MA423 - Fundamentals of Operations Research
Lecture 6: Markov Chains & Queueing Theory

Katerina Papadaki[1]

Academic year 2017/18

[1]London School of Economics and Political Science. Houghton Street London WC2A 2AE (k.p.papadaki@lse.ac.uk)

Contents

1 Markov Chains - continued **2**
1.1 Markov chains: absorption . . . 2
1.1.1 The absorption probabilities . . . 2
1.1.2 The expected number of steps . . . 5
1.1.3 The expected cost . . . 6
1.1.4 The expected number of visits to a transient state . . . 6

2 Queueing Theory **9**
2.1 Introduction . . . 9
2.2 Key Factors Influencing the Life Cycles of Queues . . . 10
2.2.1 The Queueing System . . . 10
2.2.2 Input source and arrival process . . . 11
2.2.3 Queue Discipline . . . 12
2.2.4 Service Characteristics . . . 13
2.2.5 Queue Development . . . 13
2.3 The birth-and-death process . . . 14
2.3.1 The balance equations . . . 15

Chapter 1

Markov Chains - continued

1.1 Markov chains: absorption

As in the gambler's example, absorbing states can be used to represent different final outcomes of the process. In the presence of such states, one is typically interested in the probabilities of reaching them. As observed in the gambler's example, the n-step transition matrix converged to one with different rows, with the entries in the columns corresponding to the non-absorbing states being 0. In this specific example, every state was either absorbing or transient. However, as we will see below similar behaviour is observed when there are non-absorbing recurrent states.

1.1.1 The absorption probabilities

Consider the following example that unlike the gambler's example also has non-absorbing recurrent states.

Example 1.1. Suppose we have a Markov chain with states $\{0, 1, \ldots, 10\}$ and state transition diagram as shown in Figure 1.1.

What will happen in this Markov chain in the long run? The state classes are as follows: $\{0, 1\}$, $\{2\}$, $\{3, 4\}$, $\{5, 7, 8\}$, $\{6\}$, $\{9\}$, $\{10\}$. Some of these classes consist of transient states and some of recurrent states. The classes that have transient states are $\{2\}$, $\{3, 4\}$, $\{6\}$, $\{9\}$ and the classes that have recurrent states are $\{0, 1\}$, $\{5, 7, 8\}$ and $\{10\}$. Note that the states in single-element recurrent classes such as $\{10\}$ are absorbing states, i.e. 10 is an absorbing state.

Thus in the long run the system will eventually leave the transient states and end up in one of the classes of recurrent states. We can see from the diagram that if the system is currently in transient class $\{2\}$ it might end up in either recurrent classes $\{0, 1\}$ or $\{5, 7, 8\}$. If the system is currently in transient classes $\{3, 4\}$ or $\{6\}$ it will definitely end up in recurrent class $\{5, 7, 8\}$. And if the system is in transient class $\{9\}$ it will end up in recurrent class $\{10\}$. We would like to know which state/class will the system end up given that it is currently in state i?

Before we answer such questions we introduce some definitions and notation. A set of states is called *closed* if once the system is in it then it can never leave. In example 1.1 some sets

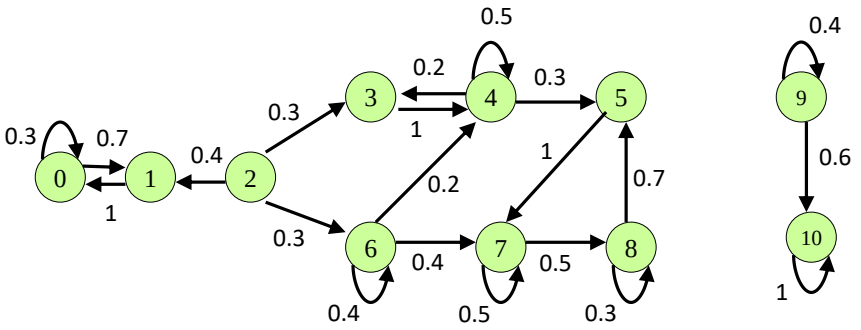

Figure 1.1: State transition diagram.

that are closed are $\{0, 1, \ldots, 10\}$, $\{0, 1, \ldots, 8\}$, $\{0, 1, 5, 7, 8\}$, $\{0, 1\}$, $\{5, 7, 8\}$ and $\{10\}$. Some sets that are not closed are $\{2\}$, $\{3, 4\}$, $\{6, 9\}$, $\{5, 7\}$, $\{6, 8\}$.

