Stochastic Operations Research 2
Part a

Ivo Adan and Jaap Wessels

Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Email: iadan@win.tue.nl, wessels@win.tue.nl

November 12, 1998

Contents

Chapter 1

Introduction

In this course we treat a number of basic queueing models. Attention is paid to basic methods for the analysis of these models, and also to applications of queueing models. Important application areas of queueing models are production systems, transportation and stocking systems, communication systems and information processing systems. Queueing models are particularly useful for the design of these system in terms of layout, capacities and control. Topics that are addressed are, e.g., the evaluation of perfomance characteristics such as the waiting time, effects of priority rules and effects of assignment rules.

The central theme in this course is the following basic approach for the analysis of queueing models. First the system is described by a Markov process with an appropriately chosen state concept. Then the balance equations for the state probabilities are formulated, and subsequently these equations are solved. For the solution of the balance equations a number of analytical and numerical approaches will be discussed. Finally the state probabilities may be used to determine performance characteristics such as, e.g., the mean waiting time and the mean number of customers in the system.

The organization is as follows. Chapter 2 first discusses a number of basic concepts and results from probability theory that we will use. In chapter 3 we briefly refresh some well known results for analysis of the linear difference and differential equations. Elementary queueing models and simple variants of these models are treated in the chapters 4 and 5. The last chapter is devoted to queueing models with priority rules.

The text contains a lot of exercises.

Chapter 2

Basic concepts

This chapter is devoted to some basic concepts from probability theory and queueing theory. These concepts can be found is every standard textbook on these topics, see e.g. [27, 36, 46].

2.1 Random variable

In this report random variables are denoted by capitals, X, Y, etc. The expected value or mean of X is denoted by $E(X)$ and its variance by $\mathrm{var}(X)$, or by $\sigma^2(X)$ where $\sigma(X)$ is the standard deviation of X.

An important quantity is the *coefficient of variation* of the positive random variable X defined as

$$c_X = \frac{\sigma(X)}{E(X)}.$$

The coefficient of variation is a measure of the variability of the random variable X. For instance, the deterministic distribution has $c_X = 0$ and the exponential distribution has $c_X = 1$.

2.2 Useful probability distributions

This section discusses a number of important distributions which have been found useful for describing random variables in many applications.

2.2.1 Exponential distribution

The density of an exponential distribution with parameter μ is given by

$$f(t) = \mu e^{-\mu t}, \qquad t > 0.$$

The distribution function equals

$$F(t) = 1 - e^{-\mu t}, \qquad t \geq 0.$$

For this distribution we have

$$E(X) = \frac{1}{\mu}, \qquad \mathrm{var}(X) = \frac{1}{\mu^2}, \qquad c_X = 1.$$

An important property of an exponential random variable X with parameter μ is the *memoryless property.* This property states that for all $x \geq 0$ and $t \geq 0$,

$$P(X > x + t | X > t) = P(X > x) = e^{-\mu x}.$$

So the remaining lifetime of X, given that X is still alive at time t, is again exponentially distributed with the same mean $1/\mu$. We often use the memoryless property in the form

$$P(X < t + \Delta t | X > t) = 1 - e^{\mu \Delta t} = \mu \Delta t + o(\Delta t), \qquad (\Delta t \to 0). \tag{2.1}$$

2.2.2 Erlang distribution

A random variable X has an *Erlang-k* ($k = 1, 2, \ldots$) distribution with mean k/μ if X is the sum of k independent random variables $X_1, \ldots, X_k$ having a common exponential distribution with mean $1/\mu$. The common notation is $E_k(\mu)$ or briefly E_k. The density of an $E_k(\mu)$ distribution is given by

$$f(t) = \mu \frac{(\mu t)^{k-1}}{(k-1)!} e^{-\mu t}, \qquad t > 0.$$

The distribution function equals

$$F(t) = 1 - \sum_{j=0}^{k-1} \frac{(\mu t)^j}{j!} e^{-\mu t}, \qquad t \geq 0.$$

The parameter μ is called the scale parameter, k is the shape parameter. A phase diagram of the E_k distribution is shown in figure 2.1.

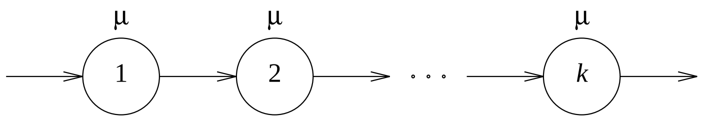

Figure 2.1: Phase diagram for the Erlang-k distribution with scale parameter μ

The mean, variance and squared coefficient of variation are equal to

$$E(X) = \frac{k}{\mu}, \qquad \mathrm{var}(X) = \frac{k}{\mu^2}, \qquad c_X^2 = \frac{1}{k}.$$

A convenient distribution arises when we mix an E_{k-1} and E_k distribution with the same scale parameters. The notation used is $E_{k-1,k}$. A random variable X has an $E_{k-1,k}(\mu)$ distribution, if X is with probability p (resp. $1-p$) the sum of $k-1$ (resp. k) independent exponentials with common mean $1/\mu$. The density of this distribution has the form

$$f(t) = p\mu\frac{(\mu t)^{k-2}}{(k-2)!}e^{-\mu t} + (1-p)\mu\frac{(\mu t)^{k-1}}{(k-1)!}e^{-\mu t}, \qquad t > 0,$$

where $0 \leq p \leq 1$. As p runs from 0 to 1, the squared coefficient of variation of the mixed Erlang distribution varies from $1/(k-1)$ to $1/k$. It will appear (later on) that this property is useful for fitting distributions on the first two moments.

2.2.3 Hyperexponential distribution

A random variable X is hyperexponentially distributed if X is with probability p_i, $i = 1, \dots, k$ an exponential X_i with mean $1/\mu_i$. The notation used is $H_k(p_1, \dots, p_k; \mu_1, \dots, \mu_k)$, or simply H_k. The density is given by

$$f(t) = \sum_{i=1}^{k} p_i\mu_i e^{-\mu_i t}, \qquad t > 0,$$

and the mean is equal to

$$E(X) = \sum_{i=1}^{k} \frac{p_i}{\mu_i}.$$

The coefficient of variation c_X of this distribution is greater or equal to 1 (see exercise 2.12.1). A phase diagram of the H_k distribution is shown in figure 2.2.

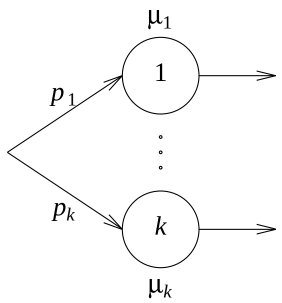

Figure 2.2: Phase diagram for the hyperexponential distribution

2.2.4 Phase-type distribution

The preceding distributions are all special cases of the *phase-type distribution*. The notation is PH. This distribution is characterized by a Markov chain with states $1, \ldots, k$ (the so-called phases) and a transition probability matrix P which is *transient*. This means that P^n tends to zero as n tends to infinity. In words, eventually you will always leave the Markov chain. The residence time in state i is exponentially distributed with mean $1/\mu_i$, and the Markov chain is entered with probability p_i in state i, $i = 1, \ldots, k$. Then X has a phase-type distribution if X is the total residence time in the preceding Markov chain, i.e. X is the total time elapsing from start in the Markov chain till departure from the Markov chain.

We mention two important classes of phase-type distributions which are *dense in the class of all non-negative distribution functions*. This is ment in the sense that for any non-negative distribution function $F(\cdot)$ a sequence of phase-type distributions can be found which pointwise converges at the points of continuity of $F(\cdot)$. The denseness of the two classes makes them very useful as a practical modelling tool. A proof of the denseness can be found in [39, 40]. The first class is the class of *Coxian distributions*, notation C_k, and the other class consists of *mixtures of Erlang distributions with the same scale parameters*. The phase representations of these two classes are shown in the figures 2.3 and 2.4

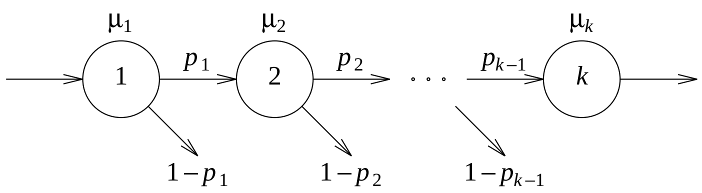

Figure 2.3: Phase diagram for the Coxian distribution

A random variable X has a Coxian distribution of order k if it has to go through up to at most k exponential phases. The mean length of phase n is μ_n, $n = 1, \ldots, k$. It starts in phase 1. After phase n it comes to an end with probability $1 - p_n$ and it enters the next phase with probability p_n. Obviously $p_k = 0$. For the Coxian-2 distribution it holds that the squared coefficient of variation is greater or equal to 0.5 (see exercise 2.12.6).

A random variable X has a mixed Erlang distribution of order k if it is with probability p_n the sum of n exponentials with the same mean $1/\mu$, $n = 1, \ldots, k$.

2.3 Fitting distributions

In practice it often occurs that the only information of random variables that is available is their mean and standard deviation, or if one is lucky, some real data. To obtain an approximating distribution it is common to fit a phase-type distribution on the mean, $E(X)$, and the coefficient of variation, c_X, of a given positive random variable X, by using the following simple approach.

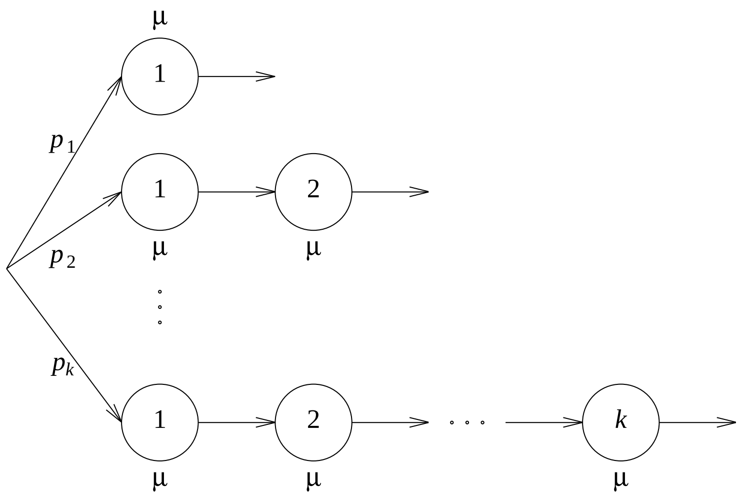

Figure 2.4: Phase diagram for the mixed Erlang distribution

In case $0 < c_X < 1$ one fits an $E_{k-1,k}$ distribution (see subsection 2.2.2). More specifically, if

$$\frac{1}{k} \leq c_X^2 \leq \frac{1}{k-1},$$

for certain $k = 2, 3, \ldots$, then the approximating distribution is with probability p (resp. $1-p$) the sum of $k-1$ (resp. k) independent exponentials with common mean $1/\mu$. By choosing (see e.g. [46])

$$p = \frac{1}{1+c_X^2}[kc_X^2 - \{k(1+c_X^2) - k^2c_X^2\}^{1/2}], \qquad \mu = \frac{k-p}{E(X)},$$

the $E_{k-1,k}$ distribution matches $E(X)$ and c_X.

In case $c_X \geq 1$ one fits a $H_2(p_1, p_2; \mu_1, \mu_2)$ distribution. The hyperexponential distribution however is not uniquely determined by its first two moments. In applications, the H_2 distribution with *balanced means* is often used. This means that the normalization

$$\frac{p_1}{\mu_1} = \frac{p_2}{\mu_2}$$

is used. The parameters of the H_2 distribution with balanced means and fitting $E(X)$ and c_X (≥ 1) are given by

$$p_1 = \frac{1}{2}\left(1 + \sqrt{\frac{c_X^2 - 1}{c_X^2 + 1}}\right), \qquad p_2 = 1 - p_1,$$

$$\mu_1 = \frac{2p_1}{E(X)}, \qquad \mu_1 = \frac{2p_2}{E(X)}.$$

It also possible to make a more sophisticated use of phase-type distributions by, e.g., trying to match the first three (or even more) moments of X or to approximate the shape of X. For the case $c_X^2 > 1$ it is shown in [50] that a three moment fit to the random variable X by a Coxian-2 distribution exists only if the third moment of X satisfies

$$E(X^3) \geq \frac{3}{2}(1 + c_X^2)^2 E(X)^3. \tag{2.2}$$

For the case $0.5 \leq c_X^2 \leq 1$ in [21] lower and upper bounds on the third moment of X are given such that a three moment fit by a C_2 distribution exists. If it exists, then the parameters μ_1, μ_2 and p_1 of the three moment fit are given by

$$\mu_1 = \frac{1}{2}\left[a_1 + \sqrt{a_1^2 - 4a_2}\right], \quad \mu_2 = \frac{1}{2}\left[a_1 - \sqrt{a_1^2 - 4a_2}\right], \quad p_1 = \frac{\mu_2(m_1\mu_1 - 1)}{\mu_1},$$

where $m_i = E(X^i)$, $i = 1, 2, 3$, and

$$a_2 = \frac{6m_1^2 - 3m_2}{6m_2^2/4 - m_1 m_3} \qquad a_1 = \frac{1}{m_1} + \frac{m_2 a_2}{2m_1}.$$

The Coxian-2 distribution can of course also be used for a two-moment fit. The following set is suggested by [32],

$$\mu_1 = 2E(X), \quad p_1 = 0.5/c_X^2, \quad \mu_2 = \mu_1 p_1.$$

In [25] a queueing model is used to explore the quality of two- and three-moment approximations, and in some cases, approximations considering the distribution shape.

Phase-type distributions may of course also naturally arise in practical applications. For example, if the processing of a job involves performing several tasks, where each task takes an exponential amount of time, then the processing time can be described by an Erlang distribution.

2.4 Poisson process

Let $N(t)$ be the number of arrivals in $[0, t]$ for a *Poisson process* with intensity λ, i.e. the time between successive arrivals is exponentially distributed with parameter λ and independent of the past. Then $N(t)$ has a *Poisson distribution* with parameter λt, so

$$P(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \qquad k = 0, 1, 2, \ldots$$

The mean, variance and coefficient of variation of $N(t)$ are equal to

$$E(N(t) = \lambda t, \qquad \text{var}(X) = \lambda t, \qquad c_{N(t)}^2 = \frac{1}{\lambda t}.$$

The Poisson process is an extremely useful process for modelling purposes in many practical applications, such as, e.g. to model arrival processes for queueing models or demand processes for inventory systems. It is empirically found that in many circumstances the arising stochastic processes can be well approximated by a Poisson process.

Next we mention two important properties of a Poisson process (see e.g. [36]).

(i) Suppose that $N_1(t)$ and $N_2(t)$ are two independent Poisson processes with respective intensities λ_1 and λ_2. Then the sum $N_1(t) + N_2(t)$ of the two processes is again a Poisson process with intensity $\lambda_1 + \lambda_2$.

(ii) Suppose that $N(t)$ is a Poisson process with intensity λ and that each arrival is marked with probability p independent of all other arrivals. Let $N_1(t)$ and $N_2(t)$ denote respectively the number of marked and unmarked arrivals in $[0, t]$. Then $N_1(t)$ and $N_2(t)$ are both Poisson processes with respective intensities λp and $\lambda(1-p)$. And these two processes are independent.

So Poisson processes remain Poisson under merging and splitting.

2.5 Queueing models and Kendall's notation

A queueing model is, among others, characterized by the arrival process of customers, the service times and the numbers of servers. Usually we assume that the interarrival times are independent and have a common distribution. The type of distribution is specified by a letter. For example, for a general distribution the letter G is used, M for the exponential distribution (M stands for Memoryless) and D for deterministic times. The same letters are used for the service times, which are also assumed to be independent and identically distributed. The basic queueing model is shown in figure 2.5. It can be used to model, e.g., machines or operators processing orders or communication equipment processing information.

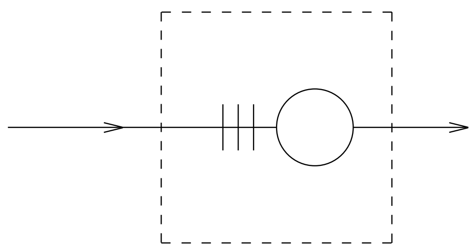

Figure 2.5: Basic queueing model

The basic model can be described by a three-part code $a/b/c$. The first letter specifies the interarrival time distribution, the second one the service time distribution and the last one the number of servers. Some examples are $M/M/1$, $M/M/c$, $M/G/1$, $G/M/1$ and $M/D/1$. The notation can be extended with an extra letter to cover other queueing models. For example, a system with exponential interarrival and service times, one server and having waiting room only for N customers (including the one in service) is abbreviated by the four letter code $M/M/1/N$.

Queueing models have been proved to very useful in many practical applications in areas such as, e.g., production systems, inventory systems and communication systems.

