i \in \{0, 1\}$, the holding times are the same as for the M/M/1 model; $a_0 = \lambda,\ a_1 = \lambda + \mu$. But when $X(t) = i \geq 2$, both remaining service times, denoted by S_{r_1} and S_{r_2}, compete to determine the next departure. Since they are independent exponentials at rate μ, we deduce that the time until the next departure is given by $\min\{S_{r_1}, S_{r_2}\} \sim exp(2\mu)$. The time until the next arrival is given by $X \sim exp(\lambda)$ and is independent of both remaining service times. We conclude that the holding time in any state $i \geq 2$ is given by $H_i = \min\{X, S_{r_1}, S_{r_2}\} \sim exp(\lambda + 2\mu)$.

For the general case of $c \geq 2$, the rates are determined analogously: $a_i = \lambda + i\mu,\ 0 \leq i \leq c,\ a_i = \lambda + c\mu,\ i > c$.

For the embedded chain: $P_{0,1} = 1$ and for $0 \leq i \leq c-1$, $P_{i,i+1} = \lambda/(\lambda + i\mu),\ P_{i,i-1} = i\mu/(\lambda + i\mu)$. Then for $i \geq c$, $P_{i,i+1} = \lambda/(\lambda + c\mu),\ P_{i,i-1} = c\mu/(\lambda + c\mu)$. This is an example of a simple random walk with state-dependent "up", "down" probabilities: at each step, the probabilities for the next increment depend on i, the current state, until $i = c$ at which point the probabilities remain constant.

5. *M/M/∞ infinite-server queue:* Here we have a M/M/c queue with $c = \infty$; a special case of the M/G/∞ queue. Letting $X(t)$ denote the number of customers in the system at time t, we see that $a_i = \lambda + i\mu,\ i \geq 0$ since there is no limit on the number of busy servers.

   For the embedded chain: $P_{0,1} = 1$ and $P_{i,i+1} = \lambda/(\lambda + i\mu),\ P_{i,i-1} = i\mu/(\lambda + i\mu),\ i \geq 1$. This simple random walk thus has state-dependent "up", "down" probabilities that continue to depend on each state i, the current state. Note how, as i increases, the down probability, $P_{i,i-1}$, increases, and approaches 1 as $i \to \infty$: when the system is heavily congested, departures occur rapidly; this model is always *stable.*

1.4 Birth and Death processes

Except for Example 2 (rat in the closed maze) all of the CTMC examples in the previous section were *Birth and Death (B&D)* processes, CTMC's that can only change state by increasing by one, or decreasing by one; $P_{i,i+1} + P_{i,i-1} = 1,\ i \in \mathcal{S}$. (In Example 2, $P_{1,3} > 0$, for example, so it is not B&D.) Here we study B&D processes more formally, since they tend to be a very useful type of CTMC. Whenever the state increases by one, we say there is a *birth*, and whenever it decreases by one we say there is a *death.* We shall focus on the case when $\mathcal{S} = \{0, 1, 2, \ldots\}$, in which case $X(t)$ can be thought of as the *population size* at time t.

For each state $i \geq 0$ we have a birth rate λ_i and a death rate μ_i: Whenever $X(t) = i$, independent of the past, the time until the next birth is a r.v. $X \sim exp(\lambda_i)$ and, independently, the time until the next death is a r.v. $Y \sim exp(\mu_i)$. Thus the holding time rates are given by $a_i = \lambda_i + \mu_i$ because the time until the next transition (change of state) in given by the holding time $H_i = \min\{X, Y\} \sim exp(\lambda_i + \mu_i)$. The idea here is that at any given time the next birth is competing with the next death to be the next transition. (We always assume here that $\mu_0 = 0$ since there can be no deaths without a population.)

This means that whenever $X(t) = i \geq 1$, the next transition will be a birth w.p. $P_{i,i+1} = P(X < Y) = \lambda_i/(\lambda_i + \mu_i)$, and a death w.p. $P_{i,i-1} = P(Y < X) = \mu_i/(\lambda_i + \mu_i)$. *Thus the embedded chain for a B&D process is a simple random walk with state dependent "up", "down" probabilities.*

When $\mu_i = 0,\ i \geq 0$, and $\lambda_i > 0,\ i \geq 0$, we call the process a pure birth process; the population continues to increase by one at each transition. The main example is the Poisson counting process (Example 1 in the previous Section), but this can be generalized by allowing each λ_i to be different. The reader is encouraged at this point to go back over the B&D Examples in the previous Section.

1.5 Explosive CTMC's

Consider a pure birth process $\{X(t)\}$, $P_{i,i+1} = 1,\ i \geq 0$, in which $a_i = \lambda_i = 2^i,\ i \geq 0$. This process spends, on average, $E(H_i) = 1/\lambda_i = 2^{-i}$ units of time in state i and then changes to state $i+1$; . Thus it spends less and less time in each state, consequently jumping to the next state faster and faster as time goes on. Since $X(t) \to \infty$ as $t \to \infty$, we now explore how fast this happens. Note that the chain will visit state i at time $H_0 + H_1 + \cdots + H_{i-1}$, the sum of the first i holding times. Thus the chain will visit *all* of the states by time

$$T = \sum_{i=0}^{\infty} H_i.$$

Taking expected value yields

$$E(T) = \sum_{i=0}^{\infty} 2^{-i} = 2 < \infty,$$

and we conclude that on average all states $i \geq 0$ have been visited by time $t = 2$, a finite amount of time! But this implies that w.p.1., all states will be visited in a finite amount of time; $P(T < \infty) = 1$. Consequently, w.p.1., $X(T+t) = \infty,\ t \geq 0$. This is an example of an *explosive* Markov chain: The number of transitions in a finite interval of time is infinite.

We shall rule out this kind of behavior in the rest of our study, and assume from now on that all CTMC's considered are non-explosive, by which we mean that the number of transitions in any finite interval of time is finite. This will always hold for any CTMC with a finite state space, or any CTMC for which there are only a finite number of distinct values for the rates a_i, and more generally whenever $\sup\{a_i : i \in \mathcal{S}\} < \infty$. Every Example given in the previous Section was non-explosive. Only the M/M/∞ queue needs some clarification since $a_i = \lambda + i\mu \to \infty$ as $i \to \infty$. But only arrivals and departures determine transitions, and the arrivals come from the Poisson process at fixed rate λ, so the arrivals can not cause an explosion; $N(t) < \infty,\ t \geq 0$. Now observe that during any interval of time, $(s, t]$, the number of departures can be no larger than $N(t)$, the total number of arrivals thus far, so they too can not cause an explosion. In short, the number of transitions in any interval $(s, t]$ is bounded from above by $2N(t) < \infty$; the non-explosive condition is satisfied. This method of bounding the number of transitions by the underlying Poisson arrival process will hold for essentially any CTMC queueing model.

1.6 Communication classes, irreducibility and recurrence

State j is said to be reachable from state i for a CTMC if $P(X(s) = j|X(0) = i) = P_{ij}(s) > 0$ for some $s \geq 0$. As with discrete-time chains, i and j are said to communicate if state j is reachable from state i, and state i is reachable from state j.

It is immediate that i and j communicate in continuous time if and only if they do so for the embedded discrete-time chain $\{X_n\}$: They communicate in continuous-time if and only if they do so at transition epochs. Thus once again, we can partition the state space up into disjoint communication classes, $\mathcal{S} = C_1 \cup C_2 \cup \cdots$, and an irreducible chain is a chain for which all states communicate ($\mathcal{S} = C_1$, one communication class). We state in passing

A CTMC is irreducible if and only if its embedded chain is irreducible.