A closed set of states is called a *minimal closed set* if it has no proper subset that is closed. Minimal closed sets are $\{0, 1\}$, $\{5, 7, 8\}$ and $\{10\}$. It turns out that in a minimal closed set all states communicate and in fact the minimal closed sets are exactly the recurrent classes. You will prove this in the exercises. Thus, the system will eventually be absorbed by a minimal closed set or recurrent class where it will stay for ever. We sometimes call minimal closed sets or recurrent classes *absorbing classes*.

We let C be the set of state classes and we let $K \subseteq C$ be the set of recurrent classes (minimal closed sets or absorbing classes). Let T be the set of transient states. For an arbitrary state i and an absorbing class k, we define the *absorption probability* by

$$f_{ik} = \text{probability of being absorbed in absorbing class } k \text{ if starting from state } i.$$

Unlike for irreducible ergodic processes, the starting state now is important. Evidently, $f_{ik} = 1$ if $i \in k$ and $f_{ik} = 0$ if $i \in k'$ and k and k' are two different absorbing classes. We can think of absorbing classes as "one absorbing state" since once we are in there we are stuck there for ever. Thus, we are only concerned with transient states and absorbing classes.

If K is the set of absorbing classes, then the absorption probabilities can be obtained by solving the following system of linear equations.

$$f_{ik} = \sum_{j=0}^{M-1} f_{jk} p_{ij}, \quad \text{for all } i \in \{0, 1, \ldots, M-1\}, k \in K,$$

$$f_{ik} = \begin{cases} 1, & \text{if } i \in k, \\ 0, & \text{if } i \in k' \neq k, \end{cases} \quad \text{for all } k, k' \in K.$$

Note that the above equations are for $i \in \{0, 1, \ldots, M-1\}$. However, when i belongs to some absorbing class $k \in K$ then the probabilities are completely determined. Thus, only when i is transient, $i \in T$, do we need to find f_{ik}. Thus each absorbing class $k \in K$, we have $|T|$ unknowns f_{ik} $i \in T$ and $|T|$ equations $f_{ik} = \sum_{j=0}^{M-1} f_{jk} p_{ij}$ for $i \in T$.

In example 1.1 let us calculate the absorption probabilities for absorbing class $k = \{5, 7, 8\}$ and all transient states $T = \{2, 3, 4, 6, 9\}$ (we could ignore 9 since it is disconnected and we know that $f_{9k} = 0$ but we keep it for completeness):

$$\begin{aligned} f_{2k} &= 0.4 f_{1k} + 0.3 f_{3k} + 0.3 f_{6k} \\ f_{3k} &= 1.0 f_{4k} \\ f_{4k} &= 0.2 f_{3k} + 0.5 f_{4k} + 0.3 f_{5k} \\ f_{6k} &= 0.2 f_{4k} + 0.4 f_{6k} + 0.4 f_{7k} \\ f_{9k} &= 0.4 f_{9k} + 0.6 f_{10,k} \end{aligned} \tag{1.1}$$

Now we can get rid of the f_{ik} when i is recurrent by substituting their values. For example, $f_{1k} = 0$ since 1 is not a member of $k = \{5, 7, 8\}$; and $f_{5k} = f_{7k} = 1$ since $5, 7 \in k = \{5, 7, 8\}$. Similarly, $f_{10,k} = 0$. Thus the above system becomes:

$$\begin{aligned} f_{2k} &= 0.3 f_{3k} + 0.3 f_{6k} \\ f_{3k} &= 1.0 f_{4k} \\ f_{4k} &= 0.2 f_{3k} + 0.5 f_{4k} + 0.3 \\ f_{6k} &= 0.2 f_{4k} + 0.4 f_{6k} + 0.4 \\ f_{9k} &= 0.4 f_{9k} \end{aligned} \tag{1.2}$$

This is a system of $|T| = 5$ equations and $|T| = 5$ unknowns. Note that the last equation implies that $f_{9k} = 0$ as expected. In a similar manner we can compute the probabilities of being absorbed by $k' = \{0, 1\}$. However, since states $0, 1, \ldots, 8$ are disconnected from $9, 10$ we can look at them as their own Markov chain, and since within states $0, 1, \ldots, 8$ there are only two absorbing classes namely $\{0, 1\}$ and $\{5, 7, 8\}$ we know that the system will end up in one of these. Thus, if we are in state 2 and we calculated the probability of being absorbed by k to be f_{2k}, we can claim that $f_{2k'} = 1 - f_{2k}$.

