1 IEOR 4701: 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 open maze. 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 $\mathcal{S} = \{\cdots, -2, -1, 0, 1, 2, \cdots\}$ (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 $\{X(t) : t \geq 0\}$ 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 an example of a transition probability for the CTMC and represents the probability that the chain will be in state j, t time units from now, given it is in state i now. 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 the present state $X(s) = i$, whatever it is, and the amount of time that has elapsed, t, since time s. In particular, $P_{ij}(t) = P(X(t) = j | X(0) = i)$. (This is the continuous-time analog of time-stationary transition probabilities, $P(X_{n+k} = j | X_n = i) = P(X_k = j | X_0 = i),\ n \geq 0,\ k \geq 0$, for discrete-time Markov chains.)

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.

[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 1.9 for details.

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. a_i can be interpreted as the rate out of state i given that $X(t) = i$; the intuitive idea being that the holding time will end, independent of the past, in the next dt units of time w.p. $a_i dt$.

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.

1.1 Chapman-Kolmogorov equations

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

Lemma 1.1 (Chapman-Kolmogorov for CTMC's) *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 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 compute the probability that the chain, now in state k, would then be in state j after an additional t time units, $P_{kj}(t)$.

1.2 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 $t \to \infty$, $N(t) \to \infty$ step by step. The graph of the sample paths of $\{N(t)\}$ are said to be step functions since they look like the steps of a stairway. This is an example of a *Pure Birth* process, since we can view each arrival as a new birth in a population in which no one ever dies.

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;

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

   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. (p_1, p_2, p_3, p_4) is called the stationary distribution for the CTMC.

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 (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 someone just departed leaving the system empty (for it is not possible for the system to be empty right after an arrival). Thus $P(X_{n+1} = 1 | X_n = 0) = 1$ since only an arrival can occur next if the system is empty. But whenever $X_n = i \geq 1$, $X_{n+1} = i+1$ 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 ($P_{0,1} = 1$).*

This CTMC is an example of a *Birth and Death* process, since we can view each arrival as a birth, and each departure as a death in a population. The *birth rate* when in state i, denoted by λ_i, is λ for all $i \geq 0$, since the time until the next birth (arrival) is always exponentially distributed with rate λ. The *death rate* when in state $i \geq 1$, denoted by μ_i, is μ since the time until the next death (departure) is always exponentially distributed with rate μ. ($\mu_0 = 0$ since there can be no death when the system is empty.) Whenever $X(t) = i$, the rate out of state i is the holding time rate, the sum of the birth and death rates, $a_i = \lambda_i + \mu_i$.

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 line (queue) and enter service at the first available free server. Once again we let $X(t)$ denote 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: $a_i = \lambda + i\mu,\ 0 \leq i \leq c,\ a_i = \lambda + c\mu,\ i > c$. This CTMC is another example of a Birth and Death process; $\lambda_i = \lambda,\ i \geq 0$ and $\mu_i = i\mu,\ 0 \leq i \leq c,\ \mu_i = c\mu,\ i > c$. Whereas the birth rate remains constant, the death rate depends on how many busy servers there are, but is never larger than $c\mu$ because that is the maximum number of busy servers possible at any given time.

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. This CTMC is yet another example of a Birth and Death process; $\lambda_i = \lambda,\ i \geq 0$ and $\mu_i = i\mu,\ i \geq 0$. Thus the death rate is proportional to the population size, whereas the birth rate remains constant.

   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 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. Note how, as i increases, the down probability increases, and approaches 1 as $i \to \infty$: when the system is heavily congested, departures occur rapidly.

1.3 Birth and Death processes

Most of our examples thus far of a CTMC were Birth and Death (B&D) processes, CTMC's that can only change state by increasing by one, or decreasing by one; $P_{i,i+1} + P_{i,i-1} = 1,\ i \in \mathcal{S}$. (The only example that was not B&D was the rat in the maze, because for example, the state can change from $i = 1$ to $j = 3$, a change of size two.) 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 $X(t)$ is called 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. $P_{i,i-1} = P(Y < X) = \mu_i/(\lambda_i + \mu_i)$.

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 embedded chain for a B&D process is, in general, a simple random walk with state dependent increment probabilities.

The reader is encouraged at this point to go back over the B&D Examples in the previous Section.