Notions of recurrence, transience and positive recurence are similar as for discrete-time chains: Let $T_{i,i}$ denote the amount of (continuous) time until the chain re-visits state i (at a later transition) given that $X(0) = i$ (defined to be ∞ if it never does return); the return time to state i. The chain will make its first transition at time H_i (holding time in state i), so $T_{ii} \geq H_i$. State i is called recurrent if, w.p.1., the chain will re-visit state i with certainty, that is, if $P(T_{ii} < \infty) = 1$. The state is called transient otherwise. This (with a little thought) is seen to be the same property as for the embedded chain (because $X(t)$ returns to state i for some t if and only if X_n does so for some n):

A state i is recurrent/transient for a CTMC if and only if it is recurrent/transient for the embedded discrete-time chain.

Thus communication classes all have the same type of states: all together they are transient or all together they are recurrent.

1.7 Positive recurrence and the existence a limiting distribution $\vec{P} = \{P_j\}$

State i is called positive recurrent if, in addition to being recurrent, $E(T_{ii}) < \infty$; the expected amount of time to return is finite. State i is called null recurrent if, in addition to being recurrent, $E(T_{ii}) = \infty$; the expected amount of time to return is infinite. Unlike recurrence, positive (or null) recurrence is not equivalent to that for the embedded chain: It is possible for a CTMC to be positive recurrent while its embedded chain is null recurrent (and vice versa). But positive and null recurrence are still class properties, so in particular:

> *For an irreducible CTMC, all states together are transient, positive recurrent, or null recurrent.*

A CTMC is called positive recurrent if it is irreducible and all states are positive recurrent. We define (when they exist, independent of initial condition $X(0) = i$) the limiting probabilities $\{P_j\}$ for the CTMC as the long-run proportion of time the chain spends in each state $j \in \mathcal{S}$:

$$P_j = \lim_{t\to\infty} \frac{1}{t} \int_0^t I\{X(s) = j | X(0) = i\} ds, \text{ w.p.1.,} \tag{1}$$

which after taking expected values yields

$$P_j = \lim_{t\to\infty} \frac{1}{t} \int_0^t P_{ij}(s) ds. \tag{2}$$

When each P_j exists and $\sum_j P_j = 1$, then $\vec{P} = \{P_j\}$ (as a row vector) is called the limiting (or stationary) distribution for the Markov chain. Letting

$$\mathbf{P}^* = \begin{pmatrix} \vec{P} \\ \vec{P} \\ \vdots \end{pmatrix} \tag{3}$$

denote the matrix in which each row is the limiting probability distribution $\vec{P}$, (2) can be expressed nicely in matrix form as

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P(s) ds = \mathbf{P}^*. \tag{4}$$

As for discrete-time Markov chains, positive recurrence implies the existence of limiting probabilities by use of the SLLN. The basic idea is that for fixed state j, we can break up the evolution of the CTMC into i.i.d. cycles, where a cycle begins every time the chain makes a transition into state j. This yields an example of what is called a *regenerative process* because we say it regenerates every time a cycle begins. The cycle lengths are i.i.d. distributed as T_{jj}, and during a cycle, the chain spends an amount of time in state j equal in distribution to the holding time H_j. This leads to

Proposition 1.1 *If $\{X(t)\}$ is a positive recurrent CTMC, then the limiting probability distribution $\vec{P} = (P_{i,j})$ as defined by Equation (1) exists, is unique, and is given by*

$$P_j = \frac{E(H_j)}{E(T_{jj})} = \frac{1/a_j}{E(T_{jj})} > 0, \;\; j \in \mathcal{S}.$$

In words: "The long-run proportion of time the chain spends in state j equals the expected amount of time spent in state j during a cycle divided by the expected cycle length (between visits to state j)".

Moreover, the stronger mode of convergence (weak convergence) holds:

$$P_j = \lim_{t\to\infty} P_{ij}(t), \;\; i, j \in \mathcal{S}. \tag{5}$$

Finally, if the chain is either null recurrent or transient, then $P_j = 0, \; j \in \mathcal{S}$; no limiting distribution exists.

Proof : Fixing state j, we can break up the evolution of the CTMC into i.i.d. cycles, where a cycle begins every time the chain makes a transition into state j. This follows by the (strong) Markov property, since every time the chain enters state j, the chain starts over again from scratch stochastically, and is independent of the past. Letting $\tau_n(j)$ denote the n^{th} time at which the chain makes a transition into state j, with $\tau_0(j) = 0$, the cycle lengths, $T_n(j) = \tau_n(j) - \tau_{n-1}(j), \; n \geq 1$, are i.i.d., distributed as the return time T_{jj}. $\{\tau_n(j) : n \geq 1\}$ forms a renewal point process because of the assumed recurrence, of the chain, and we let $N_j(t)$ denote the number of such points during $(0, t]$. From the Elementary Renewal Theorem, wp1,

$$\lim_{t\to\infty} \frac{N_j(t)}{t} = \frac{1}{E(T_{jj})}. \tag{6}$$

Letting

$$J_n = \int_{\tau_{n-1}(j)}^{\tau_n(j)} I\{X(s) = j\} ds,$$

(the amount of time spent in state j during the n^{th} cycle) we conclude that $\{J_n\}$ forms an i.i.d. sequence of r.v.s. distributed as the holding time H_j; $E(J) = E(H_j)$. Thus

$$\int_0^t I\{X(s) = j\} ds \approx \sum_{n=1}^{N_j(t)} J_n,$$

from which we obtain

$$\frac{1}{t}\int_0^t I\{X(s) = j\} ds \approx \frac{N_j(t)}{t} \times \frac{1}{N_j(t)} \sum_{n=1}^{N_j(t)} J_n.$$

Letting $t \to \infty$ yields

$$P_j = \frac{E(H_j)}{E(T_{jj})},$$

where the denominator is from (6) and the numerator is from the SLLN applied to $\{J_n\}$. $P_j > 0$ if $E(T_{jj}) < \infty$ (positive recurrence), whereas $P_j = 0$ if $E(T_{jj}) = \infty$ (null recurrence). And if transient, then $I\{X(s) = j | X(0) = i\} \to 0$ as $s \to \infty$, wp1, yielding $P_j = 0$ as well from (1).

Uniqueness of the P_j follows by the unique representation, $P_j = \frac{1/a_j}{E(T_{jj})}$.

The weak convergence in (5) holds in addition to the already established time-average convergence because the cycle-length distribution (the distribution of T_{jj} for any fixed j) is non-lattice.[2] T_{jj} has a non-lattice distribution because it is of *phase type* hence a continuous distribution. In general, a positive recurrent regenerative process with a non-lattice cycle-length distribution converges weakly. The details of this will be dealt with later when we return to a more rigorous study of renewal theory. ■

1.8 Allowing an arbitrary random initial value for $X(0)$

If $\vec{\nu}$ is a probability distribution on $\mathcal{S}$, and if $X(0) \sim \vec{\nu}$, then the distribution of $X(t)$ is given by the vector-matrix product $\vec{\nu}P(t)$; $X(t) \sim \vec{\nu}P(t),\ t \geq 0$. Recalling the definition of $\mathbf{P}^*$ in (3), note that $\vec{\nu}\mathbf{P}^* = \vec{P}$ for any probability distribution $\vec{\nu}$. Thus for a positive recurrent chain it holds more generally from (4) (by multiplying each left side by $\vec{\nu}$) that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \vec{\nu}P(s)ds = \vec{P}. \tag{7}$$

This merely means that the initial condition (e.g., the value of $X(0)$) can be random as opposed to only deterministic (e.g., $X(0) = i$) without effecting the limit; the same limiting distribution $\vec{P}$ is obtained regardless.