Let us now compute these probabilities for the gambler's problem in the general form, with the parameter q. The two absorbing states are 0 and 3 and we use these to denote their corresponding absorbing classes $\{0\}$ and $\{3\}$. After substituting $f_{00} = f_{33} = 1$, $f_{03} = f_{30} = 0$, we obtain

$$\begin{aligned} f_{10} &= (1-q) + qf_{20}, \qquad & f_{13} &= qf_{23}, \\ f_{20} &= (1-q)f_{10}, & f_{23} &= (1-q)f_{13} + q. \end{aligned}$$

One can easily compute:

$$f_{10} = \frac{1-q}{q^2-q+1}, \quad f_{13} = \frac{q^2}{q^2-q+1}, \quad f_{20} = \frac{q^2-2q+1}{q^2-q+1}, \quad f_{23} = \frac{q}{q^2-q+1}.$$

If $q = 1$ (always win), then these give $f_{13} = f_{23} = 1$; if $q = 0$ (always loose), we obtain $f_{10} = f_{20} = 1$. For the choice $q = \frac{1}{3}$, we get $f_{10} = \frac{6}{7}$, $f_{13} = \frac{1}{7}$, $f_{20} = \frac{4}{7}$, $f_{23} = \frac{3}{7}$, as we did in the previous lecture when we calculated P^n when $n \to \infty$. Note that f_{10} and f_{13} sum to 1 (as expected).

1.1.2 The expected number of steps

We can determine the probabilities for reaching the different absorbing classes. A natural question arises: how long will it take in expectation to reach one of them - that is, how long will the process last? This can be computed similarly to the formulas above. For an arbitrary state i, let

$$t_i = \text{the expected number of steps it takes to reach an absorbing class from } i,$$

We have $t_i = 0$ for every i that belongs to an absorbing class k.

Let T be the set of transient states. Then the expected number of iterations to reach the different absorbing classes can be obtained by solving the following system of linear equations.

$$\begin{aligned} t_i &= 1 + \sum_{j=0}^{M-1} t_j p_{ij}, \quad && \text{for all } i \in T, \\ t_i &= 0, \quad && \text{for all } i \notin T. \end{aligned}$$

The term 1 in the equations represents the current step, which either leads to a desired absorbing class k, or leads to another state transient state j, from where we reach an absorbing class in expectation in t_j steps. As an illustration, assume the transition probabilities from state 1 are $p_{12} = p_{13} = \frac{1}{2}$, and it takes in expectation 4 steps to reach an absorbing class from state 2, and 6 steps from state 3. Then the equation on t_1 gives

$$t_1 = 1 + \frac{1}{2}t_2 + \frac{1}{2}t_3 = 1 + \frac{1}{2}\cdot 4 + \frac{1}{2}\cdot 6 = 6.$$

Let us solve the system for the gambler's problem with $q = \frac{1}{3}$.

$$\begin{aligned} t_0 &= t_3 = 0, \\ t_1 &= 1 + \frac{2}{3} \cdot 0 + \frac{1}{3} t_2, \\ t_2 &= 1 + \frac{2}{3} t_1 + \frac{1}{3} \cdot 0, \end{aligned}$$

Solving the system gives $t_1 = \frac{12}{7}$, $t_2 = \frac{15}{7}$; hence the game will typically terminate in less than two rounds if starting with £1, and slightly more than two rounds if starting with £2.