These applications concern in particular *design problems*, where answers are required to questions like: What should be the layout?, Is the capacity sufficient? or How do we have to divide work among several capacities? An explanation why queueing models are important is that in many applications the variability in the arrival and service process have a dominant effect on the behavior of the system. This will be demonstrated by many examples treated in the following chapters.

2.6 Little's law

Little's law (see [31]) gives a very important relation between the mean sojourn time in the system $E(S)$ on one hand and the mean number of customers in the system $E(L)$ and the average number of customers entering the system per unit time, denoted by λ. Little's law states that

$$E(L) = \lambda E(S).$$

Intuitively this result can be understood as follows. Suppose that all customers pay 1 dollar per unit time while in the system. This money can be earned in two ways. The first possibility is to let pay all customers 'continuously' in time. Then the average reward earned by the system equals $E(L)$ dollar per unit time. The second possibility is to let customers pay 1 dollar per unit time for their residence in the system when they leave. In equilibrium, the average number of customers leaving the system per unit time is equal to the average number of customers entering the system. So the system earns an average reward of $\lambda E(S)$ dollar per unit time. Obviously, the system earns the same in both cases. For a rigorous proof, see [43].

To demonstrate the use of Little's law consider the basic queueing model in figure 2.5 with one server. For this model we can derive relations between several performance measures by applying Little's law to suitably defined (sub)systems. Application of Little's law to the system consisting of queue plus server yields

$$E(L) = \lambda E(S),$$

where $E(L)$ is the mean number of customers in the system, λ the arrival rate and $E(S)$ the mean sojourn time (waiting time plus service time). Applying Little's law to the queue (excluding the server) yields a relation between the queue length L^q and the waiting time W, namely

$$E(L^q) = \lambda E(W).$$

Finally, when we apply Little's law to the server only, we obtain

$$\rho = \lambda E(B),$$

where ρ is the mean number of customers at the server (which is the same as the fraction of time the server is working) and $E(B)$ the mean service time.

2.7 PASTA property

For queueing systems with Poisson arrivals, so for $M/\cdot/\cdot$ systems, the very special property holds that arriving customers find on the average the same situation in the queueing system as someone who is passing by (and not entering) at a random moment. More precisely, the fraction of customers finding on arrival the system in some state A is exactly the same as the fraction of time the system is in state A. In general this is not true (in fact, it is only true for Poisson arrivals). For instance, in a $D/D/1$ system which is empty at time 0, and with arrivals at $1, 3, 5, \ldots$ and service times 1, every arriving customer finds an empty system, whereas the fraction of time the system is empty is $1/2$.

This property of Poisson arrivals is called PASTA property, which is the acrynom for Poisson Arrivals See Time Averages. Intuitively the PASTA property may be explained by the fact that Poisson arrivals occur completely random in time, i.e., in every small time interval $[t, t+\Delta t]$ a customer arrives with probability Δt, independent of the history of the system before time t (see (2.1)). A rigorous proof of the PASTA property can be found in [54, 55].

In the following chapters we will show that in many queueing models it is possible to determine mean performance measures directly (i.e. not from the distribution of these measures) by using the PASTA property and Little's law. This technique is called the *mean value technique*.

2.8 Residual service time

In an $M/G/1$ system an arriving customer may find several customers waiting for service and one in service. What is the residual service time of the one in service? The density of the complete service time of the one in service is equal to (see e.g. [27])

$$f(x) = \frac{x f_B(x)}{\int_0^\infty y f_B(y) dy}, \qquad x > 0,$$

where $f_B(x)$ is the density of the service time distribution. So the mean service time for the job in service is equal to $E(B^2)/E(B)$. Since a customer arrives on the average in the middle of a service time, we find for the mean residual service time of the one in service that

$$E(R) = \frac{E(B^2)}{2E(B)} = \frac{1}{2}(1 + c_B^2)E(B).$$

2.9 Occupation rate

In a $G/G/1$ system with arrival rate λ and mean service time $E(B)$ the amount of work arriving per unit time equals $\lambda E(B)$. The server can handle 1 unit work per unit time. So

$\lambda E(B)$ must be less than 1, since otherwise the amount of work in the system will grow to infinity. It is common to use the notation

$$\rho = \lambda E(B)$$

where ρ is called the *occupation rate* or server utilization. In a multi server system $G/G/c$ we have to require $\lambda E(B) < c$. The occupation rate per server is $\rho = \lambda E(B)/c$ and the mean number of busy servers equals $\lambda E(B)$ (which follows from Little's law).

2.10 Generating function

Let X be a nonnegative discrete random variable with $P(X = n) = p(n)$, $n = 0, 1, 2, \ldots$. Then the generating function $P(z)$ of X is defined as

$$P(z) = \sum_{n=0}^{\infty} p(n) z^n.$$

Note that $|P(z)| \leq 1$ for all $|z| \leq 1$. Further

$$P(0) = p(0), \qquad P(1) = 1, \qquad P'(1) = E(X).$$

2.11 Laplace-Stieltjes transform

The Laplace-Stieltjes transform $\tilde{X}(s)$ of a nonnegative random variable X is defined as

$$\tilde{X}(s) = E(e^{-sX}), \qquad s \geq 0.$$

For the transform of the sum $Z = X + Y$ of two independent random variables X and Y it holds that

$$\tilde{Z}(s) = \tilde{X}(s)\tilde{Y}(s).$$

Further

$$E(X^k) = (-1)^k \left. \frac{d^k X(s)}{ds^k} \right|_{s=0}.$$

In part a of this course we will hardly use generating functions or Laplace-Stieltjes transforms. However, Part b will make extensive use of these powerful tools.

2.12 Exercises

Exercise 2.12.1
Show that the coefficient of variation of a hyperexponential distribution is greater or equal to 1.

Exercise 2.12.2
Let $X_1, \ldots, X_n$ be independent exponential random variables with mean $E(X_i) = 1/\mu_i$, $i = 1, \ldots, n$. Define

$$Y_n = \min(X_1, \ldots, X_n), \qquad Z_n = \max(X_1, \ldots, X_n).$$

Determine the distributions of Y_n and Z_n.

Exercise 2.12.3
Let $X_1, X_2, \ldots$ be independent exponential random variables with mean $1/\mu$ and let N be geometrically distributed:

$$P(N = k) = (1 - p)p^{k-1}$$

with $k = 1, 2, \ldots$ and $0 \leq p < 1$. Show that S defined as

$$S = \sum_{n=1}^{N} X_n$$

is again exponentially distributed with parameter $(1 - p)\mu$.

Exercise 2.12.4 *(Poisson process)*
Suppose that arrivals occur at $T_1, T_2, \ldots$. The interarrival times $A_n = T_n - T_{n-1}$ are independent and have common exponential distribution with mean $1/\lambda$, where $T_0 = 0$ by convention. Let $N(t)$ denote the number of arrivals is $[0, t]$ and define for $n = 0, 1, 2, \ldots$

$$p_n(t) = P(N(t) = n), \qquad t > 0.$$

(i) Determine $p_0(t)$.

(ii) Show that for $n = 1, 2, \ldots$

$$p_n'(t) = -\lambda p_n(t) + \lambda p_{n-1}(t), \qquad t > 0,$$

with initial condition $p_n(0) = 0$.

(iii) Solve the preceding differential equations for $n = 1, 2, \ldots$

Exercise 2.12.5 *(Fitting a distribution)*
Suppose that processing a job on a certain machine takes on the average 4 minutes with a standard deviation of 3 minutes. Show that if we model the processing time as a mixture of an Erlang-1 (exponential) distribution and an Erlang-2 distribution with density

$$f(t) = p\mu e^{-\mu t} + (1-p)\mu^2 t e^{-\mu t},$$

the parameters p and μ can be chosen in such a way that this distribution matches the mean and standard deviation of the processing times on the machine.

Exercise 2.12.6
Consider a random variable X with a Coxian-2 distribution with parameters μ_1 and μ_2 and branching probability p_1.

(i) Show that we may assume without loss of generality that $\mu_1 \geq \mu_2$.

(ii) Show that $c_X^2 \geq 0.5$.

Exercise 2.12.7
Let X and Y be exponentials with parameters μ and λ, respectively. Suppose that $\lambda < \mu$. Let Z be equal to X with probability λ/μ and equal to $X + Y$ with probability $1 - \lambda/\mu$. Show that Z is an exponential with parameter λ.

Exercise 2.12.8
Consider a H_2 distribution with parameters $\mu_1 > \mu_2$ and branching probabilities q_1 and q_2, respectively. Show that the C_2 distribution with parameters μ_1 and μ_2 and branching probability p_1 given by

$$p_1 = 1 - (q_1\mu_1 + q_2\mu_2)/\mu_1\,,$$

is equivalent to the H_2 distribution.

Chapter 3

Refresher

As we will see in the following chapters, the balance equations for the state probabilities form, in many cases, a system of recurrence relations or differential equations. The aim of this section is to refresh some well-known results from the theory of linear recurrence relations and linear differential equations with constant coefficients. The results can be found in standard textbooks, such as e.g. [13]. We first start to review some results for recurrence relations and then we treat differential equations. The results for these two types of equations are very similar.

3.1 Linear recurrence relations of order k

A recurrence relation of the form

$$x_{n+k} + a_1 x_{n+k-1} + a_2 x_{n+k-2} + \cdots + a_k x_n = b_n, \qquad n = 0, 1, 2, \ldots \tag{3.1}$$

is called a *linear recurrence relation with constant coefficients of order k*. If $b_n = 0$ for all n, then (3.1) is called *homogeneous*, and *inhomogeneous* otherwise. Solutions of recurrence relations are not unique. To make it unique one has to impose extra conditions. One possibility is to prescribe the starting values $x_0, \ldots, x_{k-1}$. In this case we speak of an *initial-value problem.* The solution of an initial-value problem is unique. It is also possible to impose other conditions. For example, we may want to solve a second order recurrence relation with prescribed values at two end points, say x_0 and x_N. Then we have to determine $x_1, \ldots, x_{N-1}$ given x_0 and x_N. Such a problem is usually referred to as a *boundary-value problem.* The solution of a boundary-value problem may not exist, and if it does exist, it is not necessarily unique.

Example 3.1.1 *(The Fibonacci series)*
One of the most famous and oldest examples of a recurrence relation was formulated by Leonardo Pisano in his book *Liber Abaci.* He formulates a model to describe the offspring of one pair of rabbits. In this model he assumes that each pair of rabbits, from the second month they were born, produces each month one pair of young rabbits. Denote the number

of pairs in month n by x_n. We start with one pair, so $x_0 = x_1 = 1$. Assuming that no rabbits die, the number of pairs in month $n+2$ is the sum of the number of pairs in month $n+1$ and the number of pairs born in that month, which is equal to the number of pairs in month n. So we obtain the recurrence relation

$$x_{n+2} = x_{n+1} + x_n, \qquad n = 0, 1, 2, \dots \tag{3.2}$$

with initial values $x_0 = x_1 = 1$. The relation (3.2) is a homogeneous second order linear recurrence relation with coefficients $a_1 = a_2 = -1$. We get for $x_0, x_1, x_2, \dots$ the following, so-called Fibonacci sequence

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \dots$$

The quotients x_{n+1}/x_n seem to converge to $1.618\dots$. This number is known as the golden mean, which is the desired quotient between the width and height of e.g. paintings. Does it indeed converge? To what? How can we prove that?

We now proceed with the analysis of recurrence relations. Let us first consider the homogeneous recurrence relation

$$x_{n+k} + a_1 x_{n+k-1} + a_2 x_{n+k-2} + \cdots + a_k x_n = 0, \qquad n = 0, 1, 2, \dots \tag{3.3}$$

If two sequences $x_n^(1)$ and $x_n^{(2)}$ are solutions of (3.3) and c_1 and c_2 are two complex numbers, then the sequence $\{c_1 x_n^{(1)} + c_2 x_n^{(2)}\}$ is again a solution of (3.3). This shows that the solutions form a vector space.

To show that l solutions $\{x_n^{(1)}\}, \dots, \{x_n^{(l)}\}$ are linearly independent, one has to check that the relation

$$c_1 x_n^{(1)} + \cdots + c_l x_n^{(l)} = 0, \qquad n = 0, 1, 2, \dots$$

is only possible if all coefficients $c_1, \dots, c_l$ are zero. The following result states that it suffices to check this for $n = 0, \dots, k-1$. It is a direct consequence of the fact that when $\{x_n\}$ is a solution of (3.3) with $x_0 = \cdots = x_{k-1} = 0$, then $x_n = 0$ for all n.

Lemma 3.1.2 *Let the sequences $\{x_n^{(1)}\}, \dots, \{x_n^{(l)}\}$ be solutions of (3.3). These l solutions are linearly independent if and only if the vectors*

$$\begin{pmatrix} x_0^{(1)} \\ x_1^{(1)} \\ \vdots \\ x_{k-1}^{(1)} \end{pmatrix}, \begin{pmatrix} x_0^{(2)} \\ x_1^{(2)} \\ \vdots \\ x_{k-1}^{(2)} \end{pmatrix}, \dots, \begin{pmatrix} x_0^{(l)} \\ x_1^{(l)} \\ \vdots \\ x_{k-1}^{(l)} \end{pmatrix}$$

are linearly independent.

From this lemma we immediately obtain the following result.

Theorem 3.1.3 *The set of all solutions of (3.3) forms a k-dimensional vector space over the complex field.*

This result is useful to find the general solution of (3.3). We first try to find k linearly independent solutions $\{x_n^{(1)}\}, \ldots, \{x_n^{(k)}\}$ of (3.3). Then the general solution can be expressed as a linear combination of these k basis solutions, i.e.,

$$z_n = c_1 x_n^{(1)} + \cdots + c_k x_n^{(k)}, \qquad n = 0, 1, 2, \ldots,$$

where $c_1, \ldots, c_k$ are arbitrary complex numbers. To obtain k independent solutions we try solutions of the form

$$x_n = y^n.$$

Substituting this form into (3.3) and then dividing by common factors we get

$$y^k + a_1 y^{k-1} + a_2 y^{k-2} + \cdots + a_k = 0. \tag{3.4}$$

This equation is called the *characteristic equation* of (3.3). Suppose that this equation has k distinct roots $y_1, \ldots, y_k$. Then the solutions $\{y_1^n\}, \ldots, \{y_k^n\}$ are linearly independent, which follows lemma 3.1.2 and the fact that the *VanderMonde matrix*

$$\begin{pmatrix} 1 & 1 & \ldots & 1 \\ y_1 & y_2 & \cdots & y_k \\ \vdots & \vdots & & \vdots \\ y_1^{k-1} & y_2^{k-1} & \ldots & y_k^{k-1} \end{pmatrix}$$

is non-singular. Namely, suppose that we have a linear combination of rows which vanishes, i.e. there are coefficients $c_0, \ldots, c_{k-1}$ such that

$$c_0 + c_1 y_i + \cdots + c_{k-1} y_i^{k-1} = 0, \qquad i = 1, \ldots, k.$$

This means that the polynomial

$$p(y) := c_0 + c_1 y + \cdots + c_{k-1} y^{k-1}$$

has $y_1, \ldots, y_k$ as zeros. Since the degree of $p(\cdot)$ is k this implies that $p(\cdot)$ is identical to zero, so $c_0 = \ldots = c_{k-1} = 0$. Hence the general solution of (3.3) is given by

$$z_n = c_1 y_1^n + \cdots + c_k y_k^n, \qquad n = 0, 1, 2, \ldots$$

Example 3.1.4 *(The Fibonacci series)*
The characteristic equation of recurrence relation (3.2) is

$$y^2 = y + 1, \tag{3.5}$$

with roots

$$y_1 = \frac{1}{2}(1 + \sqrt{5}), \qquad y_2 = \frac{1}{2}(1 - \sqrt{5}).$$

Hence the general solution of (3.2) is given by

$$x_n = c_1 y_1^n + c_2 y_2^n, \qquad n = 0, 1, 2, \cdots.$$

The coefficients c_1 and c_2 follow from the initial conditions $x_0 = x_1 = 1$, yielding

$$c_1 = \frac{1}{2} + \frac{1}{10}\sqrt{5}, \qquad c_2 = \frac{1}{2} - \frac{1}{10}\sqrt{5}.$$

This completes the determination of the Fibonacci series. In particular, the quotients x_{n+1}/x_n converge to $y_1 = 1.618\ldots$ as $n \to \infty$.

In case the characteristic equation (3.4) has multiple roots, we have to find for each root y with multiplicity m also m independent solutions. It is easy to verify that when we take, if y is a root of (3.4) with multiplicity m, the m sequences

$$y^n, ny^{n-1}, \binom{n}{2} y^{n-2}, \ldots, \binom{n}{m-1} y^{n-m+1}, \qquad n = 0, 1, 2, \ldots$$

as solutions, and we do so for each root, then we get k independent solutions of (3.3). Summarizing, we find the following result.