1.9 The limiting distribution yields a stationary distribution and hence a stationary version of the Markov chain

As in discrete-time, the limiting distribution $\vec{P}$ is also called a *stationary* distribution because it yields a stationary version of the chain if the chain is initially distributed as $\vec{P}$ at time $t = 0$:

Proposition 1.2 *For a positive recurrent Markov chain with limiting distribution $\vec{P}$: If $X(0) \sim \vec{P}$, then $X(t) \sim \vec{P}$ for all $t \geq 0$; that is, $\vec{P}P(t) = \vec{P},\ t \geq 0$. This means that*

$$\sum_{i\in\mathcal{S}} P_i P_{i,j}(t) = P_j,\ j \in \mathcal{S},\ t \geq 0.$$

[2] The distribution of a non-negative rv X is said to be non-lattice if there does *not* exists a $d > 0$ such that $P(X \in \{nd : n \geq 0\}) = 1$. Any continuous distribution, in particular, is non-lattice.

In fact this limiting distribution $\vec{P}$ is the only distribution (it is unique) that is stationary, that is, for which $\vec{P}P(t) = \vec{P},\ t \geq 0$. Moreover, letting $\{X^*(t) : t \geq 0\}$ denote the chain when $X(0) \sim \vec{P}$, it forms a stationary stochastic process: $\{X^*(s+t) : t \geq 0\}$ has the same distribution for all $s \geq 0$.

Proof : From the definition of $\mathbf{P}^*$ (each row is $\vec{P}$) we must equivalently show that $\mathbf{P}^*P(t) = \mathbf{P}^*$ for any $t \geq 0$. (Intuitively we are simply asserting that $P(\infty)P(t) = P(\infty)$ because $\infty + t = \infty$.)

Recalling the Chapman-Kolmogorov equations, $P(s+t) = P(s)P(t)$, and using (4), we get

$$\begin{aligned}
\mathbf{P}^*P(t) &= \Big(\lim_{u\to\infty} \frac{1}{u}\int_0^u P(s)ds\Big)P(t) \\
&= \lim_{u\to\infty} \frac{1}{u}\int_0^u P(s)P(t)ds \\
&= \lim_{u\to\infty} \frac{1}{u}\int_0^u P(s+t)ds \\
&= \lim_{u\to\infty} \frac{1}{u}\int_0^u P(s)ds \\
&= \mathbf{P}^*.
\end{aligned}$$

The second to last equality follows due to fact that adding the fixed t is asymptotically negligible: $\int_0^u P(s+t)du = \int_0^u P(s)ds + \int_u^{u+t} P(s)ds - \int_0^t P(s)$. All elements of $P(s)$ are bounded by 1, and so the last two integrals when divided by u tend to 0 as $u \to \infty$.

If a probability distribution $\vec{\nu}$ satisfies $\vec{\nu}P(t) = \vec{\nu},\ t \geq 0$, then on the one hand, since the chain is assumed positive recurrent, we have Equation (7) and hence

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \vec{\nu}P(s)ds = \vec{\nu}\mathbf{P}^* = \vec{P}. \tag{8}$$

But on the other hand $\vec{\nu}P(t) = \vec{\nu}$ implies that

$$\frac{1}{t}\int_0^t \vec{\nu}P(s)ds = \frac{1}{t}\int_0^t \vec{\nu}ds = \vec{\nu},$$

and we conclude that $\vec{\nu} = \vec{P}$; the stationary distribution is unique.

By the Markov property, a Markov process is completely determined (in distribution) by its initial state. Thus $\{X^*(s+t) : t \geq 0\}$ has the same distribution for all $s \geq 0$ because for all s, its initial state has the same distribution, $X^*(s) \sim \vec{P}$. ■

1.10 Interpretation of the $\{a_i\}$ as transition rates; the transition rate matrix (infinitesimal generator) Q

Assume here that $P_{i,i} = 0$ for all $i \in \mathcal{S}$. a_i can be interpreted as the transition rate out of state i given that $X(t) = i$; the intuitive idea being that the exponential holding time will end, independent of the past, in the next dt units of time with probability $a_i dt$. This can be made rigorous. It can be shown that for $i \neq j$

$$P'_{i,j}(0) = \lim_{h \downarrow 0} P_{i,j}(h)/h = a_i P_{i,j}. \tag{9}$$

$a_i P_{i,j}$ can thus be interpreted as the transition rate from state i to state j given that the chain is currently in state i.

When $i = j$, $P_{i,i}(h) = 1 - P(X(h) \neq i \mid X(0) = i)$ and it can be shown that

$$P'_{i,i}(0) = \lim_{h \downarrow 0} (P_{i,i}(h) - 1)/h = -a_i. \tag{10}$$

Definition 1.2 *The matrix $Q = P'(0)$ given explicitly by (9) and (10) is called the transition rate matrix, or infinitesimal generator, of the Markov chain*

For example, if $\mathcal{S} = \{0, 1, 2, 3, 4\}$, then

$$Q = \begin{pmatrix} -a_0 & a_0 P_{0,1} & a_0 P_{0,2} & a_0 P_{0,3} & a_0 P_{0,4} \\ a_1 P_{1,0} & -a_1 & a_1 P_{1,2} & a_1 P_{1,3} & a_1 P_{1,4} \\ a_2 P_{2,0} & a_2 P_{2,1} & -a_2 & a_2 P_{2,3} & a_2 P_{2,4} \\ a_3 P_{3,0} & a_3 P_{3,1} & a_3 P_{3,2} & -a_3 & a_3 P_{3,4} \\ a_4 P_{4,0} & a_4 P_{4,3} & a_4 P_{4,3} & a_4 P_{4,3} & -a_4 \end{pmatrix}$$

Note in passing that since we assume that $P_{i,i} = 0,\ i \in \mathcal{S}$, we conclude that each row of Q sums to 0.

1.11 Computing $P_{ij}(t)$: Kolmogorov backward equations

We have yet to show how to compute the transition probabilities for a CTMC, $P_{ij}(t) = P(X(t) = j | X(0) = i),\ t \geq 0$. For discrete-time Markov chains this was not a problem since $P_{ij}^{(n)} = P(X_n = j | X_0 = i),\ n \geq 1$ could be computed by using the fact that the matrix $(P_{ij}^{(n)})$ was simply the transition matrix P multiplied together n times, P^n. In continuous time however, the problem is a bit more complex; it involves setting up linear differential equations for $P_{ij}(t)$ known as the Kolmogorov backward equations and then solving. We present this derivation now.