1.1.3 The expected cost

The above computation can be easily extended to the more general setting when traversing every state i has a certain cost (or profit) c_i associated, and we are interested in computing the expected total cost (or profit) until an absorbing class is reached. We define

$$T_i = \text{the expected cost until an absorbing class is reached from } i,$$

We have $T_i = 0$ for every i belonging to an absorbing class k. The expected number of steps computed above corresponds to the special case when every state has equal cost $t_i = 1$. Replacing 1 in the above formulas by c_i we obtain the following:

Let T be the set of transient states. Then the expected cost to reach the different absorbing classes can be obtained by solving the following system of linear equations.

$$\begin{aligned} T_i &= c_i + \sum_{j=0}^{M-1} T_j p_{ij}, \quad \text{for all } i \in T, \\ T_i &= 0, \quad \text{for all } i \notin T. \end{aligned}$$

1.1.4 The expected number of visits to a transient state

Sometimes we do not want to count every step but only certain steps. For example we might want to find the expected the number of visits to a transient state before ending up in an absorbing class.

Consider the Markov chain in Figure 1.2.

The transient states are $T = \{2, 3\}$ and the absorbing classes are $\{0\}$, $\{1, \}$. Suppose we are in state 2 and we want to find the expected number of visits to state 3 before getting absorbed. How do we do that?

For example, suppose that we start at state 2 and the states visited are $2 - 3 - 3 - 2 - 3 - 3 - 1 - \ldots$ then the number of visits to state 3 here before getting absorbed by $\{1\}$ is four

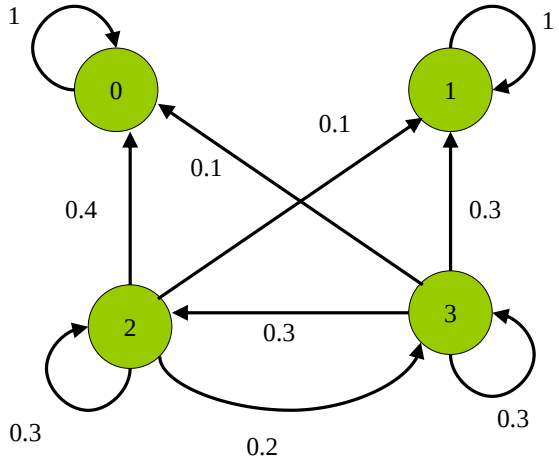

Figure 1.2: State transition diagram.

visits. If we start at state 3 and the states visited are $3-3-2-3-0$ then the number of visits to state 3 before getting absorbed by $\{0\}$ is three visits. But the equations that we used in section 1.1.2 $t_i = 1 + \sum_{j=0}^{M-1} t_j p_{ij}$ count steps not number of visits to a state.

One way to get around that is to count the steps either into state 3 or out of state 3. However, counting steps into state 3 in the case that the states visited are $3-3-2-3-0$ would count one step from $3-3$ and another step from $2-3$ but it would not count the initial visit to state 3. But if we count steps out of state 3 then this would count the step $3-3$, step $3-2$ and step $3-0$, which are exactly three steps for the three visits. Thus it is best to count the steps out of state 3. We define for $r \in T$ and $i \in \{0, 1, \ldots, M-1\}$:

v_i = the expected number of visits to state r before reaching an absorbing class from i,

Note that $v_i = 0$ if i belongs to an absorbing class because it will never visit transient state r. In the above example $r = 3$ and $T = \{2, 3\}$. Thus we find v_2 and v_3 as follows:

$$\begin{aligned} v_2 &= \quad 0.4v_0 + 0.1v_1 + 0.3v_2 + 0.2v_3 \\ v_3 &= 1 + \quad 0.1v_0 + 0.3v_1 + 0.3v_2 + 0.3v_3 \end{aligned} \tag{1.3}$$

Note that we put a 1 in the second equation because we are in state 3 and leaving state 3. In the first equation we are in state 2 and leaving state 2 and thus we do not count any of these steps. The second equation can also be written as:

$$v_3 = 0.1(1 + v_0) + 0.3(1 + v_1) + 0.3(1 + v_2) + 0.3(1 + v_3)$$

where after leaving state 3 say for state 0 we count the step we just took (one visit to 3) and the future visits from state 0, v_0.

Setting $v_0 = v_1 = 0$ we get:

$$\begin{aligned} v_2 &= \quad 0.3v_2 + 0.2v_3 \\ v_3 &= 1 + \quad 0.3v_2 + 0.3v_3 \end{aligned} \tag{1.4}$$

which gives us $v_2 = \frac{20}{43}$ and $v_3 = \frac{70}{43}$, which means that if we are in state 2 we will visit on average state 3 less than once and if we are in state 3 we will visit 3 a bit less than twice.