Theorem 3.1.5 *Let $y_1, \ldots, y_l$ be the roots of the characteristic equation (3.4) and suppose that y_i has multiplicity m_i, $i = 1, \ldots, l$. Then the sequences*

$$y_1^n, ny_1^{n-1}, \binom{n}{2} y_1^{n-2}, \ldots, \binom{n}{m_1 - 1} y_1^{n-m_1+1},$$

$$\ldots$$

$$y_l^n, ny_l^{n-1}, \binom{n}{2} y_l^{n-2}, \ldots, \binom{n}{m_l - 1} y_l^{n-m_l+1}$$

for $n = 0, 1, 2, \ldots$ form a solution basis for recurrence relation (3.3).

Now consider the inhomogeneous relation (3.1). It is readily verified that the general solution of the inhomogeneous relation is the sum of a *particular solution* of the inhomogeneous relation and the general solution of the homogeneous relation, which is a linear combination of k basis solutions. To find a particular solution one usually tries simple, suitable solutions (see the example below). Once the general solution of (3.1) is known, the coefficients for the k basis solutions are determined from the initial conditions or the boundary conditions.

Example 3.1.6 *(The Fibonacci series)*
Consider recurrence relation (3.2) with an inhomogeneous term,

$$x_{n+2} = x_{n+1} + x_n + a_n, \qquad n = 0, 1, 2, \ldots, \tag{3.6}$$

where a_n is the (mean) number of new pairs of rabbits added to the population in month $n + 2$. Below we determine a particular solution for several choices of a_n.

(i) $a_n = \text{constant} = C$;
We try to find a particular solution which is also constant. Substitution of $x_n = A$ in (3.6) yields $A = -C$.

(ii) $a_n = y^n$;
We now try to find a particular solution of the form $x_n = Ay^n$. Substitution of this form in (3.6) yields $A = 1/(y^2 - y - 1)$. Hence, a complication arises when y is a root of the characteristic equation (3.5) (see (iii) below).

(iii) $a_n = y_1^n$;
Substitution of $x_n = Ay_1^n$ does not work in this case. Instead we try $x_n = Any_1^n$. This yields $A = 1/(2y_1^2 - y_1)$.

3.2 Systems of recurrence relations

The general form of a system of linear recurrence relations with constant coefficients of the first order is, in vector-matrix notation,

$$x_{n+1} = Ax_n + b_n, \qquad n = 0, 1, 2, \ldots \tag{3.7}$$

where

$$x_n = \begin{pmatrix} x_{1,n} \\ x_{2,n} \\ \vdots \\ x_{k,n} \end{pmatrix}, \qquad b_n = \begin{pmatrix} b_{1,n} \\ b_{2,n} \\ \vdots \\ b_{k,n} \end{pmatrix}, \qquad A = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,k} \\ a_{2,1} & a_{2,2} & \ldots & a_{1,k} \\ \vdots & & & \vdots \\ a_{k,1} & a_{k,2} & \ldots & a_{k,k} \end{pmatrix}.$$

The linear recurrence relations of order k treated in the previous section may also be written as a system of recurrence relations of the first order. For instance, consider the second order relation

$$y_{n+2} + a_1 y_{n+1} + a_2 y_n = c_n, \qquad n = 0, 1, 2, \ldots$$

Define the vector x_n by

$$x_n = \begin{pmatrix} y_n \\ y_{n+1} \end{pmatrix}.$$

For the vectors x_n we get the first order relation

$$x_{n+1} = \begin{pmatrix} y_{n+1} \\ y_{n+2} \end{pmatrix} = \begin{pmatrix} y_{n+1} \\ -a_2 y_n - a_1 y_{n+1} \end{pmatrix} + \begin{pmatrix} 0 \\ c_n \end{pmatrix} = Ax_n + b_n$$

where

$$A = \begin{pmatrix} 0 & 1 \\ -a_2 & -a_1 \end{pmatrix}, \qquad b_n = \begin{pmatrix} 0 \\ c_n \end{pmatrix}.$$

In the same way higher order recurrence relations can be translated to a system of recurrence relations of the first order.

Example 3.2.1 *(Markov chain)*
Let us consider a discrete time Markov chain with N states and $N \times N$ transition probability matrix P. A minor difference with the notations introduced above is that it is common to represent the distribution at time n by a *row vector* instead of a column vector. The distribution x_n at time n satisfies

$$x_{n+1} = x_n P, \qquad n = 0, 1, 2, \ldots, \tag{3.8}$$

which is a homogeneous system of recurrence relations of the first order.

Let us consider the homogeneous system

$$x_{n+1} = A x_n, \qquad n = 0, 1, 2, \ldots \tag{3.9}$$

The solutions of the homogeneous system form a vector space and the analogue of lemma 3.1.2 holds.

Lemma 3.2.2 *Let the sequences $\{x_n^{(1)}\}, \ldots, \{x_n^{(l)}\}$ be solutions of (3.9). These l solutions are linearly independent if and only if the vectors $x_0^{(1)}, \ldots, x_0^{(l)}$ are linearly independent.*

This implies the following result.

Theorem 3.2.3 *The set of all solutions of (3.9) forms a k-dimensional vector space over the complex field.*

By iterating relation (3.9) it follows that the general solution of (3.9) is given by

$$x_n = A^n x_0, \qquad n = 0, 1, 2, \ldots$$

So, in fact, solving (3.9) amounts to determining the powers of the matrix A. Below we describe an approach to find these powers. The idea is to express the general solution of (3.9) as a linear combination of k basis solutions. We try to find a solution of (3.9) of the form

$$x_n = \lambda^n v$$

where λ is some number and v a vector. Substitution of this form into (3.9) yields

$$\lambda v = A v.$$

Hence λ is an eigenvalue and v a corresponding eigenvector. Suppose that A has k distinct eigenvalues $\lambda_1, \ldots, \lambda_k$. Since the corresponding eigenvectors $v_1, \ldots, v_k$ are linearly independent (see e.g. [28]) we can conclude that

$$x_n = c_1 \lambda_1^n v_1 + c_2 \lambda_2^n v_2 + \cdots + c_k \lambda_k^n v_k, \qquad n = 0, 1, 2, \ldots \tag{3.10}$$

is the general solution, where $c_1, \ldots, c_k$ are arbitray complex numbers.

Example 3.2.4 *(Markov chain)*
Assume that the probability transition matrix P has N distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$ with corresponding *row* eigenvectors $w_1, w_2, \ldots, w_N$. It is easy to verify that all eigenvalues of a stochastic matrix are less than or equal to 1 in absolute value. If the matrix is irreducible and aperiodic, then there is one eigenvalue equal to 1 and all other eigenvalues are less than 1 in absolute value. Let us assume that $\lambda_1 = 1$. Expressing x_0 as a linear combination of the eigenvectors w_i,

$$x_0 = \sum_{i=1}^{N} c_i w_i, \tag{3.11}$$

we get from (3.8) that

$$x_n = x_{n-1} P = x_0 P^n = \sum_{i=1}^{N} c_i w_i P^n = \sum_{i=1}^{N} c_i \lambda_i^n w_i. \tag{3.12}$$

Recall that $|\lambda_i| < 1$ for all $i > 1$, so x_n converges exponentially fast to $c_1 w_1$ which, apparently, equals the equilibrium distribution. The rate of convergence is determined by the the so-called subdominant eigenvalue, i.e., the second largest eigenvalue in absolute value.

Example 3.2.5 *(2 × 2 Markov chain)*
Consider the 2×2 matrix P,

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

The matrix P has eigenvalues 1 and $1-a-b$ with corresponding row eigenvectors (b, a) and $(1, -1)$, resp. Set $x_0 = (1, 0)$. Expressing x_0 as linear combination of the two eigenvectors,

$$x_0 = \frac{1}{a+b}(b, a) + \frac{a}{a+b}(1, -1).$$

we obtain from (3.12) that

$$x_n = \frac{1}{a+b}(b, a) + (1-a-b)^n \frac{a}{a+b}(1, -1).$$

As we see, there are two special cases, $a = b = 1$ and $a = b = 0$. In all other cases the second eigenvalue $1 - a - b$ is less than 1 in absolute value. The case $a = b = 1$ corresponds to a periodic chain (switching between state 1 and state 2). The case $a = b = 0$ gives a chain with two absorbing states, and hence, which is not irreducible.

In case λ is an eigenvalue of A with multiplicity $m > 1$, but the dimension of the eigenspace of λ is also m (so we can find m independent eigenvectors), then the general solution is again of the form (3.10). A difficulty arises when the number of independent eigenvectors is less than the multiplicity of λ, since then we do not have sufficient solutions.

In that case we call λ *defective*. For a defective eigenvalue λ we try to find extra solutions of the form $\lambda^n(nv+w)$ or even $\lambda^n(n^2v+nw+u)$ (and so on). For instance, suppose that λ is a double eigenvalue which has exactly one independent eigenvector v. Then we try to find a solution of the form

$$x_n = \lambda^n(nw+u),$$

where w and u are vectors to be determined yet. Substituting this form into (3.9) and dividing by common powers of λ leads to

$$\lambda(nw+w+u) = nAw + Au.$$

This equation is satisfied if

$$Aw = \lambda w, \qquad Au = \lambda w + \lambda u.$$

So we can choose $w = v$ and it can be show that the equation for u has a solution.

Now consider the inhomogeneous system (3.7). The general solution of the inhomogeneous system (3.7) is the sum of the general solution of the homogeneous system and a particular solution of the inhomogeneous system. The latter may be found by trying suitable solutions. For example, if the inhomogeneous term is a constant vector, say $b_n = b$, then a particular solution of (3.7) is given by the vector

$$x_n = (I-A)^{-1}b,$$

provided of course, the inverse of $I-A$ exists.

3.3 Systems of linear differential equations

The theory of systems of linear differential equations is very similar to the theory of systems of linear recurrence relations. Therefore it suffices to present a brief description of the results for the differential equations.

The general form of a system of linear differential equations with constant coefficients of the first order is, in vector-matrix notation,

$$x'(t) = Ax(t) + b(t), \qquad t \in I, \tag{3.13}$$

where I is some real interval and

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_k(t) \end{pmatrix}, \qquad b(t) = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_k(t) \end{pmatrix}, \qquad A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ a_{2,1} & a_{2,2} & \dots & a_{1,k} \\ \vdots & & & \vdots \\ a_{k,1} & a_{k,2} & \dots & a_{k,k} \end{pmatrix}.$$

We assume that $b(t)$ is a continuous vector on the interval I.

Let us first consider the homogeneous system

$$x'(t) = Ax(t), \qquad t \in I. \tag{3.14}$$

Theorem 3.3.1 *The set of all solutions of (3.14) on I forms a k-dimensional vector space over the complex field.*

If $x^{(1)}(t), \ldots, x^{(k)}(t)$ are a set of k independent solutions of (3.14), they are said to form a basis of solutions of (3.14). The general solution can be expressed as a linear combination of the basis solutions,

$$x(t) = c_1 x^{(1)}(t) + c_2 x^{(2)}(t) + \cdots + c_k x^{(k)}(t), \qquad t \in I,$$

where $c_1, \ldots, c_k$ are arbitrary complex numbers. To obtain k independent solutions of (3.14) we try to find solutions of the form

$$x(t) = e^{\lambda t} v$$

where λ is some number and v a vector. Substitution of this form into (3.14) yields, after dividing by $e^{\lambda t}$,

$$\lambda v = Av.$$

Hence λ is an eigenvalue and v a corresponding eigenvector. If A has k distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, with associated eigenvectors $v_1, \ldots, v_k$, the solutions

$$e^{\lambda_1 t} v_1, e^{\lambda_2 t} v_2, \ldots, e^{\lambda_k t} v_k$$

for $t \in I$ form a solution basis and so we can conclude that

$$x(t) = c_1 e^{\lambda_1 t} v_1 + c_2 e^{\lambda_2 t} v_2 + \cdots + c_k e^{\lambda_k t} v_k, \qquad t \in I, \tag{3.15}$$

is the general solution, where $c_1, \ldots, c_k$ are arbitray complex numbers. If λ is an eigenvalue of A with multiplicity m and the number of eigenvectors for λ is less than m, we try to find extra solutions of the form $e^{\lambda t}(t^l w_l + t^{l-1} w_{l-1} + \ldots + t w_1 + w_0)$.

Finally, by adding to the solution (3.15) a particular solution of the inhomogeneous system (3.13), we obtain the general solution of the inhomogeneous system.

3.4 Exercises

Exercise 3.4.1 *(Roulette game)*
Consider someone playing roulette in a casino. The player is very cautious and only plays pair and impair. He plays a dollar at a time. The probability p that he wins is equal to 18/37. The probability q that he looses is 19/37. So with probability p he wins a dollar, with probability q he looses a dollar. The player starts with N dollar and stops as soon as he has $2N$ dollar, or more likely, he has lost his N dollar. Denote the probability that he can go home with $2N$ dollar if now he has n dollar by x_n.

(i) Derive the following relation between the probabilities x_n,

$$x_n = px_{n+1} + qx_{n-1}, \tag{3.16}$$

with boundary conditions $x_0 = 0$ and $x_{2N} = 1$.

(ii) Solve this recurrence relation.

(iii) Compute the probability of leaving with $2N$ dollar for $N = 64$ dollar for a player playing 1, 2, 4, 8, 16, 32 and 64 dollar at a time. What is your conclusion?

Exercise 3.4.2 *(Loan)*
You are going to buy a car, but you do not have the money. So you need a loan for $K = 10000$ dollars. You pay $a = 350$ dollar a month during $N = 36$ months. What is the interest rate? Denote the interest per month by r. Further define x_n to be the remaining debt after n months.

(i) Derive the following recurrence relation for x_n,

$$x_{n+1} = (1 + r)x_n - a, \qquad n = 0, 1, ..., N - 1,$$

with boundary values $x_0 = K$ and $x_N = 0$.

(ii) Compute the interest rate r.

Exercise 3.4.3
Prove lemma 3.1.2 and the theorems 3.1.3 and 3.1.5.

Exercise 3.4.4 *(Machine with limited buffer)*
Let us consider a machine where jobs are processed one at a time and in order of arrival. The interarrival times of the jobs are exponentially distributed with mean $1/\lambda$. The processing times are exponentially distributed with mean $1/\mu$. Jobs arriving while the machine is busy wait for their turn in a buffer, if there is room, and otherwise, if the buffer is full, they are processed somewhere else. The buffer has space for one job only.
We want to find the distribution of the number of jobs at the machine (the one being

processed and the number of jobs waiting in the buffer) at time t. Due to the memoryless property of the exponential distribution (see property 2.1) the state of the system at time t can be described by the number of jobs in the system. Let $p_i(t)$ be the probability that at time t there are i jobs in the system, $i = 0, 1, 2$.

(i) Show that, when there is a job being processed, the time that elapses till something happens (an arrival or job completion) starting from time t is exponentially distributed with mean $1/(\lambda + \mu)$.

(ii) Derive, by using property (2.1), the following differential equations for $p_i(t)$, $i = 0, 1, 2$,

$$\begin{aligned}
p_0'(t) &= -\lambda p_0(t) + \mu p_1(t), \\
p_1'(t) &= \lambda p_0(t) - (\lambda + \mu) p_1(t) + \mu p_2(t), \\
p_2'(t) &= \lambda p_1(t) - \mu p_2(t).
\end{aligned}$$

(iii) Solve the differential equations in (ii) with initial condition

$$(p_0(0), p_1(0), p_2(0)) = (1, 0, 0),$$

i.e. at time 0 the system is empty.

(iv) Determine the limiting probabilities $p_i = \lim_{t\to\infty} p_i(t)$, $i = 0, 1, 2$.

(v) What is in the long run the occupation rate of the machine and the fraction of jobs that is processed somewhere else?

Chapter 4

Single-server models

In this chapter we will treat some elementary queueing models. We start to analyze models with exponential interarrival times and exponential service times. The exponential distribution allows for a very simple state description, namely the number of customers in the system. We do not have to remember when the last customer arrived nor do we have to register when the customers went into service. Since the exponential distribution is memoryless (see 2.1), this information does not yield a better prediction of the future.

Next we relax the assumption of exponential service times to phase-type service times. The use of phase-type distributions still allows a simple description of the system and the class is sufficiently rich for the modelling of random variables in many practical applications. Then we also drop the assumption of phase-type distributions and consider the case of generally distributed service times. Finally we will look at the model with non-exponential arrivals and phase-type service times.

4.1 $M/M/1/N$ queue

We first consider the $M/M/1$ queue with $N-1$ waiting positions in its queue. The service times are exponential with mean $1/\mu$ and customers arrive according to a Poisson stream with rate λ. Customers, finding on arrival all waiting positions occupied, immediately depart (and are possibly served somewhere else.)