Proposition 1.3 (Kolmogorov Backward Equations) *For a (non-explosive) CTMC with transition rate matrix $Q = P'(0)$ as in Definition 1.2, the following set of linear differential equations is satisfied by $\{P(t)\}$:*

$$P'(t) = QP(t),\ t \geq 0,\ P(0) = I, \tag{11}$$

that is,

$$P'_{ij}(t) = -a_i P_{ij}(t) + \sum_{k \neq i} a_i P_{ik} P_{kj}(t),\ i, j \in \mathcal{S},\ t \geq 0. \tag{12}$$

The unique solution is thus of the exponential form;

$$P(t) = e^{Qt}, \ t \geq 0, \tag{13}$$

where for any square matrix M,

$$e^M \overset{\text{def}}{=} \sum_{n=0}^{\infty} \frac{M^n}{n!}.$$

Proof : The Chapman-Kolmogorov equations, $P(t+h) = P(h)P(t)$, yield

$$\begin{aligned} P(t+h) - P(t) &= (P(h) - I)P(t) & (14) \\ &= (P(h) - P(0))P(t). & (15) \end{aligned}$$

dividing by h and letting $h \to 0$ then yields $P'(t) = P'(0)P(t) = QP(t)$. (Technically, this involves justifying the interchange of a limit and an infinite sum, which indeed can be justified here even when the state space is infinite.) ■

The word *backward* refers to the fact that in our use of the Chapman-Kolmogorov equations, we chose to place the h on the right-hand side in back, $P(t+h) = P(h)P(t)$ as opposed to in front, $P(t+h) = P(t)P(h)$. The derivation above can be rigorously justified for any non-explosive CTMC.

It turns out, however, that the derivation of the analogous *forward* equations $P'(t) = P(t)Q, \ t \geq 0, \ P(0) = I$, that one would expect to get by using $P(t+h) = P(t)P(h)$ can not be rigorously justified for all non-explosive CTMCs; there are examples (infinite state space) that cause trouble; the interchange of a limit and an infinite sum can not be justified.

But it does not matter, since the unique solution $P(t) = e^{Qt}$ to the backward equations is the unique solution to the forward equations, and thus both equations are valid.

> *For a (non-explosive) CTMC, the transition probabilities $P_{i,j}(t)$ are the unique solution to both the Kolmogorov backward and forward equations.*

Remark 1.1 It is rare that we can explicitly compute the infinite sum in the solution

$$P(t) = e^{Qt} = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!}.$$

But there are various numerical recipes for estimating e^{Qt} to any desired level of accuracy. For example, since
$e^M = \lim_{n\to\infty}(1 + M/n)^n$, for any square matrix M, one can choose n large and use $e^{Qt} \approx (1 + (Qt)/n)^n$.

1.12 Balance equations, rates, and positive recurrence

Consider any deterministic function $x(t),\ t \geq 0$ with values in $\mathcal{S}$. Clearly, every time $x(t)$ enters a state j, it must first leave that state in order to enter it again. Thus the number of times during the interval $(0, t]$ that it enters state j differs by at most one, from the number of times during the interval $(0, t]$ that it leaves state j. We conclude (by dividing by t and letting $t \to \infty$) that the long-run rate at which the function leaves state j equals the long-run rate at which the function enters state j. In words, "the rate out of state j is equal to the rate into state j, for each state j". We can apply this kind of result to each sample-path of a stochastic process. For a positive recurrent CTMC with limiting distribution $\vec{P} = \{P_j\}$, the rate out of state j is given by $a_j P_j$, while the rate into state j is given by $\sum_{i \neq j} P_i a_i P_{ij},\ j \in \mathcal{S}$, by interpreting the limiting probability P_j as a proportion of time and recalling Section 1.10 on transition rates. But we can show that fact directly; we do so next. And it leads to a set of equations that allow us to solve for the limiting distribution when it does exist.

A direct analysis of transition rates and how they lead to the balance equations

Letting $N_j^{(e)}(t)$ denote the number of times during $(0, t]$ that $\{X(t)\}$ entered state j, and $N_j^{(d)}(t)$ denote the number of times during $(0, t]$ that $\{X(t)\}$ departed state j, we have (recalling Equation 6) wp1 that **the long-run rate entering state j equals the long-run rate departing state j and is given by wp1**

$$\lim_{t\to\infty} \frac{N_j^{(e)}(t)}{t} = \lim_{t\to\infty} \frac{N_j^{(d)}(t)}{t} = \frac{1}{E(T_{jj})}.$$

The above is just the *elementary renewal theorem* applied to the renewal process of the times the chain visits state j, with iid interarrival times the lengths of time between visits to state j, having mean $E(T_{jj})$.

Moreover, from Proposition 1.1, we have

$$P_j = \frac{E(H_j)}{E(T_{jj})} = \frac{1/a_j}{E(T_{jj})},$$

which implies that

$$\frac{1}{E(T_{jj})} = a_j P_j,\ j \in \mathcal{S}.$$

We conclude that

Proposition 1.4 *The long-run rate entering state j equals the long-run rate departing state j and is given by the product*

$$a_j P_j,\ j \in \mathcal{S}.$$

Definition 1.3 *The* **balance equations** *for a positive recurrent CTMC are given by*

$$a_j P_j = \sum_{i \neq j} a_i P_i P_{ij}, \ j \in \mathcal{S}, \tag{16}$$

which in matrix form are given by

$$\vec{P}Q = \vec{0}, \tag{17}$$

where $Q = P'(0)$ is the transition rate matrix from Section 1.10.

From Proposition 1.4, the balance equations in words are given by **"the rate out of state j is equal to the rate into state j, for each state j".** The left hand side of (16) is the rate out of state j ($a_j P_j$). To get into state j, the chain has to come from the other states $i \neq j$, and so the right hand side sums up (over all other states $i \neq j$) the rate that the chain departs state i ($a_i P_i$) and moves next to state j ($P_{i,j}$); that yields on the right hand side the (total) rate that the chain enters state j.

Theorem 1.1 *An irreducible (and non-explosive) CTMC is positive recurrent if and only if there is a (necessarily unique) probability solution $\vec{P}$ to the balance equations $\vec{P}Q = \vec{0}$. The solution satisfies $P_j > 0, \ j \in \mathcal{S}$ and is the limiting (stationary) distribution as defined in Equations (1)-(4).*

Proof : Suppose that the chain is positive recurrent. Then from Proposition 1.1 and Proposition 1.2, there is a unique limiting probability distribution $\vec{P}$ and it is a stationary distribution; $\vec{P}P(t) = \vec{P}, \ t \geq 0$. Taking the derivative at $t = 0$ on both sides of $\vec{P}P(t) = \vec{P}$ yields

$$\vec{P}Q = \vec{0},$$

the balance equations.

Conversely, suppose that $\vec{P}$ is a probability solution to the balance equations. We will first show that any such solution must also satisfy $\vec{P}P(t) = \vec{P}, \ t \geq 0$, that is, it is a *stationary* distribution. We then will show that if an irreducible chain has such a stationary distribution, then the chain must be positive recurrent. To this end: Suppose that $\vec{P}Q = \vec{0}$. Multiplying both right sides by $P(t)$ yields $\vec{P}QP(t) = \vec{0}$, which due to the Kolmogorov Backward equations, $P'(t) = QP(t)$, is equivalent to $\vec{P}P'(t) = \vec{0}$ which is equivalent to

$$\frac{d(\vec{P}P(t))}{dt} = \vec{0}.$$

But this implies that $\vec{P}P(t)$ is a constant in t and hence that $\vec{P}P(t) = \vec{P}P(0) = \vec{P}I = \vec{P}$; $\vec{P}$ is indeed a stationary distribution. Now suppose that the chain is not positive recurrent. For an irreducible CTMC, all states together are transient, positive recurrent, or null recurrent, so the chain must be either null recurrent or transient and hence by Proposition 1.1, we have

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t P(s)ds = 0. \tag{18}$$

Multiplying both sides on the left by $\vec{P}$ yields

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \vec{P}P(s)ds = 0. \tag{19}$$