1.4 Explosive CTMC's

Consider a pure Birth process $\{X(t)\}$ 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

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

(the expected sum of all the holding times), 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.

1.5 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 easily seen that i and j communicate 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 re-visits state i, 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.*

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

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 state i to be positive recurrent for the CTMC and null recurrent for the embedded chain (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. As for discrete-time chains (where we can use "$\pi = \pi P$" to determine positive recurrence), it turns out that determining positive recurrence for an irreducible CTMC is equivalent to finding a probability solution $\{P_j : j \in \mathcal{S}\}$ to a set of linear equations called *balance equations*. These probabilities are the stationary (limiting) probabilities for the CTMC; and can be interpreted (regardless of initial condition $X(0) = i$) as the long-run proportion of time the chain spends in state j:

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

which after taking expected values yields

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

Moreover, without any further conditions (such as aperiodicity), we have

$$P_j = \lim_{t\to\infty} P_{ij}(t). \tag{3}$$

As in discrete-time,

$$\text{if } P(X(0) = j) = P_j,\ j \in \mathcal{S},\ \text{then } P(X(t) = j) = P_j,\ j \in \mathcal{S}. \tag{4}$$

That is why (P_j) is called the stationary distribution:

> *If a positive recurrent CTMC starts off with its stationary distribution, then it retains that distribution forever after.*

We will study all of this in more detail in the next sections.

1.6 Positive recurrence imples existence of P_j

As for discrete-time Markov chains, positive recurrence implies the existence of stationary 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 stationary probabilities P_j as defined by Equation (1) exist and are 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)".

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 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, and we let $N_j(t)$ denote the number of such points during $(0, t]$. From the Elementary Renewal Theorem,

$$\lim_{t\to\infty} \frac{N_j(t)}{t} = \frac{1}{E(T_{jj})}. \tag{5}$$

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 (5) and the numerator is from the SLLN applied to $\{J_n\}$. $P_j > 0$ since $E(T_{jj}) < \infty$ (positive recurrence assumption). ■

That (3) also holds, is a consequence of the fact that the cycle length distribution is continuous; the distribution of T_{jj} has a density. In general, a positive recurrent regenerative process with a continuous cycle-length distribution converges in the sense of (3). The details of this are beyond the scope of the present course.

1.7 Balance equations 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, and as we will see next, doing so for a CTMC leads to a simple way of finding the stationary probabilities, and determining positive recurrence.

Theorem 1.1 *An irreducible (and non-explosive) CTMC is positive recurrent if and only if there is a (necessarily unique) probability solution $\{P_j : j \in \mathcal{S}\}$ (e.g., $P_j \geq 0,\ j \in \mathcal{S}$ and $\sum_{j\in\mathcal{S}} P_j = 1$) to the set of "balance equations":*

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

In words, "the rate out of state j is equal to the rate into state j, for each state j"

In this case, $P_j > 0,\ j \in \mathcal{S}$ and are the stationary (limiting) probabilities for the CTMC; Equations (1)-(4) hold.

We will not prove all of Theorem 1.1 here, but will be satisfied with proving one direction: We already know from Section 1.6 that positive recurrence implies the existence of the stationary probabilities P_j. In Section 1.13 a proof is provided (using differential equations) as to why these P_j must satisfy the balance equations.

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.*

That $a_j P_j$ represents the long-run rate out of state j (that is, the long-run number of times per unit time that the chain makes a transition out of state j) is argued as follows:

P_j is the proportion of time spent in state j, and whenever $X(t) = j$, independent of the past, the remaining holding time in state j has an exponential distribution with rate a_j.

Similarly, that $\sum_{i\neq j} P_i a_i P_{ij}$ represents the long-run rate into state j is argued as follows:

The chain will enter state j from states $i \neq j$. P_i is the proportion of time spent in state i. Whenever $X(t) = i$, independent of the past, the remaining holding time in state i has an

exponential distribution with rate a_i, and then the chain will make a transition into state j with probability P_{ij}. Summing up over all state $i \neq j$ yields the result.