In the following subsection we will first study the time-dependent behavior of the system. Next we consider the limiting behavior.

4.1.1 Time-dependent behavior

The state of the system at time t can be described by the number of customers in the system (i.e. the customers in the queue and the one being served). Let $p_n(t)$ be the probability that at time t there are n customers in the system, $n = 0, 1, \ldots, N$. Based on property (2.1) we get, for $\Delta t \to 0$,

$$p_0(t + \Delta t) = (1 - \lambda\Delta t)p_0(t) + \mu\Delta t p_1(t) + o(\Delta t),$$

$$
\begin{aligned}
p_n(t+\Delta t) &= (1-\lambda\Delta t)p_{n-1}(t) - (\lambda+\mu)\Delta t p_n(t) + \mu\Delta p_{n+1}(t) + o(\Delta t),\\
p_N(t+\Delta t) &= \lambda\Delta t p_{N-1}(t) + (1-\mu\Delta t)p_N(t) + o(\Delta t).
\end{aligned}
$$

Hence, by letting $\Delta t \to 0$, we obtain the following set of differential equations for the probabilities $p_n(t)$.

$$
\begin{aligned}
p_0'(t) &= -\lambda p_0(t) + \mu p_1(t),\\
p_n'(t) &= \lambda p_{n-1}(t) - (\lambda+\mu)p_n(t) + \mu p_{n+1}(t), \qquad n = 1,\ldots,N-1,\\
p_N'(t) &= \lambda p_{N-1}(t) - \mu p_N(t),
\end{aligned}
\tag{4.1}
$$

or in vector-matrix notation,

$$p'(t) = p(t)Q, \tag{4.2}$$

where

$$p(t) = (p_0(t),\ldots,p_N(t)),$$

$$
Q = \begin{pmatrix}
-\lambda & \lambda & 0 & 0 & \cdots & 0 & 0 & 0\\
\mu & -(\lambda+\mu) & \lambda & 0 & \cdots & 0 & 0 & 0\\
0 & \mu & -(\lambda+\mu) & \lambda & \cdots & 0 & 0 & 0\\
\vdots & & & & & & & \vdots\\
0 & 0 & 0 & 0 & \cdots & \mu & -(\lambda+\mu) & \lambda\\
0 & 0 & 0 & 0 & \cdots & 0 & \mu & -\mu
\end{pmatrix}. \tag{4.3}
$$

At time 0 we assume that the system is empty. So the initial condition for (4.2) is

$$p(0) = (1, 0, \ldots, 0). \tag{4.4}$$

Equation (4.2) is a system of linear differential equations with initial-value condition (4.4). We have seen in section (3.3) that the general solution of (4.2) can be expressed in terms of the eigenvalues and eigenvectors of Q. The following lemma gives us information on the eigenvalues of Q.

Lemma 4.1.1 *The eigenvalues of Q are all real and simple; one eigenvalue is 0 and all other eigenvalues are negative and greater than or equal to $-2(\lambda+\mu)$.*

Denote these eigenvalues by $\xi_1,\ldots,\xi_N$ with $\xi_1 = 0$. We remark from a computational point of view that the eigenvalues can be easily found. They are characterized as the N roots in the interval $[-2(\lambda+\mu), 0]$ of characteristic equation

$$\mathrm{Det}(Q - \xi I) = 0,$$

where I is the $N \times N$ identity matrix. So the roots may be found by, e.g., simple bisection. Denote the associated row eigenvectors by $y_1, \ldots, y_N$. For given eigenvalue ξ_j the eigenvector y_j is a non-zero solution of the second order recurrence relation

$$\begin{aligned} -\lambda y_{j,0} + \mu y_{j,1} &= \xi_j y_{j,0}, \\ \lambda y_{j,n-1} - (\lambda + \mu) y_{j,n} + \mu y_{j,n+1} &= \xi_j y_{j,n}, \qquad n = 1, \ldots, N-1, \\ \lambda y_{j,N-1} - \mu y_{j,N} &= \xi_j y_{j,N}. \end{aligned}$$

If we set $y_{j,0} = 1$, then the eigenvector y_j is given by

$$y_{j,n} = a_j x_{j,1}^n + (1 - a_j) x_{j,2}^n, \qquad n = 0, 1, \ldots, N,$$

where $x_{j,1}$ and $x_{j,2}$ are the two roots of the quadratic equation

$$(\lambda + \mu + \xi_j) x = \lambda + \mu x^2,$$

and

$$a_j = \frac{x_{j,2} - (\lambda + \xi_j)/\mu}{x_{j,2} - x_{j,1}}.$$

Note that the eigenvector for $\xi_1 = 0$ simplifies to

$$y_{1,n} = \left(\frac{\lambda}{\mu}\right)^n, \qquad n = 1, 2, \ldots, N.$$

Once the eigenvalues and eigenvectors are known, the general solution of (4.2) can be written as

$$p(t) = c_1 y_1 + c_2 y_2 e^{\xi_2 t} + \ldots c_N y_N e^{\xi_N t}, \qquad t > 0. \tag{4.5}$$

To complete the solution the coefficients $c_1, \ldots, c_N$ have to be determined from the initial condition (4.4).

4.1.2 Limiting behavior

We now investigate what happens as t tends to infinity. From (4.5) and lemma (4.1.1) we directly obtain, by letting t tend to infinity,

$$p_n = \lim_{t \to \infty} p_n(t) = c_1 y_1 = \frac{(1-\rho)\rho^n}{1 - \rho^{N+1}}, \qquad n = 1, 2, \ldots, N,$$

where $\rho = \lambda/\mu$. Note that ρ is not the server utilization, since part of the work is done somewhere else. From this distribution we find for the mean number of customers in the system

$$E(L) = \sum_{n=0}^{N} n p_n = \frac{\rho(1 + N\rho^{N+1} - (N+1)\rho^N)}{(1-\rho)(1-\rho^{N+1})}.$$

By using Little's law we obtain for the mean time spent in the system by a customer that is served (and not sent away)

$$E(S) = \frac{E(L)}{(1 - p_0)\mu}.$$

4.2 $M/M/1$ queue

In this section we consider the case where the number of waiting positions is not limited. Let, as in the previous section, λ denote the arrival rate and μ the service rate. We have to require

$$\rho = \frac{\lambda}{\mu} < 1,$$

since, otherwise, the amount of work in the system will explode. The ρ is the utilization of the server. In the next subsection we start to look at the time-dependent behavior of the system.

4.2.1 Time-dependent behavior

Let $p_n(t)$ denote again the probability that at time t there are n customers in the system, $n = 0, 1, 2, \ldots$ They satisfy the following infinite set of differential equations.

$$\begin{aligned} p_0'(t) &= -\lambda p_0(t) + \mu p_1(t), \\ p_n'(t) &= \lambda p_{n-1}(t) - (\lambda + \mu) p_n(t) + \mu p_{n+1}(t), \qquad n = 1, 2, \ldots, \end{aligned} \tag{4.6}$$

or in vector-matrix notation,

$$p'(t) = p(t)Q,$$

where

$$p(t) = (p_0(t), p_1(t), p_2(t), \ldots),$$

and Q is a countably infinite matrix given by

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & & & & \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & & & \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & & \\ \vdots & 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & \\ & \vdots & & \vdots & & \vdots & & \end{pmatrix}.$$

We now have an *infinite system of differential equations*, and therefore we cannot apply the theory of chapter 3. An attempt to analyze this infinite system is to let N tend to infinity in solution (4.5) for the finite queue model. However, this limit is very tricky and complicated. We may end up with an infinite set of values ξ for which $Q - \xi I$ is not invertible. This set can be countably infinite, or even continuous, and the resulting solution may be expressed as an infinite sum, or an integral. An explicit solution for the probabilities $p_n(t)$ can be found in [27]. The expression presented there (see p. 77) is an infinite sum of modified Bessel functions. So already one of the simplest interesting queueing models leads to an ugly expression for the time-dependent behavior of its state probabilities. For more general systems we can only expect more complexity. Therefore, in the remainder we will focus on the *limiting or equilibrium behavior* of queueing systems, which appears to be easier to analyse.

4.2.2 Limiting behavior

One may show that as $t \to \infty$, then $p_n'(t) \to 0$ and $p_n(t) \to p_n$. So there exists a limiting distribution. Hence, from (4.6) it follows that the limiting (or equilibrium) probabilities p_n satisfy the recurrence relations

$$
\begin{aligned}
0 &= -\lambda p_0 + \mu p_1, && (4.7)\\
0 &= \lambda p_{n-1} - (\lambda+\mu)p_n + \mu p_{n+1}, \qquad n = 1, 2, \ldots && (4.8)
\end{aligned}
$$

It is also possible to derive these equations directly from a so-called *flow diagram*, as shown in figure 4.1 (see also the preceding course S1).

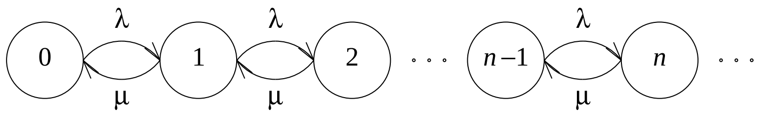

Figure 4.1: Flow diagram for the $M/M/1$ model

The arrows indicate possible transitions. The rate at which a transition occurs is λ for a transition from n to $n+1$ (an arrival) and μ for a transition from $n+1$ to n (a departure). The number of transitions per unit time from n to $n+1$, which is also called the *flow* from n to $n+1$, is equal to p_n, the fraction of time the system is in state n, times λ, the rate at arrivals occur while the system is in state n. The equilibrium equations (4.7) and (4.8) follow by equating the flow out of state n and the flow into state n.

The equations (4.8) are a second order recurrence relation with constant coefficients. Its general solution is of the form

$$p_n = c_1 x_1^n + c_2 x_2^n, \qquad n = 0, 1, 2, \ldots \tag{4.9}$$

where x_1 and x_2 are roots of the quadratic equation

$$\lambda - (\lambda + \mu)x + \mu x^2 = 0.$$

This equation has two zeros, namely $x = 1$ and $x = \lambda/\mu = \rho$. So all solutions to (4.8) are of the form

$$p_n = c_1 + c_2 \rho^n, \qquad n = 0, 1, 2, \ldots$$

There is in fact, besides (4.7), a second boundary condition, namely the condition that the sum of all probabilities is equal to 1. This condition of course directly implies that c_1 must be equal to 0. That c_1 must be equal to 0 also follows from (4.7) by substituting the solution (4.9) into (4.7).

The coefficient c_2 finally follows from the normalization equation, stating that the sum of all probabilities is equal to 1. This yield $c_2 = 1 - \rho$. So

$$p_n = (1 - \rho)\rho^n, \qquad n = 0, 1, 2, \ldots \tag{4.10}$$

Remark 4.2.1 *(alternative solution approaches)*
For this simple model there are many ways to determine the solution of the equations (4.8) and (4.7). We mention the following ones.

(i) *Recursion.* One can use (4.7) to express p_1 in p_0 yielding

$$p_1 = \rho p_1.$$

Substitution of this relation into (4.8) for $n = 1$ gives

$$p_2 = \rho^2 p_0.$$

By substituting the relations above into (4.8) for $n = 2$ we obtain p_3, and so on. Hence we can recursively express all probabilities in terms of p_0, yielding

$$p_n = \rho^n p_0, \qquad n = 0, 1, 2, \ldots.$$

The probability p_0 finally follows from normalization.

(ii) *Generating function.* Define the probability generating function

$$P(z) = \sum_{n=0}^{\infty} p_n z^n, \tag{4.11}$$

which is properly defined for z with $|z| \leq 1$. By multiplying the nth equilibrium equation with z^n and then summing the equations over all n, the equilibrium equations for p_n can be transformed to the following single equation for $P(z)$,

$$0 = \mu p_0(1 - z^{-1}) + (\lambda z + \mu z^{-1} - (\lambda + \mu))P(z).$$

The solution of this equation is

$$P(z) = \frac{p_0}{\rho z - 1} = \frac{\rho - 1}{\rho z - 1} = \sum_{n=0}^{\infty} (1 - \rho)\rho^n z^n, \tag{4.12}$$

where we used that $P(1) = 1$. Hence, by equating the coefficients of z^n in (4.11) and (4.12) we retrieve the solution (4.10). For more details on the use of generating functions we refer to part b of this course.

(iii) *Global balance principle.* This principle states that for *each set of states A*, the flow out of set A is equal to the flow into that set. In fact, the equilibrium equations (4.7)–(4.8) follow by applying this principle to a single state. But if we apply the balance principle to the set $A = \{0, 1, \ldots, n - 1\}$ we get the simple relation

$$\lambda p_{n-1} = \mu p_n, \qquad n = 1, 2, \ldots$$

Repeated application of this relation yields

$$p_n = \rho^n p_0, \qquad n = 0, 1, 2, \ldots$$

so that, after normalization, the solution (4.10) follows.

From the equilibrium probabilities we can derive expressions for the mean number of customers in the system and the mean time spent in the system. For the first one we get

$$E(L) = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\rho},$$

and by applying Little's law,

$$E(S) = \frac{1/\mu}{1-\rho}. \tag{4.13}$$

The mean waiting time follows from $E(S)$ by subtracting the mean processing time, so

$$E(W) = E(S) - 1/\mu = \frac{\rho/\mu}{1-\rho}.$$

It is also possible to derive the distribution of the waiting time. Denote by N_a the number of customers in the system just before the arrival of a customer and let X_m be the service time of the mth customer in the queue. Then we have

$$P(W > t) = P(\sum_{m=1}^{N_a} X_m > t).$$

By conditioning on N_a and using that N_a and X_m are independent it follows that

$$P(W > t) = \sum_{n=0}^{\infty} P(\sum_{m=1}^{n} X_m > t) P(N_a = n). \tag{4.14}$$

The problem is to find the probability that an arriving customer finds n customers in the system. The average number of customers per unit time finding on arrival n customers in the system is equal to λp_n. Dividing this number by the average number of customers arriving per unit time gives the desired fraction. So

$$P(N_a = n) = \frac{\lambda p_n}{\lambda} = p_n. \tag{4.15}$$

In words, the fraction of customers finding on arrival n customers in the system is equal to the fraction of time there are n customers in the system. This is known as the PASTA property (see section 2.7). Substituting (4.15) in (4.14) and using that $\sum_{m=1}^{n} X_m$ is Erlang distributed leads after some calculations to (cf. exercise 2.12.3)

$$P(W > t) = \rho e^{-\mu(1-\rho)t}, \qquad t > 0. \tag{4.16}$$

The probability that one does not have to wait is equal to $p_0 = 1 - \rho$. So the *conditional waiting time* which is the waiting time given that it is greater than zero is exponentially distributed with parameter $\mu(1-\rho)$.

In fact, the means $E(L)$ and $E(S)$ can also be directly determined, i.e. without knowing the probabilities $p(n)$, by combining Little's law and the PASTA property. Based on PASTA we know that the average number of customers in the system seen an arriving customer equals $E(L)$ and each of them (also the one in service) has a (residual) service time with mean $1/\mu$. The customer further has to wait for its own service time. Hence

$$E(S) = E(L)\frac{1}{\mu} + \frac{1}{\mu}$$

and together with

$$E(L) = \lambda E(S)$$

we find expression (4.13). As we mentioned earlier, this approach is called the mean value technique.

If we look at the expressions for $E(L)$ and $E(S)$ we see that both quantities grow to infinity as ρ approaches unity. The dramatic behavior is caused by the variation in the arrival and service process. This type of behavior with respect to ρ is characteristic for almost every queueing system. In table 4.1 we list for increasing values of ρ the mean mean waiting time and some waiting time probabilities.

We can conclude that the server utilization and the variation in the arrival and service process are *essential* to the behavior of the system.

ρ	$E(W)$		$P(W>t)$		
		t	5	10	20
0.5	1		0.04	0.00	0.00
0.8	4		0.29	0.11	0.02
0.9	9		0.55	0.33	0.12
0.95	19		0.74	0.58	0.35

Table 4.1: Performance characteristics for the $M/M/1$ with mean service time 1

4.3 $M/E_r/1$ queue

In the preceding we evaluated a system with exponential service times. In this section we show that other types of variation in the service times may be evaluated with the same basic approach if we use phase-type distributions for the service times (see section 2.2.4). We demonstrate the approach for Erlang service times. Other phase-type service times can be analyzed along the same lines (see exercises 4.6.11 and 4.6.12).

We consider a single server. Customers arrive according to a Poisson stream with rate λ and are treated in order of arrival. The service times are Erlang-r distributed with mean r/μ (see section 2.2.2). So the service times can be thought of as consisting of r independent exponential phases with mean $1/\mu$. The Erlang distribution can be used to model service times with a low coefficient of variation (less than one), but it can also arise naturally. For instance, if a job has to pass, stage by stage, through a series of r independent production stages, where each stage takes a exponentially distributed time. For stability we have to require

$$\rho = \frac{\lambda r}{\mu} < 1. \tag{4.17}$$

There are two ways in which one can describe the state of the system: a natural way, giving a two-dimensional description, and another way that is found in most textbooks and leads to a one-dimensional state.