Chapter 2

Queueing Theory

2.1 Introduction

Queuing Theory is the study of queues, that is, situations in which units (possibly customers) arrive for a service and must wait if the service facility is occupied. When the service rate is too low or the arrival rate is too great, excess waiting is the result; if the service capacity is high relative to the arrival rate, idle capacity is the result. In order to balance the costs associated with these states, it is necessary to examine and quantify or qualify the behaviour characteristics of the system. These are usually random variables, and we are interested in their expected values.

A queue will have one or more servers, customers who arrive in some known pattern, and a service time distribution. The servers will usually form parallel channels, some or all of which may be specialised (for example, "cash payment only" or "less than ten items" in a supermarket). Queues may also be linked in series, usually when a customer must pass through several servers before completing service. An assembly line is an example of a series queuing system. Different queues may interfere with each other, as at traffic lights.

There will usually be a known inter-arrival time distribution for the "customers", but the arrival intervals may or may not be independent. For example, arrival intervals at traffic lights are often dependent. Arrivals may also occur in batches, to enter a theatre for example. Service times may also be dependent, such as when the server is a machine which has to be set up for each job, and the set up time depends on both the last job completed and the job which is about to be started. In the real world there may be a system for reducing the service time in emergencies or if the queue gets too long, and there may be a priority rule applying in the queue.

The customer population may be finite or infinite, and may be split into different categories (e.g. at the airport there are queues for arriving passengers which separate "U.K. or European Community citizens" from "other nationals"). The customers may react to the queue in various ways: balking, colluding, jockeying or reneging. Where one queue feeds into another the inter-departure pattern from the first is of interest, because it is the inter-arrival pattern for the second.

Lastly here most queuing systems are stochastic processes. Any particular sequence of arrivals and departures (even an infinite sequence) is simply a realisation of this stochastic process. The system may exhibit steady state behaviour after the influence of the initial conditions has died away. There may also be an initial period during which the queue shows some memory of its starting condition.

Queuing Theory, so far, cannot analyse all these possible queuing situations. It is not an optimisation technique, but it does derive analytical solutions to stochastic systems. Where the system gets beyond the range of current theory, or too complex, simulation techniques can be used to estimate results or properties about the system. Simulation has gradually become more widely used as computing power has made this more feasible.

2.2 Key Factors Influencing the Life Cycles of Queues

Three key factors influence the life cycles of individual queues:

- The arrival patterns of the 'customers', or 'items'.
- The logic of the queue behaviour.
- The characteristics of the service facility.

Almost every queuing system is a stochastic process, so that any one particular sequence of arrivals and departures gives one realisation (or simulation) of this stochastic process. Usually we wish to identify the expected values of the system characteristics, not to predict the future. We will be interested in questions such as the expected number of customers in the system at a particular point in time, or the expected time that a customer will spend in the system, or the percentage of the day that a server will be idle (i.e. not serving).

2.2.1 The Queueing System

- **The queuing system** (which in these notes we will also refer to simply as *the system*) is comprised of the people queuing to be served, and the people being served.

  For example, suppose there are currently ten people in a shop: nine customers and the shop keeper, who is serving one of the nine customers. Six of the other customers are still wandering around the shop, choosing goods to buy, and the other two are queuing, waiting to be served. Then the system, in this case, consists only of three people: the customer who is being served and the two who are queuing. The system does not include the server (or servers) or the customers who are still shopping.

- **State of the system** = number of customers in the queueing system. This is the number of customers queuing plus the number of customers being served.

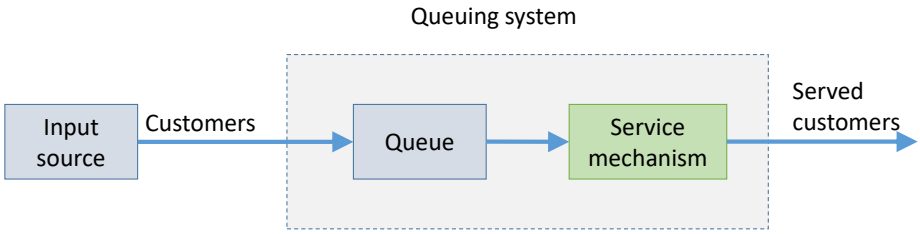

Figure 2.1: The queueing system.