The idea here is that whenever $X(t) = j$, independent of the past, the probability that the chain will make a transition (hence leave state j) within the next (infinitesimal) interval of length dt is given by $a_j dt$ (it is here we are using the Markov property). Thus a_j is interpreted as the instantaneous rate out of state j given that $X(t) = j$. Similarly whenever $X(t) = i$, independent of the past, the probability that the chain will make a transition from i to j within the next (infinitesimal) interval of length dt is given by $a_i P_{ij} dt$ (it is here we are using the Markov property), so $a_i P_{ij}$ is interpreted as the instantaneous rate into state j given that $X(t) = i$. (All of this is analogous to the Bernoulli trials interpretation of a Poisson process at rate λ; λdt is the probability of an arrival ocurring in the next dt time units, independent of the past.)

It is important to understand that whereas "the rate out of state j equals the rate into state j" statement would hold even for non-Markov chains (it is a deterministic sample-path statement, essentially saying "what goes up must come down"), the fact that these rates can be expressed in a simple form merely in terms of $a_j P_j$ and $P_i a_i P_{ij}$ depends crucially on the Markov property. Thus, in the end, it is the Markov propery that allows us to solve for the stationary probabilities in such a nice algebraic fashion, a major reason why CTMC's are so useful in applications. See Section 1.13 for the proof of why positive recurrence implies that the P_j must satisfy the balance equations.

Final Comment

We point out that $\pi = \pi P$ for a discrete-time MC also has the same "rate out = rate in" interpretation, and therefore can be viewed as discrete-time balance equations. To see why, first observe that since time is discrete, the chain spends exactly one unit of time in a state before making a next transition. Thus π_j can be interpreted not only as the long-run proportion of time the chain spends in state j but also as the long-run rate (long-run number of times per unit time) at which the chain enters state j and as the long-run rate at which the chain leaves state j. (Every time the chain enters state j, in order to do so again it must first leave state j.) $\pi_i P_{ij}$ thus can be interpreted as the rate at which the chain makes a transition $i \to j$, and so $\sum_{i \in \mathcal{S}} \pi_i P_{ij}$ can be interpreted as the rate at which the chain makes a transition into state j. So

$$\pi_j = \sum_{i \in \mathcal{S}} \pi_i P_{ij},$$

says that "the rate out of state j equals the rate into state j".

1.8 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. *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, at time t. Since $P_{01} = 1 = P_{10}$, (after all, the chain alternates forever between being empty and having one customer) 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, $\lambda P_0 = \mu P_1$ (since the other one is identical to this one): Whenever $X(t) = 0$, λ is the rate out of state 0, and whenever $X(t) = 1$, μ is the rate into state 0 (equivalently, out of state 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 $1/(1+\rho)$.

2. *A three state chain:* Consider a CTMC with three states $0, 1, 2$ in which whenever it is in a given state, it is equally likely next to move to any one of the remaining two states. Assume that a_0, a_1, a_2 are given non-zero holding-time rates.

   This chain is clearly irreducible and has a finite state space, so it is positive recurrent by Theorem 1.2. The balance equations are

$$\begin{aligned} a_0 P_0 &= (1/2) a_1 P_1 + (1/2) a_2 P_2 \\ a_1 P_1 &= (1/2) a_0 P_0 + (1/2) a_2 P_2 \\ a_2 P_2 &= (1/2) a_0 P_0 + (1/2) a_1 P_1. \end{aligned}$$

   Using $P_0 + P_1 + P_2 = 1$, one can solve (details left to the reader).

3. *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:

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

   These equations are derived 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

$$\lambda P_j = \mu P_{j+1},\ j \geq 0, \tag{6}$$

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 queueing model is *stable*, *unstable* otherwise. Stability intuitively means that the queue length doesn't grow and get out of control over time, but instead reaches an equilibrium in distribution.

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

$$\begin{aligned} L &= \sum_{j=0}^{\infty} j p_j & (7)\\ &= \sum_{j=0}^{\infty} j(1-\rho)\rho^j & (8)\\ &= \frac{\rho}{1-\rho}. & (9) \end{aligned}$$

Then, from $L = \lambda w$, we can solve for the average sojourn time of a customer, $w = L/\lambda$ or

$$w = \frac{1/\mu}{1-\rho}. \tag{10}$$

Using (10) we can go on to compute average delay d (time spent in line before entering service), because $w = d + 1/\mu$, or $d = w - 1/\mu$;

$$d = \frac{\rho}{\mu(1-\rho)}. \tag{11}$$

4. *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

$$\lambda_j P_j = \mu_{j+1} P_{j+1},\ j \geq 0,\ \textit{the Birth and Death balance equations.} \tag{12}$$