But using the already established $\vec{P}P(t) = \vec{P}$ we have $\frac{1}{t}\int_0^t \vec{P}P(s)ds = \frac{1}{t}\int_0^t \vec{P}ds = \vec{P}$ and we end with a contradiction $\vec{P} = 0$ ($\vec{P}$ is a probability distribution by assumption). Finally, from Proposition 1.2 we know that there can only be one stationary distribution for a positive recurrent chain, the limiting distribution as defined in Equations (1)-(4), so we conclude that $\vec{P}$ here is indeed the limiting distribution. ■

As for discrete-time Markov chains, when the state space is finite, we obtain a useful and simple special case:

Theorem 1.2 *An irreducible CTMC with a finite state space is positive recurrent; there is always a unique probability solution to the balance equations.*

Proof : Suppose (without loss of generality) the state space is $\mathcal{S} = \{1, 2, \dots, b\}$ for some integer $b \geq 1$. We already know that the chain must be recurrent because the embedded chain is so. We also know that the embedded chain is positive recurrent because for finite state discrete-time chains irreducibility implies positive recurrence. Let $\tau_{1,1}$ denote the discrete return time to state 1, and let $T_{1,1}$ denote the corresponding continuous return time. We know that $E(\tau_{1,1}) < \infty$. Also, $T_{1,1}$ is a random sum of $\tau_{1,1}$ holding times, starting with H_1. Let $a^* = \min\{a_1, \dots, a_b\}$. Then $a^* > 0$ and every holding time H_i satisfies $E(H_i) \leq 1/a^* < \infty$, $i \in \{1, 2, \dots, b\}$. Letting $\{Y_n\}$ denote iid exponential rvs at rate a^*, independent of $\tau_{1,1}$, we conclude (Wald's Equation) that

$$E(T_{1,1}) \leq E\Big(\sum_{n=1}^{\tau_{1,1}} Y_n\Big) = E(\tau_{1,1})(1/a^*) < \infty.$$

■

1.13 Examples of setting up and solving balance equations

Here we apply Theorems 1.1 and 1.2 to a variety of models. In most cases, solving the resulting balance equations involves recursively expressing all the P_j in terms of one particular one, P_0 (say), then solving for P_0 by using the fact that $\sum_{j\in\mathcal{S}} P_j = 1$. In the case when the state space is infinite, the sum is an infinite sum that might diverge unless further restrictions on the system parameters (rates) are enforced.

1. *FIFO M/M/1 queue:* $X(t)$ denotes the number of customers in the system at time t. Here, irreducibility is immediate since as pointed out earlier, the embedded chain

is a simple random walk (hence irreducible), so, from Theorem 1.1, we will have positive recurrence if and only if we can solve the balance equations (16):

$$\begin{aligned}
\lambda P_0 &= \mu P_1 \\
(\lambda + \mu)P_1 &= \lambda P_0 + \mu P_2 \\
(\lambda + \mu)P_2 &= \lambda P_1 + \mu P_3 \\
&\vdots \\
(\lambda + \mu)P_j &= \lambda P_{j-1} + \mu P_{j+1},\ j \geq 1.
\end{aligned}$$

These equations can also be derived from scratch as follows: Given $X(t) = 0$, the rate out of state 0 is the arrival rate $a_0 = \lambda$, and the only way to enter state 0 is from state $i = 1$, from which a departure must occur (rate μ). This yields the first equation. Given $X(t) = j \geq 1$, the rate out of state j is $a_j = \lambda + \mu$ (either an arrival or a departure can occur), but there are two ways to enter such a state j: either from state $i = j - 1$ (an arrival occurs (rate λ) when $X(t) = j - 1$ causing the transition $j - 1 \to j$), or from state $i = j + 1$ (a departure ocurrs (rate μ) when $X(t) = j$ causing the transition $j + 1 \to j$). This yields the other equations.

Note that since $\lambda P_0 = \mu P_1$ (first equation), the second equation reduces to $\lambda P_1 = \mu P_2$ which in turn causes the third equation to reduce to $\lambda P_2 = \mu P_3$, and in general the balance equations reduce to

$$\lambda P_j = \mu P_{j+1},\ j \geq 0, \tag{20}$$

which asserts that

for each j, the rate from j to $j + 1$ equals the rate from $j + 1$ to j,

or

$$P_{j+1} = \rho P_j,\ j \geq 0,$$

from which we recursivly obtain $P_1 = \rho P_0,\ P_2 = \rho P_1 = \rho^2 P_0$ and in general $P_j = \rho^j P_0$. Using the fact that the probabilities must sum to one yields

$$1 = P_0 \sum_{j=0}^{\infty} \rho^j,$$

from which we conclude that there is a solution if and only if the geometric series converges, that is, if and only if $\rho < 1$, equivalently $\lambda < \mu$, "the arrival rate is less than the service rate", in which case $1 = P_0(1 - \rho)^{-1}$, or $P_0 = 1 - \rho$.

Thus $P_j = \rho^j(1 - \rho),\ j \geq 0$ and we obtain a geometric stationary distribution.

Summarizing:

> *The FIFO M/M/1 queue is positive recurrent if and only if $\rho < 1$ in which case its stationary distribution is geometric with paramater ρ; $P_j = \rho^j(1-\rho),\ j \geq 0$. (If $\rho = 1$ it can be shown that the chain is null recurrent, and transient if $\rho > 1$.)*

When $\rho < 1$ we say that the M/M/1 queue is *stable*, *unstable* otherwise. Stability intuitively means that the queue length doesn't grow without bound over time.

When the queue is stable, we can take the mean of the stationary distribution to obtain the average number of customers in the system

$$
\begin{aligned}
l &= \lim_{t\to\infty} \frac{1}{t}\int_0^t X(s)ds && (21)\\
&= \sum_{j=0}^{\infty} jP_j && (22)\\
&= \sum_{j=0}^{\infty} j(1-\rho)\rho^j && (23)\\
&= \frac{\rho}{1-\rho}. && (24)
\end{aligned}
$$

2. *Birth and Death processes:* The fact that the balance equations for the FIFO M/M/1 queue reduced to "for each state j, the rate from j to $j+1$ equals the rate from $j+1$ to j" is not a coincidence, and in fact this reduction holds for any Birth and Death process. For in a Birth and Death process, the balance equations are:

$$
\begin{aligned}
\lambda_0 P_0 &= \mu_1 P_1\\
(\lambda_1+\mu_1)P_1 &= \lambda_0 P_0 + \mu_2 P_2\\
(\lambda_2+\mu_2)P_2 &= \lambda_1 P_1 + \mu_3 P_3\\
&\vdots\\
(\lambda_j+\mu_j)P_j &= \lambda_{j-1}P_{j-1} + \mu_{j+1}P_{j+1},\ j \geq 1.
\end{aligned}
$$

Plugging the first equation into the second yields $\lambda_1 P_1 = \mu_2 P_2$ which in turn can be plugged into the third yielding $\lambda_2 P_2 = \mu_3 P_3$ and so on. We conclude that for any Birth and Death process, the balance equations reduce to

$$\lambda_j P_j = \mu_{j+1}P_{j+1},\ j \geq 0,\ \textit{the Birth and Death balance equations.} \qquad (25)$$

Solving recursively, we see that

$$P_j = P_0 \frac{\lambda_0 \times \cdots \times \lambda_{j-1}}{\mu_1 \times \cdots \times \mu_j} = P_0 \prod_{i=1}^{j} \frac{\lambda_{i-1}}{\mu_i},\ j \geq 1.$$

Using the fact that the probabilities must sum to one then yields:

An irreducible Birth and Death process is positive recurrent if and only if

$$\sum_{j=1}^{\infty}\prod_{i=1}^{j}\frac{\lambda_{i-1}}{\mu_i}<\infty,$$

in which case

$$P_0=\frac{1}{1+\sum_{j=1}^{\infty}\prod_{i=1}^{j}\frac{\lambda_{i-1}}{\mu_i}},$$

and

$$P_j=\frac{\prod_{i=1}^{j}\frac{\lambda_{i-1}}{\mu_i}}{1+\sum_{j=1}^{\infty}\prod_{i=1}^{j}\frac{\lambda_{i-1}}{\mu_i}},\ j\geq 1. \tag{26}$$

For example, in the M/M/1 model,

$$1+\sum_{j=1}^{\infty}\prod_{i=1}^{j}\frac{\lambda_{i-1}}{\mu_i}=\sum_{j=0}^{\infty}\rho^j,$$

which agrees with our previous analysis.