4.3.1 Natural description

The natural way to describe the state of a nonempty system is by the pair (n, m) where n denotes the number of customer in the system and m the remaining number of service phases of the customer in service. The empty state is denoted by 0. This leads to the flow diagram in figure 4.2.

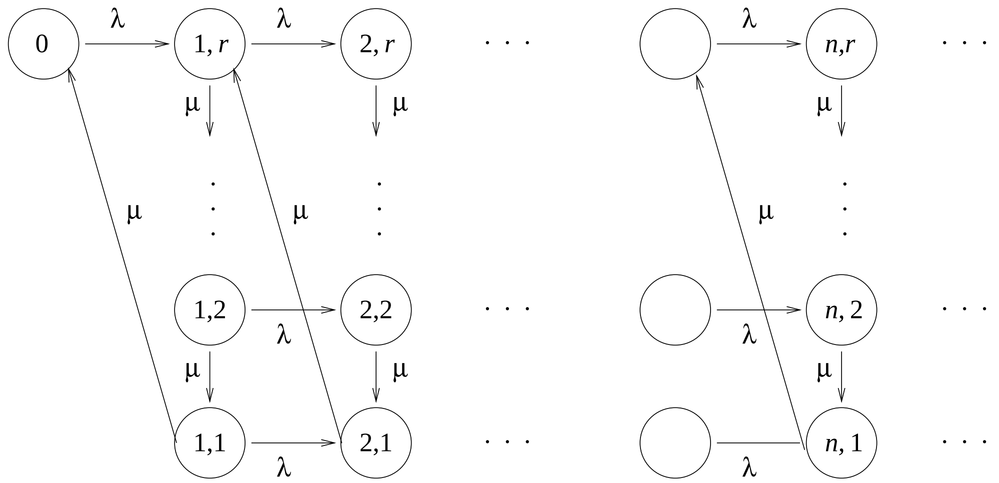

Figure 4.2: Two dimensional flow diagram for the $M/E_r/1$ model

Let $p(n, m)$ denote the probability that the system is in state (n, m). Then we can write down the following balance equations for $n > 1$,

$$\begin{aligned} p(n,m)(\lambda+\mu) &= p(n-1,m)\lambda + p(n,m+1)\mu, \qquad m < r, && (4.18)\\ p(n,r)(\lambda+\mu) &= p(n-1,r)\lambda + p(n+1,1)\mu, && (4.19) \end{aligned}$$

and for $n \leq 1$ we have

$$\begin{aligned}
p(1,m)(\lambda+\mu) &= p(1,m+1)\mu, \qquad m<r, &(4.20)\\
p(1,r)(\lambda+\mu) &= p(0)\lambda+p(2,1)\mu, &(4.21)\\
p(0)\lambda &= p(1,1)\mu. &(4.22)
\end{aligned}$$

We will discuss two approaches to solves these equations. The first one is the matrix-geometric approach, the other uses basis solutions.

4.3.2 Matrix-geometric approach

The structure of the set of equilibrium equations (4.18)–(4.19) is closely related to the similar set of equations (4.8) for the $M/M/1$. This becomes even more striking if we replace $p(n,m)$ by the vector

$$\overline{p}_n = (p(n,1), p(n,2), \cdots, p(n,r)).$$

Then, in vector-matrix notation, the equations (4.18)–(4.19) are

$$\overline{0} = \overline{p}_{n-1}A_0 + \overline{p}_n A_1 + \overline{p}_{n+1}A_2, \qquad n>1, \tag{4.23}$$

where the $r \times r$ matrices A_0, A_1 and A_2 satisfy

$$A_0 = \begin{pmatrix} \lambda & 0 & \cdots & & 0 \\ 0 & \lambda & & & \\ \vdots & & \ddots & & \vdots \\ & & & \lambda & 0 \\ 0 & & \cdots & 0 & \lambda \end{pmatrix},$$

$$A_1 = \begin{pmatrix} -(\lambda+\mu) & 0 & \cdots & & 0 \\ \mu & -(\lambda+\mu) & & & \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & & -(\lambda+\mu) & 0 \\ 0 & \vdots & 0 & \mu & -(\lambda+\mu) \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 0 & \cdots & 0 & \mu \\ \vdots & & & 0 \\ & & & \vdots \\ 0 & \cdots & & 0 \end{pmatrix}.$$

The equations (4.23) can be simplified by eliminating $\overline{p}_{n+1}$. This can be achieved by applying the global balance principle to flows between levels, where *level* n is defined as the set of states $\{(n,1),\ldots,(n,r)\}$. By equating the flow between the levels n and $n+1$ with $n \geq 1$ one obtains

$$p(n+1,1)\mu = (p(n,1)+\cdots+p(n,r))\lambda,$$

or

$$\overline{p}_{n+1}A_2 = \overline{p}_n A_3, \tag{4.24}$$

with

$$A_3 = \begin{pmatrix} 0 & \cdots & 0 & \lambda \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \lambda \end{pmatrix}.$$

Hence, substitution of (4.24) into (4.23) yields

$$\overline{0} = \overline{p}_{n-1}A_0 + \overline{p}_n(A_1 + A_3),$$

so

$$\overline{p}_n = \overline{p}_{n-1}R, \qquad n = 2,3,\ldots, \tag{4.25}$$

with

$$R = -A_0(A_1+A_3)^{-1}.$$

Note that the inverse of $A_1 + A_3$ exists, since this is a transient generator. From (4.25) we obtain the *matrix-geometric solution*

$$\overline{p}_n = \overline{p}_1 R^{n-1}, \qquad n = 2,3,\ldots$$

The matrix R is the so-called *rate matrix*. The vector $\overline{p}_1$ can be expressed in $p(0)$ by substituting relation (4.24) into the equations (4.20)–(4.21) for $n=1$, yielding

$$\overline{p}_1 = (0,\ldots,0,\lambda p(0))(A_1+A_3)^{-1}.$$

The probability $p(0)$ finally follows from the normalization equation.

More general the matrix-geometric approach works for Markov processes with the same type of transition structure as in figure 4.2. Such Markov processes are called *quasi-birth-death processes* (QBDs). More precisely, a QBD is a Markov process with state space $\{(n,m); n = 0,1,2,\ldots, m = 1,\ldots,r\}$, which, when the states are ordered as $(0,1),\ldots,(0,r),(1,1),\ldots,(1,r),(2,1),\ldots$, has a generator Q of the following form

$$Q = \begin{pmatrix} B_{00} & B_{01} & 0 & 0 & \cdots \\ B_{10} & A_1 & A_0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix},$$

where the matrices B_{00}, B_{01}, B_{10}, A_0, A_1 and A_2 have dimensions $r \times r$. The structure of Q implies that in a single transition the process can move only to the next higher or lower level (as we have also seen for the $M/M/1$ queue), which explains the name birth and death process. Provided the equilibrium probabilities $p(n,m)$ of the QBD exist, it can be shown that the equilibrium probability vectors $\overline{p}_n$ are given by the matrix-geometric form

$$\overline{p}_n = \overline{p}_1 R^{n-1}, \qquad n = 1, 2, \ldots$$

where the rate matrix R is the minimal nonnegative solution of the matrix quadratic equation

$$0 = A_0 + RA_1 + R^2 A_2. \tag{4.26}$$

The rate matrix R has the following probabilistic interpretation. The component $R(i,j)$ is the expected number of visits to the state $(n+1, j)$ before the first return to level n, given that the process starts in state (n, i).

The boundary probability vectors $\overline{p}_0$ and $\overline{p}_1$ follow from the balance equations at the levels 0 and 1 and the normalization equation,

$$\begin{aligned} \overline{p}_0 B_{00} + \overline{p}_1 B_{10} &= 0, \\ \overline{p}_0 B_{01} + \overline{p}_1 (A_1 + RA_2) &= 0, \\ \overline{p}_0 \overline{e} + \overline{p}_1 (I - R)^{-1} \overline{e} &= 1, \end{aligned}$$

where $\overline{e}$ is the column vector with ones and I is the identity matrix. The same results hold for QBSs with a bit more complex boundary behavior, i.e. for which the number of states in level 0 (the boundary states) is less or greater than r. For an excellent and extensive study of the matrix-geometric theory the reader is referred to [34].

An essential step in numerical computation of the matrix-geometric solution is the evaluation of the rate matrix R. The standard way to compute R is by successive substitutions, based on an alternative form of the matrix quadratic equation (4.26), namely

$$R = \Delta A_0 + R(I + \Delta A_1) + R^2 \Delta A_2,$$

where $\Delta \geq 0$ is chosen sufficiently small such that $I + \Delta(A_0 + A_1 + A_2) \geq 0$. Then the matrices R_n generated by the iteration scheme

$$\begin{aligned} R_0 &= 0, \\ R_{n+1} &= \Delta A_0 + R_n(I + \Delta A_1) + R_n^2 \Delta A_2, \qquad n = 0, 1, \ldots, \end{aligned}$$

monotonically increase to R as n tends to infinity. More sophisticated and efficient algorithms for the determination of R for QBDs can be found in [29].

Somtimes it is possible to explicitly determine R. For example this is the case for the $M/E_r/1$ queue. Two general cases for which the rate matrix R can be determined explicitly are described in [35].

4.3.3 Basis solutions

We first try to find solutions of (4.18)–(4.19) of the type

$$p(n,m) = y^n v(m), \qquad n = 0,1,2,\dots, \quad m = 1,\dots,r. \tag{4.27}$$

Then we linearly combine these solutions to also satisfy the boundary equations (4.20)–(4.22). If we substitute the form (4.27) into equation (4.18) then we get

$$(\lambda+\mu)yv(k) = \lambda v(k) + \mu y v(k+1),$$

from which follows

$$v(k+1) = \frac{(\lambda+\mu)y-\lambda}{\mu y} v(k), \qquad k = 1,2,\dots,r-1.$$

Hence

$$v(k) = x^{k-1}v(1), \qquad k = 1,\dots,r, \tag{4.28}$$

where we introduced

$$x = \frac{(\lambda+\mu)y-\lambda}{\mu y}. \tag{4.29}$$

From equation (4.19) we get

$$(\lambda+\mu)yv(r) = \lambda v(r) + \mu y^2 v(1),$$

So, substitution of $v(r) = x^{r-1}v(1)$ yields

$$(\lambda+\mu)yx^{r-1} = \lambda x^{r-1} + \mu y^2,$$

or

$$[(\lambda+\mu)y-\lambda]x^{r-1} = \mu y^2.$$

From this we see that

$$x^r = y. \tag{4.30}$$

Hence, inserting (4.30) into (4.29) we obtain

$$(\lambda+\mu)x^r = \lambda + \mu x^{r+1}. \tag{4.31}$$

Clearly one of the solutions for x is $x = 1$ which implies $y = 1$ and thus is not a useful solution. Using Rouché's theorem and stability condition (4.17) one may show that the other r roots $x_1,\dots,x_r$ say are all distinct and inside the unit circle (see exercise 4.6.7). Of course, then the corresponding values for $y_1, ..., y_r$ are also inside the unit circle. From this we can conclude that the overall solution for the probabilities $p(n,m)$ satisfies

$$p(n,m) = \sum_{i=1}^{r} A_i v_i(m) y_i^n,$$

which by (4.28) and (4.30) simplifies to

$$p(n,m) = \sum_{i=1}^{r} A_i x_i^{nr+k}, \qquad n = 0,1,2,\ldots, \quad m = 1,\ldots,r. \tag{4.32}$$

The unknown coefficients A_i and $p(0)$ can be determined from the r boundary conditions (4.20)–(4.22) together with the normalization equation.

Remark 4.3.1 *(relation with the matrix-geometric solution)*
The values y_i are the eigenvalues of the rate matrix R, and the associated row-eigenvector for eigenvalue y_i is the vector $(v_i(1), v_i(2), \ldots, v_i(r))$, $i = 1,\ldots,r$.

4.3.4 Phase description

The second way to describe the state is by counting the number of phases of work in the system. The number $nr + k$ that we met in the preceding section is just that number minus n (for the job in service we have k phases of work instead of n). There is a one to one correspondence between the number of phases in the system and the pair (n,k). The number of phases in the system is equal to the number of jobs waiting in the queue multiplied by r plus the number of not yet completed phases of the job in service.

The analysis for the one-dimensional system is of course closely related to the one in the previous section. For the one-dimensional description we get the flow diagram of figure 4.3.

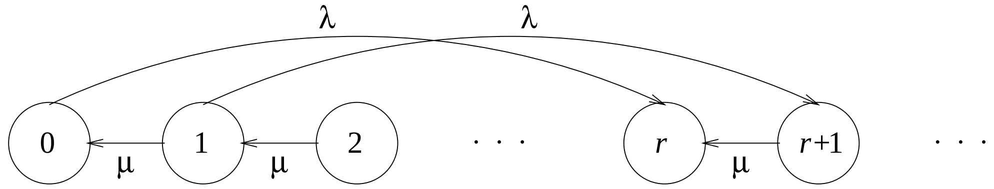

Figure 4.3: One-dimensional flow diagram for the $M/E_r/1$ model

Let p_n be the equilibrium probability of n phases work in the system. From the flow diagram we obtain the following set of balance equations for p_n.

$$p_0\lambda = p_1\mu, \tag{4.33}$$
$$p_n(\lambda+\mu) = p_{n-r}\lambda + p_{n+1}\mu, \qquad n = 1,2,\ldots \tag{4.34}$$

where by convention

$$p_{-1} = p_{-2} = \cdots = p_{-r+1} = 0. \tag{4.35}$$

The balance equations form a recurrence relation of order $r+1$. So its general solution is

$$p_n = \sum_{k=1}^{r+1} B_k x_k^n\,, \qquad n = 0,1,2,\ldots$$

where the factors x_k are the roots of the equation

$$(\lambda + \mu)x^r = \lambda + \mu x^{r+1},$$

provided this equation has $r+1$ distinct roots. Note that this equation is the same as (4.31). So there are indeed $r+1$ distinct roots. One of them equals 1, say $x_{r+1} = 1$, and the others are all inside the unit circle.

Since the probabilities p_n have to add up to one, we have to set $B_{r+1} = 0$. It is convenient to rewrite p_n in the form of a sum of r geometric distributions, yielding

$$p_n = \sum_{k=1}^{r} C_k(1-x_k)x_k^n\,, \qquad n = 0, 1, 2, \ldots \tag{4.36}$$

The coefficients C_k can be determined from convention (4.35) and the normalization condition. Note that since the balance equations are dependent, equation (4.33) may be omitted. Substitution of (4.36) into (4.35) directly leads to the following set of equations for the coefficients C_k.

$$C_1(1-x_1)y_1^n + \cdots + C_r(1-x_r)y_r^n = 0\,, \qquad n = 1, \ldots, r-1\,,$$

where $y_k = 1/x_k$. These equations are of *VanderMonde-type* and therefore can be solved explicitly using Cramer's rule (see exercise 4.6.9). Then we get

$$C_k = \frac{C}{\prod_{j=1}^{r}(1-y_j)} \frac{\prod_{j\neq k}(1-y_j)}{\prod_{j\neq k}(y_k - y_j)}\,, \qquad k = 1, \ldots, r\,,$$

for some constant C. This constant can be determined from the normalization condition, which leads to (see exercise 4.6.10)

$$C = \prod_{j=1}^{r}(1-y_j). \tag{4.37}$$

Hence the coefficients C_k are given by

$$C_k = \frac{\prod_{j\neq k}(1-y_j)}{\prod_{j\neq k}(y_k - y_j)}\,, \qquad k = 1, \ldots, r.$$

This completes the determination of the equilibrium probabilities p_n. Of course, it is easily verified that expression (4.32) found for the probabilities $p(n,k)$ is equivalent to the one for the probabilities p_n.

We finally remark that the results in this section can be easily extended to the case that the processing times are mixtures of Erlang distributions with the same scale parameters.