2.2.2 Input source and arrival process

The input source of the queuing system is the set of potential customers who may enter the system. We will assume that the **input source has infinite size**. This is both for mathematical convenience (calculations are far easier in the infinite case) and because the assumption is often realistic, namely in the case where the population is very large.

We need to specify the probability distribution according to which customers join the queue. This is specified by the probability distribution of the *inter-arrival times*, that is, the time duration between two consecutive arrivals. We denote by $a(t)$ the probability density function of the inter-arrival times.

A common assumption, which will be made throughout this notes, is that the arrival process is a *Poisson process*. Let us denote by λ the arrival rate, that is, the expected number of arrivals per unit time (hour/day). In a Poisson process, arrivals satisfy the following two assumptions:

a) **Memorylessness:** arrivals defined on non-overlapping time intervals are independent. In other words, the number of arrivals in an interval after time t is independent on the number of arrivals before time t.

b) For $\delta > 0$ "very small" and for every time $t \in [0, 1)$, the probability of having more than one arrival in the time interval $[t, t+\delta)$ is negligible. The probability of exactly one arrival occurring in $[t, t+\delta)$ is essentially $\lambda\delta$, and the probability of 0 arrivals is essentially $1-\lambda\delta$.

Assume that the arrival process follows assumptions a) and b), and that λ is the arrival rate. Let $X(t)$ be the random variable denoting the number of arrivals in the time interval $[0, t)$. Then the probability of exactly n arrivals in $[0, t)$ follows the *Poisson Distribution* with parameter λ, that is,

$$\mathbb{P}[X(t) = n] = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

The mean number of arrivals in the time interval $[0, t)$ is

$$\mathbb{E}[X(t)] = \lambda t.$$

The inter-arrival times follow the exponential distribution with mean $1/\lambda$:

$$a(t) = \lambda e^{-\lambda t}.$$

For example, if on average three text messages might arrive at your mobile phone per hour between 09.00 and midnight, then $\lambda = 3$, and the expected inter-arrival time is $1/\lambda = 1/3$ of an hour, that is, 20 minutes.

We are assuming that the arrivals are independent on previous arrivals, and that at most one customer arrives at any given time. In practice there can be bulk arrivals (at a restaurant, for example). Often arrivals depend on the number of people already in the queue (for example, people might balk if the queue is too long), or on the time since the last arrival (e.g. buses and trains).

2.2.3 Queue Discipline

The queue discipline describes the method used to determine the order in which customers are served. Here we mention a few common examples:

- **FIFO** first in - first out, also known as "first come - first served".
- **LIFO** last in - first out.
- **Random** next customer in queue to be served is selected at random. For example, the machine operator might reach into a box full of parts and select one at random.
- **Priority queuing** customers are classified at arrival into different categories, each with a different priority level. Within each category, customers are served according to a FIFO discipline but servers are shared between categories: for a customer to be served in a category not only does the customer have to be the first in his queue but there must not be any other customers in categories of higher priority. This is routinely done in Accident & Emergency departments.

There may be a limit on queue size, or customers may leave the queue after a certain time if they still haven't been served. The simplest queue discipline and the one we assume here is "first come, first served". When calculating expected number of customers in the system the queue discipline is unimportant, but when looking at waiting times then the queue discipline needs to be taken into account. The expected waiting times will be the same but the variance of the waiting times will vary based on the queue discipline. For example, a LIFO queue will have higher variance for waiting times than the FIFO queue.

2.2.4 Service Characteristics

We denote by μ the *service rate*. This is the average number of customers served by the server per unti time if the server is continuously working. We will assume that service times follow an exponential distribution with parameter μ, that is,

$$s(t) = \mu e^{-\mu t}.$$

An advantage of the assumption above is the following property of the exponential distribution.

> The minimum of k independent exponentially distributed random variables with parameters $\mu_1, \ldots, \mu_k$ is an exponential random variable with parameter $\mu_1 + \mu_2 + \cdots + \mu_k$.