We note in passing that the statement "for each state j, the rate from j to $j+1$ equals the rate from $j+1$ to j" holds for any deterministic function $x(t),\ t\geq 0$, in which changes of state are only of magnitude 1; up by 1 or down by 1. Arguing along the same lines as when we introduced the balance equations, every time this kind of function goes up from j to $j+1$, the only way it can do so again is by first going back down from $j+1$ to j. Thus the number of times during the interval $(0,t]$ that it makes an "up" transition from j to $j+1$ differs by at most one, from the number of times during the interval $(0,t]$ that it makes a "down" transition from $j+1$ to j. We conclude (by dividing by t and letting $t\to\infty$) that the long-run rate at which the function goes from j to $j+1$ equals the long-run rate at which the function goes from $j+1$ to j. Of course, as for the balance equations, being able to write this statement simply as $\lambda_j P_j=\mu_{j+1}P_{j+1}$ crucially depends on the Markov property

3. *M/M/1 loss system:* This is the M/M/1 queueing model, except there is no waiting room; any customer arriving when the server is busy is "lost", that is, departs without being served. In this case $\mathcal{S}=\{0,1\}$ and $X(t)=1$ if the server is busy and $X(t)=0$ if the server is free. $P_{01}=1=P_{10}$; the chain is irreducible. Since the state space is finite we conclude from Theorem 1.2 that the chain is positive recurrent for any $\lambda>0$ and $\mu>0$. We next solve for P_0 and P_1. We let $\rho=\lambda/\mu$. There is only one balance equation, $\lambda P_0=\mu P_1$. So $P_1=\rho P_0$ and since $P_0+P_1=1$, we conclude that $P_0=1/(1+\rho)$, $P_1=\rho/(1+\rho)$. So the long-run proportion of time that the server is busy is $\rho/(1+\rho)$ and the long-run proportion of time that the server is free (idle) is $1/(1+\rho)$.

4. *M/M/∞ queue:* $X(t)$ denotes the number of customers (busy servers) in the system at time t. Being a Birth and Death process we need only consider the Birth and Death balance equations (25) which take the form

$$\lambda P_j = (j+1)\mu P_{j+1},\ j \geq 0.$$

Irreducibility follows from the fact that the embedded chain is an irreducible simple random walk, so positive recurrence will follow if we can solve the above equations.

As is easily seen by recursion, $P_j = \rho^j/j!P_0$. Forcing these to sum to one (via using the Taylor's series expansion for e^x), we obtain $1 = e^{\rho}P_0$, or $P_0 = e^{-\rho}$. Thus $P_j = e^{-\rho}\rho^j/j!$ and we end up with the Poisson distribution with mean ρ:

> *The M/M/∞ queue is always positive recurrent for any $\lambda > 0$, $\mu > 0$; its stationary distribution is Poisson with mean $\rho = \lambda/\mu$.*

The above result should not be surprising, for we already studied (earlier in this course) the more general M/G/∞ queue, and obtained the same stationary distribution. But because we now assume exponential service times, we are able to obtain the result using CTMC methods. (For a general service time distribution we could not do so because then $X(t)$ does not form a CTMC; so we had to use other, more general, methods.)

5. *M/M/c loss queue:* This is the M/M/c model except there is no waiting room; any arrival finding all c servers busy is lost. This is the $c-$server analog of Example 3. With $X(t)$ denoting the number of busy servers at time t, we have, for any $\lambda > 0$ and $\mu > 0$, an irreducible B&D process with a finite state space $\mathcal{S} = \{0, \ldots, c\}$, so positive recurrence follows from Theorem 1.2. The B&D balance equations (25) are

$$\lambda P_j = (j+1)\mu P_{j+1},\ 0 \leq j \leq c-1,$$

or $P_{j+1} = P_j\rho/(j+1),\ 0 \leq j \leq c-1$; the first c equations for the FIFO M/M/∞ queue. Solving we get $P_j = \rho^j/j!P_0, 0 \leq j \leq c$, and summing to one yields

$$1 = P_0\Big(1 + \sum_{j=1}^{c}\frac{\rho^j}{j!}\Big) = P_0\Big(\sum_{j=0}^{c}\frac{\rho^j}{j!}\Big),$$

yielding

$$P_0 = \Big(\sum_{j=0}^{c}\frac{\rho^j}{j!}\Big)^{-1}.$$

Thus

$$P_j = \frac{\rho^j}{j!}\Big(\sum_{n=0}^{c}\frac{\rho^n}{n!}\Big)^{-1},\ 0 \leq j \leq c. \tag{27}$$

In particular

$$P_c = \frac{\rho^c}{c!}\Big(\sum_{n=0}^{c} \frac{\rho^n}{n!}\Big)^{-1}, \tag{28}$$

the proportion of time that all servers are busy. Later we will see from a result called *PASTA*, that P_c is also the proportion of lost customers, that is, the proportion of arrivals who find all c servers busy. This turns out to be a very famous/celebrated queueing theory result because the solution in (27), in particular the formula for P_c in (28), turns out to hold even if the service times are *not* exponential (the M/G/c-loss queue), a result called *Erlang's Loss Formula*.

6. *Population model with family immigration:* Here we start with a general B&D process (birth rates λ_i, death rates μ_i), but allow another source of population growth, in addition to the births. Suppose that at each of the times from a Poisson process at rate γ, independently, a family of random size B joins the population (immigrates). Let $b_i = P(B = i),\ i \geq 1$ denote corresponding family size probabilities. Letting $X(t)$ denote the population size at time t, we no longer have a B&D process now since the arrival of a family can cause a jump larger than size one. The balance equations ("the rate out of state j equals the rate into state j") are:

$$\begin{aligned}
(\lambda_0+\gamma)P_0 &= \mu_1 P_1 \\
(\lambda_1+\mu_1+\gamma)P_1 &= (\lambda_0+\gamma b_1)P_0 + \mu_2 P_2 \\
(\lambda_2+\mu_2+\gamma)P_2 &= \gamma b_2 P_0 + (\lambda_1 + \ \gamma b_1)P_1 + \mu_3 P_3 \\
&\vdots \\
(\lambda_j+\mu_j+\gamma)P_j &= \lambda_j P_{j-1} + \mu P_{j+1} + \sum_{i=0}^{j-1} \gamma b_{j-i} P_i,\ j \geq 1.
\end{aligned}$$

The derivation is as follows: When $X(t) = j$, any one of three events can happen next: A death (rate μ_j), a birth (rate λ_j) or a family immigration (rate γ). This yields the rate out of state j. There are j additional ways to enter state j, besides a birth from state $j-1$ or a death from state $j+1$, namely from each state $i < j$ a family of size $j-i$ could immigrate (rate γb_{j-i}). This yields the rate into state j.