Solving recursively, we see that

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

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

$$\sum_{j=1}^{\infty} \frac{\lambda_0 \times \cdots \times \lambda_{j-1}}{\mu_1 \times \cdots \times \mu_j} < \infty,$$

in which case

$$P_0 = \frac{1}{1 + \sum_{j=1}^{\infty} \frac{\lambda_0 \times \cdots \times \lambda_{j-1}}{\mu_1 \times \cdots \times \mu_j}},$$

and

$$P_j = \frac{\frac{\lambda_0 \times \cdots \times \lambda_{j-1}}{\mu_1 \times \cdots \times \mu_j}}{1 + \sum_{j=1}^{\infty} \frac{\lambda_0 \times \cdots \times \lambda_{j-1}}{\mu_1 \times \cdots \times \mu_j}},\ j \geq 1. \tag{13}$$

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

$$1 + \sum_{j=1}^{\infty} \frac{\lambda_0 \times \cdots \times \lambda_{j-1}}{\mu_1 \times \cdots \times \mu_j} = \sum_{j=0}^{\infty} \rho^j,$$

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 as $\lambda_j P_j = \mu_{j+1} P_{j+1}$ crucially depends on the Markov property

5. *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 (12) 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.)

6. *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 1. 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 (12) 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

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

yielding

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

Thus

$$P_j = \frac{\rho^j}{j!}\Big(1 + \sum_{n=1}^{c} \frac{\rho^n}{n!}\Big)^{-1},\ 0 \le j \le c. \tag{14}$$

In particular

$$P_c = \frac{\rho^c}{c!}\Big(1 + \sum_{n=1}^{c} \frac{\rho^n}{n!}\Big)^{-1}, \tag{15}$$

the proprtion 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 important result because the solution in (14), in particular the formula for P_c in (15), holds even if the service times are not exponential (the M/G/c-loss queue), a famous queueing result called *Erlang's Loss Formula*.

7. *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 $b_i = P(B = i),\ i \ge 1$ 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:

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

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.

1.9 Transitions back into the same state; $P_{jj} > 0$.

In our study of CTMC's we have inherently been assuming that $P_{jj} = 0$ for each $j \in \mathcal{S}$, but this is not necessary. The minor complication stems from making sense of wether or not a transition

from j to j is to be interpreted as having left/entered state j, "did the chain really leave (or enter) state j when it made such a transition?" This would seem to be an important issue when setting up balance equations. But it turns out that as long as one is consistent (on both sides of the equations), then the same equations arise in the end. We illustrate with a simple example: A CTMC with two states, $0, 1$, in which $P_{ij} = 1/2,\ i, j = 0, 1$. a_0 and a_1 are given non-zero holding-time rates. It is important to observe that, 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 into and out of state j, then the balance equations are

$$\begin{aligned} a_0 P_0 &= (1/2)a_0 P_0 + (1/2)a_1 P_1 \\ a_1 P_1 &= (1/2)a_0 P_0 + (1/2)a_1 P_1. \end{aligned}$$

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

$$a_0 P_0 = a_1 P_1.$$

If we instead interpret a transition $j \to j$ as neither a transition into or out of state j, then the balance equation is (they both are identical)

$$(1/2)a_0 P_0 = (1/2)a_1 P_1,$$

which simplifies to the same one equation

$$a_0 P_0 = a_1 P_1.$$

So, it makes no difference. This is how it works out for any CTMC.

1.10 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. There is one exogenous Poisson arrival process at rate λ which gets partitioned with probabilities p_1 and $p_2 = 1 - p_1$ (yielding rates

$\lambda_1 = p_1\lambda$ and $\lambda_2 = p_2\lambda$ to the nodes respectively). Type 1 arrivals join the queue at node 1 and type 2 do so at node 2. This is equivalent to each node having its own independent Poisson arrival process. Whenever a customer completes service at node i, they next go to the queue at node j with probability Q_{ij}, independent of the past, $i = 1, 2,\ j = 0, 1, 2.$ $j = 0$ refers to departing the system. So typically, a customers 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 $p_2 = 0$, an example of a *feedforward* network. In general, $Q = (Q_{ij})$ is called the routing transition matrix, and represents the transition matrix of a Markov chain. We always assume that states 1 and 2 are transient, and state 0 is absorbing. 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

$$P_{n,m} = (1 - \alpha_1)\alpha_1^n \times (1 - \alpha_2)\alpha_2^m,\ n \geq 0,\ m \geq 0,$$

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

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

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 $\lambda_i E(N_i)$ represents the *total arrival rate* to the i^{th} node. So $\alpha_i < 1, i = 1, 2$, 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.