Remark 4.3.2 *(root finding)*
We have seen that the problem of finding the probabilities boils down to finding the roots of the polynomial equation (4.31). Standard methods to find the zeros of a polynomial of degree $r+1$ may suffer from numerical instability if r is large, and a coefficient of variation

of 0.1 already gives rise to $r = 100$. Here the roots can be efficiently determined as follows. Equation (4.31) can be rewritten as

$$x^r = \frac{\lambda}{\lambda + \mu(1-x)}.$$

Raising both sides to the power $1/r$ we obtain r fixed point equations,

$$x = \phi F(x), \tag{4.38}$$

where ϕ satisfies $\phi^r = 1$ and

$$F(x) = \sqrt[r]{\lambda/(\lambda + \mu(1-x))}.$$

By applying Rouché's Theorem it can be proved for each feasible ϕ that equation (4.38) has exactly one root with $|x| < 1$ (cf. exercise 4.6.7). It is easily verified that $F(x)$ is convex for $0 \leq x \leq 1$. This implies, together with $F(0) > 0$, $F(1) = 1$ and $F'(1) = 1/\rho > 1$, that there is a unique χ with $0 < \chi < 1$ satisfying

$$\chi = F(\chi). \tag{4.39}$$

In exercise 4.6.8 it is shown that $\phi F(x)$ is a *contraction* on $|x| \leq \chi$. Hence, for given ϕ, the root of (4.38) with $|x| < 1$ can be determined by successive substitutions, i.e.,

$$\begin{aligned} x^{(0)} &= 0, \\ x^{(k+1)} &= \phi F(x^{(k)}), \qquad k = 0, 1, \ldots \end{aligned}$$

The sequence $x^{(0)}, x^{(1)}, \ldots$ converges exponentially fast to the desired root.

4.3.5 The waiting time in the $M/E_r/1$ queue

To determine the mean and distribution of the waiting time note that the solution (4.36) is a generalization of the one for $M/M/1$ model (not one, but a sum of geometric distributions). Based on this observation, results for the mean and distribution of the waiting time in the $M/M/1$ can be readily generalized to the case of Erlang service times, yielding (cf. (4.16)

$$P(W > t) = \sum_{k=1}^{r} C_k x_k e^{-\mu(1-x_k)t}, \qquad t > 0, \tag{4.40}$$

and

$$E(W) = \sum_{k=1}^{r} C_k \frac{x_k}{\mu(1-x_k)}. \tag{4.41}$$

By using the PASTA property and Little's law it is also possible to directly derive an expression for $E(W)$. Given that on arrival of a customer the server is busy, the mean residual service time $E(R)$ of the customer in service is equal to

$$\begin{aligned} E(R) &= \frac{1}{r}\frac{r}{\mu} + \frac{1}{r}\frac{r-1}{\mu} + \cdots + \frac{1}{r}\frac{1}{\mu} \\ &= \frac{r+1}{2\mu}. \end{aligned}$$

Namely, with probability $1/r$ the server works on phase i of the service time. The probability that the server is busy on arrival is equal to ρ. An arriving customer further has to wait for all customers in the queue. According to the PASTA property the mean number of customers waiting in the queue is equal to $E(L^q)$. Hence we find

$$E(W) = E(L^q)\frac{r}{\mu} + \rho E(R).$$

By Little's law we further have

$$E(L^q) = \lambda E(W).$$

From the two equations above it follows that

$$\begin{aligned} E(W) &= \frac{\rho E(R)}{1-\rho} \\ &= \frac{\rho}{1-\rho}\frac{r+1}{2\mu}. \end{aligned}$$

In table 4.2 we list for varying values of ρ and r the mean waiting time and some waiting time probabilities. The squared coefficient of variation of the service time is denoted by c_B^2. We see that the variation in the service times is important to the behavior of the system.

ρ	r	c_B^2	$E(W)$		$P(W>t)$		
				t	5	10	20
0.8	1	1	4		0.29	0.11	0.02
	2	0.5	3		0.21	0.05	0.00
	4	0.25	2.5		0.16	0.03	0.00
	10	0.1	2.2		0.12	0.02	0.00
0.9	1	1	9		0.55	0.33	0.12
	2	0.5	6.75		0.46	0.24	0.06
	4	0.25	5.625		0.41	0.18	0.04
	10	0.1	4.95		0.36	0.14	0.02

Table 4.2: Performance characteristics for the $M/E_r/1$ with mean service time 1

4.4 The $M/G/1$ queue

In this section we consider the case of general service times. The state description of the system at time t is now more complicated than in the models with phase-type distributions. The state can be described by the pair (n, s) where n is the number of customers in the system and s is a continuous variable denoting the elapsed service time of the customer in service. It is easier to observe the system just after departure epochs. Then all customers in the system are 'fresh' and thus $s = 0$. The distribution of the number of customers just after a departure can be determined elegantly using transform techniques. This will be treated in part b of this course. Here we will restrict our attention to the mean performance, which can be obtained directly with the mean-value technique.

We consider a single server. Customers arrive according to a Poisson stream with rate λ and are treated in order of arrival. The service times are independent and identically distributed with mean $E(B)$ and coefficient of variation c_B. For stability we have to require

$$\rho = \lambda E(B) < 1. \tag{4.42}$$

For the mean waiting time $E(W)$ it holds that

$$E(W) = E(L^q)E(B) + \rho E(R),$$

where the first term is the mean time to clear the queue and the second one is the mean amount of work at the server. The mean residual service time is given by (see section 2.8)

$$E(R) = \frac{1}{2}(1 + c_B^2)E(B).$$

Together with Little's law (applied to the queue)

$$E(L^q) = \lambda E(W),$$

we get

$$\begin{aligned} E(W) &= \frac{\rho E(R)}{1-\rho} \\ &= \frac{\rho E(B)}{2(1-\rho)}(1 + c_B^2). \end{aligned} \tag{4.43}$$

This formula is known as the famous Pollaczek-Khintchine formula for the mean waiting time. The formula for the Laplace-Stieltjes transform of the waiting time will be derived in part b. From (4.43) we find for the mean sojourn time $E(S)$ and the mean number of jobs in the system $E(L)$ that

$$E(S) = E(W) + E(B) = \frac{\rho E(R)}{1-\rho} + E(B)$$

and

$$E(L) = \lambda E(S) = \frac{\rho^2}{2(1-\rho)}(1 + c_B^2) + \rho.$$

An important conclusion from these results is that the mean performance only depends on the mean and standard deviation of the service time. It is *insensitive* to the distribution of the service time. Hence, in practice, the mean performance can be evaluated by simply collecting information on the mean and standard deviation of the service time.

4.5 $G/PH/1$ queue

In practice Poisson arrivals often provide a good description of the arrival process, but there are also many cases of interest in which Poisson arrivals are not appropriate. Think for instance of appointment systems or communication systems where the input may have a strong bursty character (during long periods no traffic and for short periods a lot of traffic) In this section we deal with systems with generally distributed interarrival times and we show in this case how the waiting time distribution can be found.

Consider a single server. The interarrival times have an arbitrary interarrival time distribution $A(t)$ with mean $1/\lambda$. The service time distribution is a mixture of Erlang distributions with the same scale parameters. That is, with probability q_n it is the sum of n exponentials with the same mean $1/\mu$, $n = 1, 2, \ldots, r$. The phase representation of this distribution is illustrated in figure 2.4. So the mean service time $E(B)$ is equal to

$$E(B) = \sum_{n=1}^{r} q_n \frac{n}{\mu}.$$

For stability we assume

$$\lambda E(B) < 1.$$

At an arbitrary point t in time the state of the system can be described by the pair (x, n) where x is the time elapsed till the last arrival before t and n is the number of uncompleted phases work in the system. Since the first variable is continuous this description leads to a complicated analysis. Therefore we will look at special points in time, namely at arrival instants. On arrival the state description simplifies to n, since $x = 0$.

So we consider the system at arrival instants, or in other words, we look at the *imbedded Markov chain* just before an arrival. The states are the number of uncompleted phases work in the system just before an arrival and the transition probabilities are given by

$$p_{nm} = \begin{cases} 0\,, & n \geq 0\,,\ \ m \geq n+r+1\,, \\ q_1\beta_{n+1-m} + q_2\beta_{n+2-m} + \cdots + q_r\beta_{n+r-m}\,, & n \geq 0\,,\ \ 1 \leq m \leq n+r\,, \end{cases}$$

where β_k is defined as the probability that k service stages are completed during an interarrival time, so

$$\beta_k = \int_0^\infty \frac{(\mu t)^k}{k!} e^{-\mu t} dA(t)\,, \qquad k \geq 0\,.$$

The remaining transition probabilities p_{n0} satisfy of course

$$p_{n0} = 1 - \sum_{m=1}^{\infty} p_{nm}.$$

Note that the imbedded Markov chain is irreducible and aperiodic. Since we assumed that the system is stable, the equilibrium probabilities $p_n, n = 0, 1, \ldots$ exist. Below the balance equations in the states $n > 0$ are formulated. The one in $n = 0$ is omitted, because of its minor importance to the analysis.

$$\begin{aligned} p_n &= \sum_{m=0}^{\infty} p_{n-1+m} q_1 \beta_m + \sum_{m=0}^{\infty} p_{n-2+m} q_2 \beta_m + \cdots + \sum_{m=0}^{\infty} p_{n-r+m} q_r \beta_m, \\ & \qquad n = 1, 2, \ldots , \end{aligned} \tag{4.44}$$

where by convention

$$p_{-1} = p_{-2} = \cdots = p_{-r+1} = 0. \tag{4.45}$$

The probability p_n can be interpreted as the probability that an arbitrary customer finds on arrival i phases work in the system. So once we know these probabilities we can calculate the waiting time of a customer.

The approach to solve the equilibrium equations is to seek solutions of equation (4.44) of the form

$$p_n = x^n \tag{4.46}$$

and then to linearly combine these solutions to also satisfy the convention (4.45) and the normalization condition. Note that, since the equilibrium equations are dependent, the equation in $n = 0$ will be satisfied automatically. Inserting (4.46) into (4.44) and then dividing by x^{n-r} yield

$$\begin{aligned} x^r &= (x^{r-1} q_1 + x^{r-2} q_2 + \cdots + q_r) \sum_{m=0}^{\infty} x^m \beta_m \\ &= (x^{r-1} q_1 + x^{r-2} q_2 + \cdots + q_r) \tilde{A}(\mu(1-x)) \, , \end{aligned} \tag{4.47}$$

where $\tilde{A}(s)$ is the Laplace-Stieltjes transform of $A(t)$, i.e.,

$$\tilde{A}(s) = E(e^{-As}) = \int_0^{\infty} e^{-\mu t} dA(t), \qquad Re\, s > 0.$$

Clearly, because of the normalization condition, only solutions with $|x| < 1$ are useful. Applying Rouché's theorem it can be shown that equation (4.47) has exactly r roots inside the unit circle. For convenience, we assume that the roots of (4.47) with $|x| < 1$ are simple, so that we have r different roots. Hence, we find r solutions of the form $p_n = x^n$ with $|x| < 1$ satisfying equation (4.44). These solutions are labeled x_k^n, $k = 1, \ldots, r$. It now remains to determine coefficients C_k such that the linear combination

$$p_n = \sum_{k=1}^{r} C_k (1 - x_k) x_k^n \, , \qquad n = 0, 1, 2, \ldots$$

satisfies the convention (4.45) and the normalization condition. This proceeds along exactly the same lines as in section 4.3.4. We can conclude that the arrival distribution for the $G/PH/1$ model is of the same form as the one for the $M/E_r/1$ model (see (4.36)), and thus also the waiting time distribution is given by expression (4.40).

When the service times have a pure Erlang distribution, it can be shown that under certain conditions for the interarrival time distribution the roots x_k can be found by successive substitutions similar as explained in remark 4.3.2 (see [4]).

4.6 Exercises

Exercise 4.6.1
Prove lemma (4.1.1).

Exercise 4.6.2 *(bulk arrivals)*
In a work station orders arrive according to a Poisson arrival process with arrival rate λ. An order consists of a number of independent jobs. The number N of jobs in an order is geometrically distributed:

$$P(N = k) = (1 - p)p^{k-1}$$

with $k = 1, 2, \ldots$ and $0 \leq p < 1$. Each job requires an exponentially distributed amount of processing time with mean $1/\mu$.

(i) Derive the distribution of the workload of a job.

(ii) Determine the distribution of the number of orders in the system.

Exercise 4.6.3 *(variable production rate)*
Consider a work station where jobs arrive according to a Poisson process with arrival rate λ. The jobs have an exponentially distributed service time with mean $1/\mu$. So the service completion rate (the rate at which jobs depart from the system) is equal to μ.
If the queue length drops below the threshold Q_L the service completion rate is lowered to μ_L. If the queue length reaches Q_H, where $Q_H \geq Q_L$, the service rate is increased to μ_H. (L stands for low, H for high.) Determine the queue length distribution and the mean time spent in the system.

Exercise 4.6.4 *(a fast and a slow machine)*
Consider two parallel machines with a common (infinite) buffer where jobs arrive according to a Poisson stream with rate λ. The processing times are exponentially distributed with mean $1/\mu_1$ on machine 1 and $1/\mu_2$ on machine 2 ($\mu_1 > \mu_2$). Jobs are processed in order of arrival. A job who arrives when both machines are idle is assigned to the fast machine. To guarantee that the system can handle the amount of work offered per unit time we assume $\rho < 1$ where

$$\rho = \frac{\lambda}{\mu_1 + \mu_2}\,.$$

(i) Determine the distribution of the number of jobs in the system.

(ii) Use this distribution to derive the mean number of jobs in the system.

(iii) When is it better to not use the slower machine at all?

(iv) Calculate for the following two cases the mean number of jobs in the system with and without the slow machine.

(a) $\lambda = 2$, $\mu_1 = 5$, $\mu_2 = 1$;

(b) $\lambda = 3$, $\mu_1 = 5$, $\mu_2 = 1$.

For more details concerning this problem the interested reader is referred to [37]. There it is shown that one should not remove the slow machine if $r > 0.5$ where $r = \mu_2/\mu_1$. When $0 \le r < 0.5$ the slow machine should be removed (and the resulting system is stable) whenever $\rho \le \rho_c$, where

$$\rho_c = \frac{2 + r^2 - \sqrt{(2 + r^2)^2 + 4(1 + r^2)(2r - 1)(1 + r)}}{2(1 + r^2)}.$$

Exercise 4.6.5 *(base-stock policy)*
For a given product one has a stock of N items in order to directly satisfy customer demand. The demand (for one item at a time) is a Poisson process with rate λ per week. When demand for an item occurs, one directly orders a new one at the supplier. This ordering policy is called *base-stock policy*. The lead time is exponentially distributed with mean $1/\mu$ weeks. Each demand which cannot be satisfied directly from stock on hand is lost.
This inventory system is completely characterized by the number of outstanding orders. Let p_n, $n = 0, \ldots, N$ denote the probability that there are n outstanding orders.

(i) Determine the probabilities p_n.

(ii) What is the fraction of customer demand which is lost?

(iii) Denote the fraction of customer which is lost when the (maximal) stock level is N by B_N. Derive a recurrence relation to determine $B_0, B_1, B_2, \ldots$

Suppose that $\lambda = \mu = 2$.

(iv) Determine the minimum stock level such that at least 90% of customer demand can be satisfied from stock on hand.

Exercise 4.6.6 *(mean clearance time)*
Consider a work station where jobs arrive according to a Poisson process with rate λ. The jobs have an exponentially distributed service time with mean $1/\mu$. Assume that $\lambda < \mu$. Let C_n be the time till the work station is empty again if there are now n jobs present in the system, $n = 0, 1, 2, \ldots$ (so $C_0 = 0$).

(i) Show that the means $E(C_n)$ satisfy the following recurrence relation,

$$E(C_n) = \frac{1}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} E(C_{n-1}) + \frac{\lambda}{\lambda + \mu} E(C_{n+1}), \qquad n = 1, 2, \ldots$$

(ii) Show that

$$E(C_1) = \frac{1}{\mu - \lambda}.$$

(iii) Solve the recursion relation for $E(C_n)$.

Exercise 4.6.7 *(Rouché's theorem)*
Rouché's theorem reads as follows.
Let the bounded region D have as its boundary a contour C. Let the functions $f(z)$ and $g(z)$ be analytic both in D and on C, and assume that $|f(z)| < |g(z)|$ on C. Then $f(z)+g(z)$ has in D the same number of zeros as $g(z)$, all zeros counted according to their multiplicity.

(i) Use Rouché's theorem to prove that the polynomial equation (4.31) has exactly r (possibly complex) roots x with $|x| < 1$.
(Hint: take the left-hand side of (4.31) as $f(z)$, its right-hand side as $g(z)$ and take as contour C the circle with center 0 and radius $1 - \epsilon$ with ϵ small and positive.)

(ii) Show that all roots are simple.

Exercise 4.6.8 *(contraction mapping)*
Show for the function $\phi F(z)$ in remark 4.3.2 that for all $|x|, |y| \leq \chi$,

(i) $|\phi F(x)| \leq \chi$,

(ii) $|\phi F(x) - \phi F(y)| \leq |x - y| F'(\chi)$,

where χ is defined by (4.39). Since $F'(\chi) < 1$, properties (i) and (ii) imply that $\phi F(x)$ is a contraction mapping on $|x| \leq \chi$.