1.14 Transitions back into the same state; $P_{i,i} > 0$.

In our study of CTMC's we have inherently been assuming that $P_{i,i} = 0$ for each $i \in \mathcal{S}$, but this is not necessary as we illustrate here.

Suppose that $0 < P_{i,i} < 0$. Assume $X_0 = i$ and let K denote the total number of transitions (visits) to state i before making a transition out to another state. Since $X_0 = i$, we count this initial visit as one such visit. Then $P(K = n) = (1-p)^{n-1}p,\ n \geq 1$, where $p = 1 - P_{i,i}$. Letting Y_n denote iid exponential rvs at rate a_i (the holding time rate), we can represent the total holding time H_T in state i as an independent geometric sum

$$H_T = \sum_{n=1}^{K} Y_n.$$

In particular $E(H_T) = E(K)/a_i = 1/pa_i$. In fact $H_T \sim exp(pa_i)$ as is easily seen by deriving its Laplace transform:

$$E(e^{sH_T}) = \frac{pa_j}{pa_i + s}, \ s \geq 0.$$

(Condition on K first.)

Thus, we can reset $a_i = pa_i$, reset $P_{i,i} = 0$ and reset $P_{i,j} = P_{i,j}/p$ for $j \neq i$. This yields the same CTMC $\{X(t)\}$ (e.g., it has the same distribution), but for which $P_{i,i} = 0$.

In any case, even if we keep $P_{i,i} > 0$, as long as one is consistent (on both sides of the balance equations), then the same balance equations arise in the end. We illustrate with a simple example: A CTMC with two states, $0, 1$, and embedded chain transition matrix

$$P = \begin{pmatrix} 0.25 & 0.75 \\ .20 & 0.80 \end{pmatrix}.$$

$a_0 > 0$ and $a_1 > 0$ are given non-zero holding-time rates. By definition, a_i is the holding time rate when in state i, meaning that after the holding time $H_i \sim exp(a_i)$ is completed, the chain will make a transition according to the transition matrix $P = (P_{ij})$. If we interpret a transition $j \to j$ as both a transition out of and into state j, then the balance equations are

$$\begin{aligned} a_0 P_0 &= (0.25) a_0 P_0 + (0.20) a_1 P_1 \\ a_1 P_1 &= (0.75) a_0 P_0 + (0.80) a_1 P_1. \end{aligned}$$

As the reader can check, these equations reduce to the one equation

$$(0.75) a_0 P_0 = (0.20) a_1 P_1,$$

which is what we get if we were to instead interpret a transition $j \to j$ as neither a transition into or out of state j. Resetting the parameters as explained above means resetting $a_0 = (0.75)a_0$, $a_1 = (0.20)a_1$ and P to

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

So, it makes no difference as far as $\{X(t)\}$ is concerned[3]. This is how it works out for any CTMC.

[3] But there might be other associated stochastic processes that will become different by making this change. For example, in queueing models, allowing $P_{i,i} > 0$ might refer to allowing customers to return to the end of the queue for another round after completing service. By resetting $P_{i,i} = 0$, we are forcing the customer to re-enter service immediately for the extra round instead of waiting at the end of the queue. This of course would effect quantities of interest such as *average waiting time*.

1.15 Poisson Arrivals See Time Averages (PASTA)

For a stable M/M/1 queue, let π^a_j denote the long-run proportion of arrivals who, upon arrival, find j customers already in the system. If $X(t)$ denotes the number in system at time t, and t_n denotes the time of the n^{th} Poisson arrival, then

$$\pi^a_j \overset{\text{def}}{=} \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I\{X(t_n-) = j\},$$

where $X(t_n-)$ denotes the number in system found by the n^{th} arrival.

On the one hand, $\lambda\pi^a_j$ is the long-run rate (number of times per unit time) that $X(t)$ makes a transition $j \to j+1$. After all, arrivals occur at rate λ, and such transitions can only happen when arrivals find j customers in the system. On the other hand, from the B&D balance equations (20), λP_j is also the same rate in question. Thus $\lambda\pi^a_j = \lambda P_j$, or

$$\pi^a_j = P_j,\ j \geq 0,$$

which asserts that

> *the proportion of Poisson arrivals who find j customers in the system is equal to the proportion of time there are j customers in the system.*

This is an example of *Poisson Arrivals See Time Averages (PASTA)*, and it turns out that PASTA holds for any queueing model in which arrivals are Poisson, no matter how complex, as long as a certain (easy to verify) condition, called *LAC*, holds. (Service times do not need to have an exponential distribution, they can be general, as in the $M/G/\infty$ queue.) Moreover, PASTA holds for more general quantities of interest besides number in system. For example, the proportion of Poisson arrivals to a queue who, upon arrival, find a particular server busy serving a customer with a remaining service time exceeding x (time units) is equal to the proportion of time that this server is busy serving a customer with a remaining service time exceeding x. In general, PASTA will not hold if the arrival process is not Poisson.

To state PASTA more precisely, let $\{X(t) : t \geq 0\}$ be any stochastic process, and $\psi = \{t_n : n \geq 0\}$ a Poisson process. Both processes are assumed on the same probability space. We have in mind that $X(t)$ denote the state of some "queueing" process with which the Poisson arriving "customers" are interacting/participating. The state space $\mathcal{S}$ can be general such as multi-dimensional Euclidean space. We assume that the sample-paths of $X(t)$ are right-continuous with left-hand limits. [4]

The *lack of anticipation condition* (*LAC*) that we will need to place on the Poisson process asserts that for each fixed $t > 0$, the future increments of the Poisson process

[4] A function $x(t),\ t \geq 0$, is right-continuous if for each $t \geq 0$, $x(t+) \overset{\text{def}}{=} \lim_{h\downarrow 0} X(t+h) = x(t)$. It has left-hand limits if for each $t > 0$, $x(t-) \overset{\text{def}}{=} \lim_{h\downarrow 0} x(t-h)$ exists (but need not equal $x(t)$). If $x(t-) \neq x(t+)$, then the function is said to be discontinuous at t, or have a *jump* at t. Queueing processes typically have jumps at arrval times and departure times.

after time t, $\{N(t+s)-N(t) : s \geq 0\}$, be independent of the joint past, $\{(N(u), X(u)) : 0 \leq u \leq t\}$. This condition is stronger than the independent increments property of the Poisson process, for it requires that any future increment be independent not only of its own past but of the past of the queueing process as well. If the Poisson process is completely independent of the queueing process, then LAC holds, but we have in mind the case when the two processes are dependent via the arrivals being part of and participating in the queueing system.

Let $f(x)$ be any bounded real-valued function on $\mathcal{S}$, and consider the real-valued process $f(X(t))$. We are now ready to state PASTA. (The proof, ommitted, is beyond the scope of this course.)

Theorem 1.3 (PASTA) *If the Poisson process satisfies LAC, then w.p.1.,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(X(t_n-)) = \lim_{t\to\infty} \frac{1}{t} \int_0^t f(X(s))ds,$$

in the sense that if either limit exists, then so does the other and they are equal.