1.11 Poisson Arrivals See Time Averages (PASTA)

For a stable M/M/1 queue, let π_j^a 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

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

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

On the one hand, $\lambda\pi_j^a$ 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 (6), λP_j is also the same rate in question. Thus $\lambda\pi_j^a = \lambda P_j$, or

$$\pi_j^a = P_j,\ j \geq 0,$$

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. For example PASTA holds for the M/G/c queue, the M/G/c loss queue, and essentially any queueing network in which arrivals are Poisson.

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. [2]

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 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 6 in Section 1.8, the M/M/c-loss queue, where we first mentioned PASTA in the context of Erlang's Loss Formula.

[2] A function $x(t)$, $t \geq 0$, is right-continuous if for each $t \geq 0$, $x(t+) \stackrel{\text{def}}{=} \lim_{h\downarrow 0} X(t+h) = x(t)$. It has left-hand limits if for each $t > 0$, $x(t-) \stackrel{\text{def}}{=} \lim_{h\downarrow 0} x(t-h)$ 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.

Final Remarks on PASTA

1. To see why Poisson arrivals are needed to assert that arrivals see time averages, consider a single-server queue in which all interarrival times are exactly of length 2, and all service times are exactly of length 1, a deterministic queue. Note that every customer completes service before the next arrival, and so every arrival finds the system empty. This means that $\pi_0^a = 1$, the proportion of arrivals who find an empty system is 1. But since $\rho = 0.5$ ($\lambda = 0.5$ and $\mu = 1$), the proportion of time the system is empty is $P_0 = 1 - \rho = 0.5$. So $\pi_0^a \neq P_0$ here. This is the norm rather than the exception, unless arrivals are Poisson.

2. The simple proof of PASTA that we presented for the M/M/1 queue ($\lambda\pi_j^a = \lambda P_j$) works essentially for any queueing model (multidimensional complex networks even) for which the only source of customer arrivals is a Poisson process at rate λ. Letting $X(t)$ denote the number of customers in the system at time t (in total, no matter where/how they are distributed within the system), all that is needed is: (1) Only a Poisson arrival can cause $X(t)$ to jump from j to $j+1$. (2) For any $n \geq 0$, given $X(t) = n$, λ is the (instantaneous) rate at which the next arrival will occur, independent of the past. (1) implies that $\lambda\pi_j^a$ is the rate at which $\{X(t)\}$ makes a transition $j \to j+1$, and (2) implies that λP_j is this very same rate.

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

We have yet to show how to compute the transition probabilities $P_{ij}(t) = P(X(t) = j|X(0) = i),\ t \geq 0$ for a CTMC. 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.

For a function $x(t),\ t \geq 0$, the derivative

$$x'(t) \stackrel{\text{def}}{=} \lim_{h \to 0} \frac{x(t+h) - x(t)}{h},$$

can be described by

$$x'(t) = \frac{x(t+dt) - x(t)}{dt}, \tag{16}$$

where dt denotes an interval length h that is infinitesimally small.

Moreover, for a (non-explosive) CTMC,

$$P_{ik}(dt) = a_i P_{ik} dt,\ k \neq i \in \mathcal{S}, \tag{17}$$

$$P_{ii}(dt) = 1 - a_i dt, \tag{18}$$

which says that given $X(0) = i$, the probability of a trandsition $i \to k \neq i$ in the next dt time units is the rate of such a transition, conditional on $X(0) = i$, multiplied by dt (remember the Poisson process and more generally the balance equations), and the probability that no transition takes place is 1 minus the probability that a transition does take place.