Exercise 4.6.9 *(VanderMonde determinant)*
Let $\alpha_1, \alpha_2, \ldots$ be a sequence of numbers and define the determinant Δ_n for $n = 1, 2, \ldots$ by

$$\Delta_n = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \vdots & & & \vdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \cdots & \alpha_n^{n-1} \end{vmatrix}.$$

The determinant Δ_n is called a *VanderMonde determinant* of order n.

(i) Derive the following recurrence relation for the determinants Δ_n,

$$\Delta_n = \Delta_{n-1} \prod_{m=1}^{n-1} (\alpha_m - \alpha_n), \qquad n = 1, 2, \ldots$$

(ii) Show that the solution of this recurrence relation is given by

$$\Delta_n = \prod_{1 \leq i < j \leq n} (\alpha_j - \alpha_i).$$

Exercise 4.6.10
Prove expression (4.37) for the constant C.
(Hint: consider the polynomial

$$p(x) = \sum_{k=1}^{r} \frac{\prod_{j \neq k}(x - y_j)}{\prod_{j \neq k}(y_k - y_j)},$$

and show that it is identical to 1.)

Exercise 4.6.11 *(Hyper-exponential processing times)*
Consider a single machine where orders arrive according to a Poisson stream with rate λ. The processing times are hyperexponentially distributed. With probability q_1 the processing time is exponential with mean $1/\mu_1$ and with probability q_2 it is exponential with mean $1/\mu_2$ ($q_1 + q_2 = 1$). The hyperexponential distribution can be used to model processing times with a coefficient of variation greater than 1 (see section 2.3), but it can also arise naturally (when there are two types of jobs).
The state of this system can be described by a pair (n, m) where n is the number of jobs in the system and m indicates the type of job being processed. If the job being processed is exponential with parameter μ_1 (resp. μ_2) then $m = 1$ (resp. $m = 2$).

(i) Describe the flow diagram and give the set of balance equations for the equilibrium state probabilities.

(ii) Solve these equations.

(iii) Determine the mean number of jobs in the system and the mean time spent in the system.

(iv) Derive an expression for the distribution of the time spent in the system.

Exercise 4.6.12 *(Coxian processing times)*
Consider the same system as in the previous exercise, except that now the the processing times are Coxian distributed of order 2 (see figure 2.3). The processing time consists of an exponential phase with mean $1/\mu_1$ which is followed with probability q by a second exponential phase with mean $1\mu_2$. The Coxian-2 distribution can be used to model processing times with a coefficient of variation greater than 0.5 (see section 2.3).

(i) How can the state of this system be characterized?

(ii) Describe the flow diagram and give the set of balance equations for the state probabilities.

(iii) Find the solution to these equations.

(iv) Determine the mean number of jobs in the system and the mean time spent in the system.

(v) Derive an expression for the distribution of the time spent in the system.

Exercise 4.6.13 *(Two phase production)*
A machine produces products in two phases. The first phase is standard and the same for all products. The second phase is customer specific (the finishing touch). The first (resp. second) phase takes an exponential time with mean $1/\mu_1$ (resp. $1/\mu_2$) hour. Orders for the production of one product arrive according to a Poisson stream with rate λ orders per hour. Orders are processed in order of arrival.
The state of this system can be described by the pair (n, m) where n is the number of orders in the system and m indicates the production phase (so m is 1 or 2).

(i) Formulate and solve the balance equations for the state probabilities.

Suppose that $\mu_1 = 6, \mu_2 = 30$ and $\lambda = 3$.

(ii) Determine the mean production lead time of an order.

To reduce the production lead time one decides to start already with the production of phase 1 when the system is empty. If upon completion of phase 1 no order arrived yet, the machine stops the production and waits for the first order to arrive. When the first order arrives the machine can directly start with phase 2.

(iii) Formulate and solve the balance equations for this slightly modified system.

(iv) Determine the reduction in the mean production lead time. Try to explain this reduction.

Exercise 4.6.14
Consider a machine where jobs are being processed. The mean production time is 4 minutes and the standard deviation is 3 minutes. We have seen in exercise 2.12.5 that the production times can be modelled as a mixture of an exponential distribution and an Erlang-2 distribution with the same scale parameters.
The mean number of jobs arriving per hour is 10. Suppose that the interarrival times are exponentially distributed.

(i) Determine the mean waiting time of the jobs.

Now suppose that the interarrival times are uniformly distributed between 4 and 8 minutes. The state of this system just before the arrival of an order can be described by n, where n is the number of uncompleted production phases in the system.

(ii) Formulate the balance equations for the (equilibrium) arrival probabilities.

(iii) Determine the arrival probabilities.

(iv) Determine the mean waiting time of the jobs. Compare the answer with (i) and explain the result.

(v) What is the fraction of the orders which has to wait longer than 10 minutes?

Exercise 4.6.15
Orders arrive according to a Poisson process, so the interarrival times are independent and exponentially distributed. Each order has to pass, stage by stage, through a series of R independent production stages. Each stage takes a exponentially distributed time. The total production time, the workload, of the order is the sum of R independent, identically distributed random variables. So the distribution of the workload is the R-stage Erlangian distribution.

(i) The state of this system (i.e., the work station together with the queue in front of it) can be characterized by the number of orders in the system and by the number of not yet completed stages for the order in the work station.
Describe the flow diagram and give the set of equations for the equilibrium state probabilities.

(ii) The state of the system at a certain time can also be described by the total number of production stages that have to be completed before all the orders in the system are ready.
Give the flow diagram and formulate the equations for the equilibrium state probabilities.

(iii) Define

$$\begin{aligned} p_n &= P(\text{total number of orders in the system} = n), \\ q_k &= P(\text{total number of production stages in the system } = k). \end{aligned}$$

Give the relation between these two probabilities.

Exercise 4.6.16
Orders arrive in so called bulks. Each bulk consists of R independent orders. The bulks themselves arrive according to a Poisson process, with independent exponentially distributed interarrival times. The sequence in which orders of one bulk are processed in the work station is unimportant. All orders of a bulk have to wait until all orders of bulks that have arrived earlier are completed. The workload of each order is exponentially distributed. The state of the system is simply characterized by the number of orders in the system.
Describe the flow diagram for this situation. Compare the result with that of the previous exercise, part (ii).

Exercise 4.6.17
Two types of orders arrive according to independent Poisson processes with arrival rates λ_1 and λ_2 respectively. Type i orders require an exponentially distributed time with mean $1/\mu_i$, $i = 1, 2$.

Describe the state of the system, give the flow diagram and the balance equations for the equilibrium state probabilities. Determine the mean production lead time of the type 1 orders and of the type 2 orders.

Exercise 4.6.18
Consider an $M/G/1$ queue, where the server successfully completes a service time with probability p. If a service time is not completed successfully, it has to be repeated until it is successful. Determine the mean sojourn time of a customer in the following two cases:

(i) The repeated service times are identical.

(ii) The repeated service times are independent, and thus (possibly) different.

Exercise 4.6.19
Consider a queueing system with a single server. Every 4 minutes a customer arrives. The service times are exponentially distributed with a mean of 2 minutes. Determine the distribution of the waiting time and the distribution of the sojourn time.

Exercise 4.6.20
The distribution of the interarrival times is given by

$$A(t) = \frac{13}{14}(1 - e^{-3t}) + \frac{11}{14}(1 - e^{-2t}).$$

The service times are exponential with mean $1/6$, and there is one server.

(i) Determine the number of customers in the system on arrival.

(ii) Determine the distribution of the sojourn time.

Chapter 5

Variations on the $M/G/1$ theme

In this chapter we treat some variations of $M/G/1$ model. In the first section we consider a production system with setup times. Next we will analyze an unreliable production system. In the final section we revisit the problem with a fast and a slow machine which has been introduced in exercise 4.6.4.

5.1 Machine with setup times

Consider a single machine where jobs are being processed in order of arrival. Since it is expensive to keep the machine in operation while there are no jobs, the machine is turned off as soon as the system is empty. When a new job arrives the machine is turned on again, but it takes some time till the machine is ready for processing. So turning off the machine leads to longer production leadtimes. The question arises whether it is better to keep the machine continuously in operation and taking the extra costs for granted. Some simple models for answering this question are formulated and studied below.

5.1.1 Exponential processing and setup times

Suppose that the jobs arrive according to a Poisson stream with rate λ and that the processing times are exponential with mean $1/\mu$. For stability we require that $\rho = \lambda/\mu < 1$. The setup time of the machine is also exponential with mean $1/\theta$. Due to the exponential assumptions the state of the system can be characterized by the pair (n, m) where n is the number of jobs in the system and m denotes the state of the machine. If $m = 1$ the machine is in operation, if $m = 0$ the machine is off or in its setup phase. The flow diagram of this system is shown in figure 5.1.

By equating in each state the flow out of and into that state we obtain the following set of balance (or equilibrium) equations.

$$p(0,0)\lambda = p(1,1)\mu, \tag{5.1}$$

$$p(n,0)(\lambda+\theta) = p(n-1,0)\lambda, \qquad n = 1, 2, \ldots \tag{5.2}$$

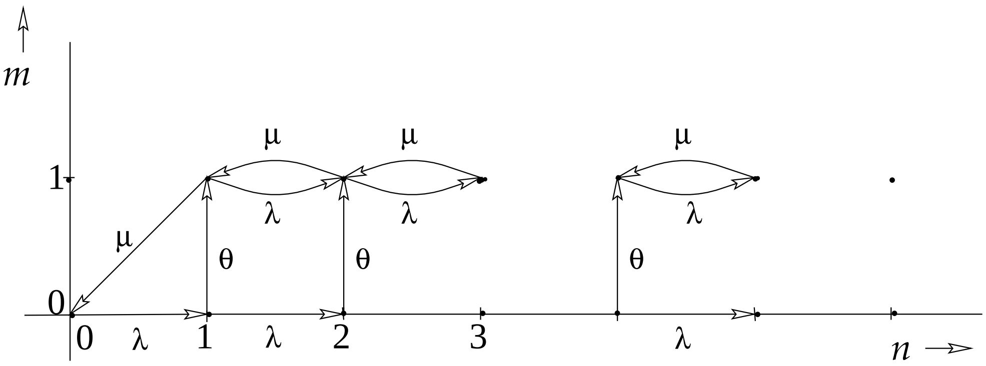

Figure 5.1: Flow diagram for the $M/M/1$ model with setup times

$$p(1,1)(\lambda+\mu) = p(1,0)\theta + p(2,1)\mu, \tag{5.3}$$
$$p(n,1)(\lambda+\mu) = p(n,0)\theta + p(n+1,1)\mu + p(n-1,1)\lambda, \qquad n = 2,3,\ldots \tag{5.4}$$

These balance equations can be solved as follows. Equation (5.2) is a first order difference equation which has as solution

$$p(n,0) = p(0,0)x^n, \qquad n = 0,1,\ldots \tag{5.5}$$

with

$$x = \frac{\lambda}{\lambda+\theta}.$$

Note that ρ is the fraction of time the machine is processing jobs. Hence the fraction of time the machine is not processing jobs is given by

$$\sum_{n=0}^{\infty} p(n,0) = 1-\rho,$$

whence with (5.5),

$$p(0,0) = (1-\rho)(1-x).$$

To find the probabilities $p(n,1)$ we substitute (5.5) into (5.4) yielding

$$p(n,1)(\lambda+\mu) = (1-\rho)(1-x)x^n\theta + p(n+1,1)\mu + p(n-1,1)\lambda. \tag{5.6}$$

This is a second order *inhomogeneous* recurrence relation. Its general solution is given by

$$p(n,1) = A(1-\rho)(1-x)x^n + B_1 y_1^n + B_2 y_2^n, \qquad n = 1,2,\ldots \tag{5.7}$$

The first term is a particular solution to the inhomogeneous relation, where

$$A = \frac{\theta x}{(\lambda+\mu)x - \mu x^2 - \lambda} = \frac{\lambda+\theta}{\mu-\lambda-\theta}.$$

For convenience we assume $\theta \neq \mu - \lambda$, so that the denominator above does not vanish. In case this assumption does not hold the particular solution is of the form nx^n, see exercise 5.5.1. The second term is the general solution of the homogeneous equation in which the coefficients B_1 and B_2 are free and y_1 and y_2 are the roots of

$$(\lambda + \mu)y = \mu y^2 + \lambda.$$

Hence $y_1 = \lambda/\mu = \rho$ and $y_2 = 1$. Since the probabilities $p(n,m)$ have to add up to one, we have to set $B_2 = 0$. Finally substitution of (5.5) and (5.7) into equation (5.3) yields $B_1 = -A$. Remark that since the equilibrium equations are dependent the remaining equation in (0,0) is also satisfied. This completes the solution of the balance equations. Summarizing we have found

$$\begin{aligned} p(n,0) &= (1-\rho)(1-x)x^n, \qquad n = 0,1,\ldots \\ p(n,1) &= A(1-\rho)(1-x)(x^n - \rho^n), \qquad n = 1,2,\ldots \end{aligned}$$

From the equilibrium distribution we can derive the mean number of jobs in the system, yielding

$$\begin{aligned} E(L) &= \sum_{n=1}^{\infty} n(p(n,0) + p(n,1)) \\ &= (1-\rho)(1-x)\left[\frac{x}{(1-x)^2} + \frac{Ax}{(1-x)^2} - \frac{A\rho}{(1-\rho)^2}\right], \end{aligned}$$

and by using Little's law we obtain the mean production leadtime $E(S)$. It is also possible to derive the distribution of the production leadtime. By conditioning on the state seen on arrival and using PASTA we get

$$\begin{aligned} P(S > t) &= \sum_{n=0}^{\infty} p(n,0)P(Y + X_1 + \ldots + X_{n+1} > t) \\ &+ \sum_{n=1}^{\infty} p(n,1)P(X_1 + \ldots + X_{n+1} > t), \end{aligned}$$

where Y is an exponential (residual) setup time with mean $1/\theta$ and $X_1, X_2, \ldots$ are exponential processing times with mean $1/\mu$. Substituting the expressions for $p(n,0)$ and $p(n,1)$ and using that a geometric sum of exponential random variables is again exponential (see exercise 2.12.3), we get

$$P(S > t) = \frac{1}{\mu - \lambda - \theta}\left[\mu(1-\rho)e^{-\theta t} - \theta e^{-\mu(1-\rho)t}\right]. \tag{5.8}$$

Now we have seen how to calculate, for instance, the mean and distribution of the production lead time, one should be able to decide what to do: keep the machine continuously in operation or to turn it off when there is no work.

The means $E(L)$ and $E(S)$ can also be determined by combining the PASTA property and Little's law, in the same way as we have seen for the $M/M/1$ system. Based on PASTA we know that the average number of jobs in the system seen an arriving job equals $E(L)$ and each of them (also the one being processed) has a (residual) processing time with mean $1/\mu$. With probability $1-\rho$ the machine is not in operation on arrival, so that the job also has to wait for the setup phase with mean $1/\theta$. Further the job has to wait for its own processing time. Hence

$$E(S) = (1-\rho)\frac{1}{\theta} + E(L)\frac{1}{\mu} + \frac{1}{\mu}$$

and together with

$$E(L) = \lambda E(S)$$

we find

$$E(S) = \frac{1/\mu}{1-\rho} + \frac{1}{\theta}\,.$$

So the mean production lead time is equal to the one in the system where the machine is always on, plus an extra delay caused by turning off the machine when there is no work. In fact, it can be shown (by using, e.g., a sample path argument) that the extra delay is an exponential time with mean $1/\theta$.

5.1.2 General processing and setup times

We now consider the model with general processing times and general setup times. The arrivals are still Poisson with rate λ. The mean and coefficient of variation of the processing time are denoted by $E(B)$ and c_B respectively. $E(T)$ and c_T are the mean and coefficient of variation of the setup time. For stability we require that $\rho = \lambda E(B) < 1$. Below we show that the mean waiting time can be determined by using the mean-value technique.

The mean waiting time $E(W)$ of a job satisfies

$$\begin{aligned} E(W) &= E(L^q)E(B) + \rho E(R_B) \\ &+ \Pr(\text{Machine is off on arrival})E(T) \\ &+ \Pr(\text{Machine is in setup phase on arrival})E(R_T), \end{aligned} \tag{5.9}$$

where $E(R_B)$ and $E(R_T)$ are the mean residual processing and residual setup time, so

$$E(R_B) = \frac{1}{2}(1+c_B^2)E(B), \qquad E(R_T) = \frac{1}{2}(1+c_T^2)E(T).$$

To find the probability that on arrival the machine is off (i.e. not working and not in the setup phase), note that by PASTA, this probability is equal the faction of time that the machine is off. Since a period in which the machine is not processing jobs consists of an interarrival time followed a setup time, we get

$$\Pr(\text{Machine is off on arrival}) = (1-\rho)\frac{1/\lambda}{1/\lambda + E(T)}.$$

Similarly we find

$$\Pr(\text{Machine is in setup phase on arrival}) = (1-\rho)\frac{E(T)}{1/\lambda + E(T)}.$$

Substituting these relations into (5.9) and using Little's law stating that

$$E(L^q) = \lambda E(W)$$

we finally obtain that

$$E(W) = \frac{\rho E(R_B)}{1-\rho} + \frac{1/\lambda}{1/\lambda + E(T)}E(T) + \frac{E(T)}{1/\lambda + E(T)}E(R_T).$$

Note that the first term at the right-hand side is equal to the mean waiting time in the $M/G/1$ without setup times, the other terms express the extra delay due to the setup times.