A standard example when $X(t)$ is the number of customers in a queue, would be to let f denote an indicator function; $f(x) = I\{x = j\}$, so that $f(X(t)) = I\{X(t) = j\}$, and $f(X(t_n-)) = I\{X(t_n-) = j\}$. This would, for example, yield $\pi_j^a = P_j$ for the M/M/1 queue.

The reader should now go back to Example 5 in Section 1.13, the M/M/c-loss queue, where we first mentioned PASTA in the context of Erlang's Loss Formula.

1.16 Multi-dimensional CTMC's

So far we have assumed that a CTMC is a one-dimensional process, but that is not necessary. All of the CTMC theory we have developed in one-dimension applies here as well (except for the Birth and Death theory). We illustrate with some two-dimensional examples, higher dimensions being analogous.

1. *Tandem queue:* Consider a queueing model with two servers in tandem: Each customer, after waiting in line and completing service at the first single-server facility, immediately waits in line at a second single-server facility. Upon completion of the second service, the customer finally departs. in what follows we assume that the first facility is a FIFO M/M/1, and the second server has exponential service times and also serves under FIFO, in which case this system is denoted by

   $$FIFO\ M/M/1/ \longrightarrow /M/1.$$

   Besides the Poisson arrival rate λ, we now have two service times rates (one for each server), μ_1 and μ_2. Service times at each server are assumed i.i.d. and independent of each other and of the arrival process.

Letting $X(t) = (X_1(t), X_2(t))$, where $X_i(t)$ denotes the number of customers in the i^{th} facility, $i = 1, 2$, it is easily seen that $\{X(t)\}$ satisfies the Markov property. This is an example of an irreducible two-dimensional CTMC. Balance equations (rate out of a state equals rate into the state) can be set up and used to solve for stationary probabilities. Letting $P_{n,m}$ denote the long-run proportion of time there are n customers at the first facility and m at the second (a joint probability),

$$\lambda P_{0,0} = \mu_2 P_{0,1},$$

because the only way the chain can make a transion into state $(0, 0)$ is from $(0, 1)$ (no one is at the first facility, exactly one customer is at the second facility, and this one customer departs (rate μ_2)). Similarly when $n \geq 1,\ m \geq 1$,

$$(\lambda + \mu_1 + \mu_2) P_{n,m} = \lambda P_{n-1,m} + \mu_1 P_{n+1,m-1} + \mu_2 P_{n,m+1},$$

because either a customer arrives, a customer completes service at the first facility and thus goes to the second, or a customer completes service at the second facility and leaves the system. The remaining balance equations are also easily derived. Letting $\rho_i = \lambda/\mu_i,\ i = 1, 2$, it turns out that the solution is

$$P_{n,m} = (1 - \rho_1)\rho_1^n \times (1 - \rho_2)\rho_2^m,\ n \geq 0,\ m \geq 0,$$

provided that $\rho_i < 1,\ i = 1, 2$. This means that as $t \to \infty$, $X_1(t)$ and $X_2(t)$ become independent r.v.s. each with a geometric distribution. This result is quite surprising because, after all, the two facilities are certainly dependent at any time t, and why should the second facility have a stationary distribution as if it were itself an M/M/1 queue? (For example, why should departures from the first facility be treated as a Poisson process at rate λ?) The proof is merely a "plug in and check" proof using Theorem 1.2: Plug in the given solution (e.g., treat it as a "guess") into the balance equations and verify that they work. Since they do work, they are the unique probability solution, and the chain is positive recurrent.

It turns out that there is a nice way of understanding part of this result. The first facilty is an M/M/1 queue so we know that $X_1(t)$ by itself is a CTMC with stationary distribution $P_n = (1 - \rho_1)\rho_1^n,\ n \geq 0$. If we start off $X_1(0)$ with this stationary distribution ($P(X_1(0) = n) = P_n,\ n \geq 0$), then we know that $X_1(t)$ will have this same distribution for all $t \geq 0$, that is, $\{X_1(t)\}$ is stationary. It turns out that when stationary, the departure process is itself a Poisson process at rate λ, and so the second facility (in isolation) can be treated itself as an M/M/ 1 queue when $\{X_1(t)\}$ is stationary. This at least explains why $X_2(t)$ has the geometric stationary distribution, $(1 - \rho_2)\rho_2^m,\ m \geq 0$, but more analysis is required to prove the independence part.

2. *Jackson network:*

Consider two FIFO single-server facilities (indexed by 1 and 2), each with exponential service at rates μ_1 and μ_2 respectively. For simplicity we refer to each facility as a "node". Each node has its own queue with its own independent Poisson arrival process at rates λ_1 and λ_2 respectively. Whenever a customer completes service at node $i = 1, 2$, they next go to the queue at node $j = 1, 2$ with probability Q_{ij}, independent of the past, or depart the system with probability $Q_{i,0}$, where the state 0 refers to departing the system, and we require that $Q_{0,0} = 1$, an absorbing state. We always assume that states 1 and 2 are transient, and state 0 is absorbing. So typically, a customer gets served a couple of times, back and forth between the two nodes before finally departing. In general, we allow *feedback*, which means that a customer can return to a given node (perhaps many times) before departing the system. The tandem queue does not have feedback; it is the special case when $Q_{1,2} = 1$ and $Q_{2,0} = 1$ and $\lambda_2 = 0$, an example of a *feedforward* network. In general, $Q = (Q_{ij})$ is called the routing transition matrix, because it represents the transition matrix of a Markov chain. Letting $X(t) = (X_1(t), X_2(t))$, where $X_i(t)$ denotes the number of customers in the i^{th} node, $i = 1, 2$, $\{X(t)\}$ yields an irreducible CTMC. Like the tandem queue, it turns out that the stationary distribution for the Jackson network is of the product form

$$P_{n,m} = (1 - \rho_1)\rho_1^n \times (1 - \rho_2)\rho_2^m, \ n \geq 0, \ m \geq 0,$$

provided that $\rho_i < 1, \ i = 1, 2$. Here

$$\rho_i = \frac{\lambda_i}{\mu_i} E(N_i),$$

where $E(N_i)$ is the expected number of times that a customer attends the i^{th} facility. $E(N_i)$ is completely determined by the routing matrix Q: Each customer, independently, is routed according to the discrete-time Markov chain with transition matrix Q, and since 0 is absorbing (and states 1 and 2 transient), the chain will visit each state $i = 1, 2$ only a finite number of times before getting absorbed. Notice that $\alpha_i = \lambda_i E(N_i)$ represents the *total arrival rate* to the i^{th} node. So $\rho_i < 1, i = 1, 2$, just means that the total arrival rate must be smaller than the service rate at each node. As with the tandem queue, the proof can be carried out by the "plug in and check" method. The α_i can be computed equivalently as the solution to the *flow* equations:

$$\alpha_i = \lambda_i + \sum_j \alpha_j Q_{j,i}, \ i, j \in \{1, 2\}.$$

Letting $Q_T = (Q_{j,i}), \ i, j \in \{1, 2\}$, denote the 2×2 matrix without the absorbing state 0 included, the flow equations in matrix form are

$$\vec{\alpha} = \vec{\lambda} + \vec{\alpha} Q_T,$$

with solution

$$\vec{\alpha} = \vec{\lambda}(I - Q_T)^{-1}.$$

We recognize that $(I - Q_T)^{-1} = S = (s_{i,j})$, where $s_{i,j}$ denotes the expected number of times the discrete-time chain visits transient state j given it started in transient state $i,\ i = 1, 2$.