This assumption is very useful for service times in queuing systems with multiple servers. For example, we will use the above property often in the case where there are multiple independent servers all having an exponential service time distribution with the same parameter μ. In this case, if n people are being served at the same time, the service time of the system is the minimum of the service times among the n busy servers, and the service rate of the system is the sum of the service rates, that is, $n\mu$ (assuming all the servers are busy).

Another characteristic of the exponential distribution is that it is *memoryless*, that is the probability distribution of the remaining time until the next event occurs (arrival or service completion) is always the same regardless of how much time has already passed. For arrivals this means that the arrival of the next customer is uninfluenced by the when the last arrival occurred. However, for service times this means that however long the service has been going on the probability of completion is always the same. While mathematical convenient, the assumption of exponentially distributed service times is often not realistic.

2.2.5 Queue Development

A queuing system is described by the number of people in the system at any given time.

Let us denote by $p_n(t)$ the probability that there are n customers in the system at time t, for $n = 0, 1, 2, 3, \ldots$. Typically, after an *Initial Phase* (e.g. when the shop opens and for a short

time thereafter), the system goes through a *Transient Phase* where the distributions $p_n(t)$ may vary considerably with time t, but then this typically stabilizes, approaching a steady state. This is the case if the limit probability

$$p_n = \lim_{t \to +\infty} p_n(t)$$

exists for all n. The probability distribution p_n, $n = 0, 1, 2, 3, \ldots$ is the *steady state distribution*. This is the probability that in the long run the system will have n customers in it.

In the most general case, arrival and service rates depend on the number of people in the system. We will use the following notation.

λ_n = mean arrival rate (expected number of arrivals per unit time) at state n, $n = 0, 1, 2, 3, \ldots$.

μ_n = mean service rate (expected number of services completed per unit time) at state n, $n = 1, 2, 3, \ldots$.

Note that μ_0 is not defined, because no customer is served if the system is empty. Common scenarios are the one where λ_n is increasing in n (e.g. customers being attracted to a successful restaurant) or decreasing on n (e.g. customers entering a supermarket may balk (i.e. leave) if they observe that long queues are building up).

2.3 The birth-and-death process

In these notes we will assume that the inputs and outputs of the model behave according to a *birth-and-death process*, where a birth corresponds to an arrival in the system, and a death corresponds to a serviced customer.

The birth-and-death process is based on the following assumptions

Assumptions of birth-and-death process

1. In every state n, $n = 0, 1, 2, \ldots$, the inter-arrival times are exponential distributed with mean λ_n^{-1}.
2. In every state n, $n = 1, 2, \ldots$, the service times are exponential distributed with mean μ_n^{-1}.
3. The inter-arrivals time and service time are mutually independent.

Note that the last assumption implies that, if we reach state n, the next state transition is from n to $n + 1$ (birth) if the inter-arrival time is smaller than the service time, or from n to $n - 1$ (death) if the service time is smaller than the inter-arrival time.

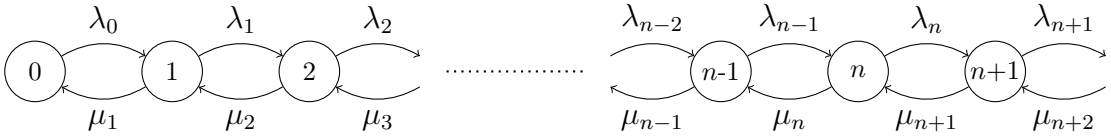

Figure 2.2: Representation of the birth-death process.

2.3.1 The balance equations

Recall that, in the steady state, the probability of being in state n at time t is a constant p_n independent of the time t. This means the following:

> **Rate in = Rate out principle.** In the steady state, for any state $n = 0, 1, 2, \ldots$ of the system, the mean rate at which the system enters a state n should be equal to the mean rate at which the system leaves state n

If the mean rate into n was greater than the mean rate out of n then the probability p_n would keep increasing (as n is increasingly visited more often) and thus the system would not be in equilibrium /steady-state.

As we shall see, the above principle will give rise to *balance equations* which fully describe the steady state probabilities $(p_n)_{n=0,1,2,\ldots}$.