The Chapman-Kolmogorov equations, for all $t \geq 0,\ s \geq 0,\ i, j \in \mathcal{S}$,

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

with $s = dt$, together with (17)-(18) yield

$$\begin{aligned} P_{ij}(t+dt) - P_{ij}(t) &= -P_{ij}(t) + \sum_{k \in \mathcal{S}} P_{ik}(dt) P_{kj}(t) && (19) \\ &= -P_{ij}(t)(1 - P_{ii}(dt)) + dt \sum_{k \neq i} a_i P_{ik} P_{kj}(t) && (20) \\ &= dt\Big[-P_{ij}(t) a_i + \sum_{k \neq i} a_i P_{ik} P_{kj}(t)\Big]. && (21) \end{aligned}$$

From (16) we thus conclude that

$$P'_{ij}(t) = -a_i P_{ij}(t) + \sum_{k \neq i} a_i P_{ik} P_{kj}(t),\ i, j \in \mathcal{S}, \tag{22}$$

known as the *Kolmogorov backward equations*. These are a set of linear differential equations and thus can be solved accordingly (as we shall do below). The word *backward* refers to the fact that in our use of the Chapman-Kolmogorov equations, we chose to place the $s = dt$ first and the t second, that is, the dt was placed "in back". The derivation above can be rigorously justified for any non-explosive CTMC.

If we define $r_{ik} = a_i P_{ik},\ k \neq i,\ r_{ii} = -a_i$, then in matrix form (22) becomes

$$P'(t) = RP(t),\ P(0) = I,\ \textit{backward equations in matrix form}, \tag{23}$$

where $P(t) = (P_{ij}(t)),\ P'(t) = (P'_{ij}(t)),\ R = (r_{ij})$, and I is the identity matrix (recall that $P_{ii}(0) = 1,\ P_{ij}(t) = 0,\ j \neq i$).

The unique solution is thus of the exponential form;

$$P(t) = e^{Rt},\ t \geq 0, \tag{24}$$

where

$$e^{Rt} \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} \frac{(Rt)^n}{n!}.$$

It is rare that we can explicitly compute the inifinite sum, but there are various numerical recipes for estimating e^{Rt} to any desired level of accuracy. For example, since $e^M = \lim_{n \to \infty}(1 + M/n)^n$, for any matrix M, one can choose n large and use $e^{Rt} \approx (1 + (Rt)/n)^n$.

1.13 Kolmogorov forward equations

Note that $P(t) = e^{Rt}$ is also the unique solution to $P'(t) = P(t)R,\ P(0) = I$, known as the *Kolmogorov forward equations* because they can be derived by putting the $s = dt$ in front

instead of in back when using the Chapman Kolmogorov equations (together with (16)-(18)):

$$\begin{aligned} P_{ij}(t+dt) - P_{ij}(t) &= -P_{ij}(t) + \sum_{k\in\mathcal{S}} P_{ik}(t)P_{kj}(dt) && (25)\\ &= -P_{ij}(t) + P_{ij}(t)P_{jj}(dt) + \sum_{k\neq j} P_{ik}(t)P_{kj}(dt) && (26)\\ &= dt\Big[-P_{ij}(t)a_j + \sum_{k\neq j} P_{ik}(t)a_k P_{kj}\Big], && (27) \end{aligned}$$

yielding

$$P'_{ij}(t) = -a_j P_{ij}(t) + \sum_{k\neq j} a_k P_{kj} P_{ik}(t),\ i,j\in\mathcal{S}, \tag{28}$$

which is $P'(t) = P(t)R,\ P(0) = I$ in matrix form. Although it turns out that the above method of derivation of the forward equations is not always justified (whereas our analogous derivation for the backward equations is justified), it does not matter, since the unique solution $P(t) = e^{Rt}$ 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 $P_{ij}(t)$ are the unique solution to both the Kolmogorov backward and forward equations.*

As an application, suppose the CTMC is positive recurrent. Then (recall (3)) for all $i,j\in\mathcal{S}$, $\lim_{t\to\infty} P_{ij}(t) = P_j$ which in turn implies that $\lim_{t\to\infty} P'_{ij}(t) = 0$. Thus from the forward equations (28) we obtain

$$0 = -a_j P_j + \sum_{i\neq j} a_i P_{ij} P_i,\ j\in\mathcal{S},$$

the balance equations. We conclude that if a (non-explosive) CTMC is positive recurrent, then the stationary probabilities must satisfy the balance equations (this then proves one direction of Theorem 1.1).