Let us first concentrate on the balance equations for the state $n = 0$. Looking at Figure 2.2 we can see that we can only enter state $n = 0$ from state $n = 1$. The steady-state probability p_1 of being in state 1 represents the proportion of time that it would be possible for the process to enter state 0. Now given that the process is in state 1, the mean rate of entering state 0 is μ_1. In other words, for each unit of time spend in state 1, the expected number of times that it would leave state 1 to enter state 0 is μ_1. Since, from any other state this mean rate is 0, if we multiply μ_1 by p_1 we get the mean number of times that the process enters state 0 per unit time, which is the mean entering rate of 0: $\mu_1 p_1$.

Similarly, the mean leaving rate for 0 is: $p_0\lambda_0$. This is the proportion of time that the process is at 0 (p_0) multiplied by the mean number of times that the process leaves state 0 for state 1 per unit time (λ_0). This gives the mean number of times that the process leaves state 0 per unit time (since from state 0 it cannot go to any other state except 1).

Using the Rate in = Rate out principle we have:

$$p_0\lambda_0 = p_1\mu_1.$$

Similarly consider the general state n. Taking into account all four possible switches of state relative to state n, represented in Figure 2.2, there are two entering cases and two leaving cases, which must balance.

The mean entering rate for n is: $p_{n-1}\lambda_{n-1} + p_{n+1}\mu_{n+1}$ since we can enter state n from both states $n-1$ and $n+1$. And the mean leaving rate for n is: $p_n\lambda_n + p_n\mu_n$ since we can leave state n to enter states $n-1$ and $n+1$ either by a departure (death) or arrival (birth).

Hence we have:

$$p_{n-1}\lambda_{n-1} + p_{n+1}\mu_{n+1} = p_n\lambda_n + p_n\mu_n.$$

State	Balance equation
$n = 0$	$p_0\lambda_0 = p_1\mu_1.$
$n = 1$	$p_0\lambda_0 + p_2\mu_2 = p_1\lambda_1 + p_1\mu_1$
$n = 2$	$p_1\lambda_1 + p_3\mu_3 = p_2\lambda_n + p_2\mu_n$
$\vdots$	
n	$p_{n-1}\lambda_{n-1} + p_{n+1}\mu_{n+1} = p_n\lambda_n + p_n\mu_n$
$\vdots$	

Table 2.1: Balance equations for the steady state of the birth-death process.

We can use the balance equation for $n = 0$ in Table 2.1 to express the unknown p_1 in terms of the unknown p_0:

$$p_1 = \frac{\lambda_0}{\mu_1}p_0.$$

Thus once we know the value of p_0, and knowing the values of λ_0 and μ_1, we will be able to calculate p_1.

Similarly, substituting the value of p_1 obtained above into the balance equation for $n = 1$ in Table 2.1 we obtain

$$p_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2}p_0.$$

Similarly, substituting the values of p_1 and p_2 obtained above into the balance equation for $n = 1$ in Table 2.1 we obtain

$$p_3 = \frac{\lambda_0\lambda_1\lambda_2}{\mu_1\mu_2\mu_3}p_0.$$

Proceeding in this way, it is not difficult to see the following.

In the steady state, for $n = 1, 2, 3, \ldots$, the probability that n customers are in the system is

$$p_n = C_n p_0, \tag{2.1}$$

where we define

$$C_n := \frac{\lambda_0\lambda_1\lambda_2\cdots\lambda_{n-1}}{\mu_1\mu_2\mu_3\cdots\mu_n}.$$

Since $p_0 + p_1 + p_2 + p_3 + \cdots + p_n + \cdots = 1$, it follows that

$$p_0\left(1 + C_1 + C_2 + C_3 + \cdots\right) = 1,$$

therefore we have shown the following.

In the steady state, the probability that no customers are in the system is

$$p_0 = (1 + C_1 + C_2 + C_3 + \cdots)^{-1} . \tag{2.2}$$

Once the values of the arrival and departing rates are known, and if the steady state exists, the value p_0 can in principle be calculated using (2.2), and the values of p_n for $n = 1, 2, 3, \ldots$ are then completely determined by (2.1).

Many queuing systems are based on the birth-and-death process. This means that the input and the service mechanism satisfy assumptions 1, 2, 3 of the death-and-birth process. We will consider several cases, depending on whether there is only one server or multiple servers and on whether or not the arrival rates depend on the number of customers in the system.