5.1.3 Threshold setup policy

Instead of switching on the machine upon arrival of the first job, it may be better to wait till the number of jobs reaches some threshold value, N say. This situation can be analyzed along the same lines as in the previous section. The mean waiting time now satisfies

$$\begin{aligned} E(W) &= E(L^q)E(B) + \rho E(R_B) \\ &+ \sum_{i=1}^{N} \Pr(\text{Arriving job is number } i)\left(\frac{N-i}{\lambda} + E(T)\right) \\ &+ \Pr(\text{Machine is in setup phase on arrival})E(R_T), \end{aligned} \tag{5.10}$$

The probability that an arriving job is the i-th one in a cycle can be determined as follows. A typical production cycle is displayed in figure 5.2.

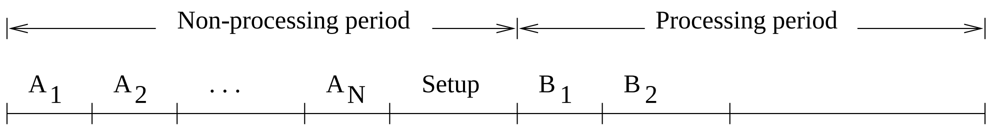

Figure 5.2: Production cycle in case the machine is switched on when there are N jobs

The probability that a job arrives in a non-processing period is equal to $1-\rho$. Such a period now consists of N interarrival times followed by a setup time. Hence, the probability that a job is the i-th one, given that the job arrives in a non-processing period, is equal to $1/\lambda$ divided by the mean length of a non-processing period. So

$$\Pr(\text{Arriving job is number } i) = (1-\rho)\frac{1/\lambda}{N/\lambda + E(T)}, \qquad i = 1, \ldots, N,$$

and similarly,

$$\Pr(\text{Machine is in setup phase on arrival}) = (1-\rho)\frac{E(T)}{N/\lambda + E(T)}.$$

Substituting these relations into (5.10) we obtain, together with Little's law, that

$$E(W) = \frac{\rho E(R_B)}{1-\rho} + \frac{N/\lambda}{N/\lambda + E(T)}\left(\frac{N-1}{2\lambda} + E(T)\right) + \frac{E(T)}{1/\lambda + E(T)}E(R_T).$$

5.2 Unreliable machine

In this section we consider an unreliable machine processing jobs. The machine breaks down at random instants, whether it is processing or not. What is the impact of these breakdowns on the production leadtimes? To obtain some insight in the effects of the breakdowns we formulate and study some simple models in the following subsections.

5.2.1 Exponential processing and down times

Suppose that jobs arrive according to a Poisson stream with rate λ. The processing times are exponential with mean $1/\mu$. The time between two breakdowns is exponential with mean $1/\eta$ and the down time is also exponential with mean $1/\theta$.

The state of the system can be characterized by the pair (n, m) where n is the number of jobs in the system and m denotes the state of the machine. If $m = 1$ the machine is up, if $m = 0$ the machine is down and in repair. The flow diagram of this system is shown in figure 5.3.

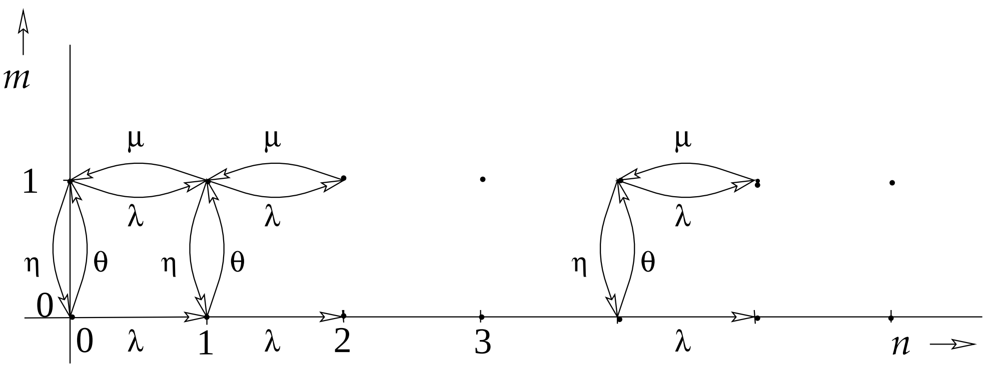

Figure 5.3: Flow diagram for the $M/M/1$ model with random breakdowns

What is the condition under which the machine can handle the amount of work offered per unit time? Let ρ_U be the fraction of time the machine is up, so

$$\rho_U = \frac{1/\eta}{1/\eta + 1/\theta} = \frac{1}{1+\eta/\theta}.$$

Then we have to require that

$$\frac{\lambda}{\mu} < \rho_U. \tag{5.11}$$

From the flow diagram we obtain the following set of balance equations.

$$\begin{aligned}
p(0,0)(\lambda+\theta) &= p(0,1)\eta, && (5.12)\\
p(n,0)(\lambda+\theta) &= p(n-1,0)\lambda + p(n,1)\eta, \qquad n = 1,2,\ldots && (5.13)\\
p(0,1)(\lambda+\eta) &= p(0,0)\theta + p(1,1)\mu, && (5.14)\\
p(n,1)(\lambda+\eta+\mu) &= p(n,0)\theta + p(n+1,1)\mu + p(n-1,1)\lambda, \quad n = 1,2,\ldots && (5.15)
\end{aligned}$$

These balance equations can be solved as follows. We first try to find solutions of (5.13) and (5.15) of the form

$$\begin{aligned}
p(n,0) &= x^n u,\\
p(n,1) &= x^n v.
\end{aligned}$$

Substituting this form into (5.13) and (5.15) yields

$$\begin{aligned}
xu(\lambda+\theta) &= u\lambda + xv\eta,\\
xv(\lambda+\eta+\theta) &= ux\theta + x^2 v\mu + v\lambda.
\end{aligned}$$

Set $u = 1$. Then we get from the first equation

$$v = \frac{(\lambda+\theta)x - \lambda}{\eta x}. \tag{5.16}$$

Substitution of this relation into the second equation leads to

$$x((\lambda+\theta)x-\lambda)(\lambda+\eta+\theta) = \eta\theta x^2 + ((\lambda+\theta)x-\lambda)\mu x^2 + ((\lambda+\theta)x-\lambda)\lambda$$

which can be rewritten as

$$(x-1)(\mu(\lambda+\theta)x^2 - \lambda(\lambda+\theta+\eta+\mu)x + \lambda^2) = 0.$$

The root $x = 1$ is not useful, since we must be able to normalize the solution. It can be easily shown that condition (5.11) implies that the quadratic equation above has two roots x_1 and x_2 satisfying $0 < x_1 < x_2 < 1$. Hence the solution of (5.13) and (5.15) is

$$\begin{aligned}
p(n,0) &= c_1 x_1^n + c_2 x_2^n, && (5.17)\\
p(n,1) &= c_1 v_1 x_1^n + c_2 v_2 x_2^n,
\end{aligned}$$

where v_i is given by (5.16) with $x = x_i$, $i = 1,2$ and the coefficients c_1 and c_2 are to be determined yet. These coefficients follow from the boundary conditions (5.12) and (5.14) together with the normalization condition. Inserting (5.17) into (5.12) yields

$$\frac{c_1}{x_1} + \frac{c_2}{x_2} = 0.$$

When c_1 and c_2 satisfy this relation, it can be verified that equation (5.14) is also satisfied (in fact, this is due to the dependence of the balance equations). From the condition that the probabilities have to add up to one, we obtain

$$\frac{c_1(1+v_1)}{1-x_1} + \frac{c_2(1+v_2)}{1-x_2} = 1$$

The two equations for c_1 and c_2 have as solution

$$c_1 = Cx_1, \qquad c_2 = -Cx_2,$$

where

$$C = \left\{ \frac{x_1(1+v_1)}{1-x_1} - \frac{x_2(1+v_2)}{1-x_2} \right\}^{-1}.$$

This completes the determination of the probabilities $p(n,m)$. Based on the solution for $p(n,m)$ it is easy to find the mean number of jobs in the system,

$$E(L) = C\left(\frac{(1+v_1)x_1^2}{(1-x_1)^2} - \frac{(1+v_2)x_2^2}{(1-x_2)^2} \right),$$

and $E(S)$ follows by using Little's law.

In table 5.1 we investigate the impact of the variability in the availability of the machine on the mean production leadtime. The mean production time is 1 hour ($\mu = 1$). The average number of jobs that arrives in a week (40 hours) is 32, so $\lambda = 0.8$ jobs per hour. The fraction of time the machine is available is kept constant at 90%, so $\rho_U = 0.9$. The rate η at which the machine breaks down is varied from (on the average) every 10 minutes till once a week. In the former case the mean down time is 1.1 minute, in the latter case it is more dramatic, namely nearly half a day (4.4 hours). The results show that the variation in the availability of the machine is essential to the behavior of the system.

η	θ	$E(S)$
6	54	10.02
3	27	10.03
1	9	10.1
0.125	1.125	10.8
0.0625	0.5625	11.6
0.025	0.225	14

Table 5.1: The mean production leadtime in hours as a function of the break-down rate η per hour for fixed availability of 90%

The mean production leadtime can also be determined directly (without using the probabilities $p(n,m)$) by using the mean-value technique. This is shown in the next subsection for general processing and down times.

5.2.2 General processing and down times

We now consider the model with general processing times and general down times. The arrivals are still Poisson with rate λ. The mean and coefficient of variation of the processing time are denoted by $E(B)$ and c_B respectively. The time between two breakdowns is exponential with mean $1/\eta$ and $E(D)$ and c_D are the mean and coefficient of variation of the down time. For stability we require that (cf. 5.11)

$$\lambda E(B) < \frac{1}{1+\eta E(D)}.$$

Below we show that the mean waiting time and the mean production leadtime can be determined by using the mean-value technique.

We first introduce the generalized processing time, which is defined as the processing time plus the down times occurring in that processing time. Denote the generalized processing time by G. Then we have

$$G = B + \sum_{i=1}^{N(B)} D_i,$$

where $N(B)$ is the number of break-downs during the processing time B and D_i is the i-th down time. For the mean of G we get by conditioning on B and $N(B)$ that

$$\begin{aligned}
E(G) &= \int_{x=0}^{\infty} \sum_{n=0}^{\infty} E(G|B=x, N(B)=n) e^{-\eta x} \frac{(\eta x)^n}{n!} f_B(x)dx \\
&= \int_{x=0}^{\infty} \sum_{n=0}^{\infty} (x + nE(D)) e^{-\eta x} \frac{(\eta x)^n}{n!} f_B(x)dx \\
&= \int_{x=0}^{\infty} \sum_{n=0}^{\infty} (x + x\eta E(D)) e^{-\eta x} \frac{(\eta x)^n}{n!} f_B(x)dx \\
&= E(B) + E(B)\eta E(D),
\end{aligned}$$

and similarly,

$$E(G^2) = E(B^2)(1+\eta E(D))^2 + E(B)\eta E(D^2).$$

A typical production cycle is shown in figure 5.4. The non-processing period consists of cycles which start with an up time. When the machine is up two things can happen: a job arrives, in which case the machine starts to work, or the machine goes down and has to be repaired. Hence, an up time is exponential with mean $1/(\lambda+\eta)$ and it is followed by a processing period (i.e. a job arrives) with probability $\lambda/(\lambda+\eta)$ or otherwise, it is followed by a down time. During a processing period the machine works on jobs with generalized processing times (until all jobs are cleared). For the mean waiting time it holds that

$$\begin{aligned}
E(W) &= E(L^q)E(G) + \rho_G E(R_G) \\
&+ \Pr[\text{Arrival in a down time in a non-processing period}]E(R_D), \qquad (5.18)
\end{aligned}$$

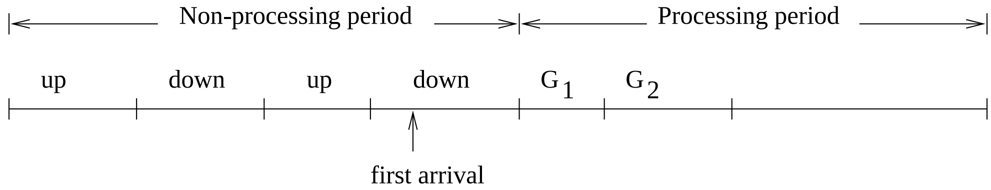

Figure 5.4: Production cycle of an unreliable machine

where ρ_G is the fraction of time the machine works on generalized jobs and $E(R_G)$ is the mean residual generalized processing time, so

$$\rho_G = \lambda E(G), \qquad E(R_G) = \frac{E(G^2)}{2E(G)}.$$

The probability that a job arriving in a non-processing period finds that the machine is down, is given by

$$\frac{E(D)\eta/(\lambda+\eta)}{1/(\lambda+\eta)+E(D)\eta/(\lambda+\eta)} = \frac{E(D)\eta}{1+E(D)\eta},$$

Hence, since a job arrives with probability $1-\rho_G$ in a non-processing period, we have

$$\Pr[\text{Arrival in a down time in a non-processing period}] = (1-\rho_G)\frac{E(D)\eta}{1+E(D)\eta}.$$

Substitution of this relation into (5.18) and using Little's law yields

$$E(W) = \frac{\rho_G E(R_G)}{1-\rho_G} + (1-\rho_G)\frac{E(D)\eta}{1+E(D)\eta}E(R_D),$$

and the mean production lead time follows from

$$E(S) = E(W) + E(G).$$

5.3 An $M/G/1$ queue with an exceptional first customer in a busy period

The life of the server in an $M/G/1$ queues is an alternating sequence of periods during which no work is done, the so-called idle periods, and periods during which the server helps customers, the so-called busy periods. In some applications the service time of the first customer in a busy period is different from the service times of the other customers served in thhe busy period. For example, consider the setup problem in section 5.1 again. This problem may, alternatively, be described by an $M/G/1$ queue in which the first job in a busy period has an exceptional service time, namely the setup time plus the actual processing time. The problem in section 5.2 can also be described by such a model. Here we can take the generalized processing times as the service times. However, on arrival of the first job in a busy period the machine can be down. Then the (generalized) servicing of that job cannot immediately start, but one has to wait till the machine is repaired. This repair time can be included in the service time of the first job. The determination of the distribution, or the moments of the service time of the first job is in this case more difficult then in the setup problem. Below we show how, in the general setting of an $M/G/1$ with an exceptional first customer, the mean sojourn time can be found by using the mean-value approach.

Let B_f and R_f denote the service time and residual service time, respectively, of the first customers in a busy period. For the mean waiting time we have

$$E(W) = E(L^q)E(B) + \rho_f E(R_f) + \rho E(R), \tag{5.19}$$

where ρ_f is the fraction of time the server works on first customers, and ρ is the fraction of time the server works on other (ordinary) customers. The probability that an arriving customer is the first one in a busy cycle is given by $1 - \rho_f - \rho$. Hence, the number of first customers arriving per unit of time is equal to $\lambda(1 - \rho_f - \rho)$. Thus ρ_f and ρ satisfy

$$\begin{aligned} \rho_f &= \lambda(1 - \rho_f - \rho)E(B_f), \\ \rho &= \lambda(\rho_f + \rho)E(B), \end{aligned}$$

from which it follows that

$$\rho_f = \frac{\lambda E(B_f)(1 - \lambda E(B))}{1 + \lambda E(B_f) - \lambda E(B)}, \qquad \rho = \frac{\lambda E(B_f)\lambda E(B))}{1 + \lambda E(B_f) - \lambda E(B)}.$$

Together with Little's law we obtain from (5.19) that

$$E(W) = \frac{\rho_f E(R_f) + \rho E(R)}{1 - \lambda E(B)},$$

and for the mean sojourn time we finally get

$$E(S) = E(W) + (1 - \rho_f - \rho)E(B_f) + (\rho_f + \rho)E(B).$$

5.4 A fast and a slow machine

We now revisit the problem studied in exercise 4.6.4. Consider two parallel machines with a common (infinite) buffer where jobs arrive according to a Poisson stream with rate λ. The processing times are exponentially distributed with mean $1/\mu_1$ on machine 1 and $1/\mu_2$ on machine 2 ($\mu_1 > \mu_2$). Jobs are processed in order of arrival. For stability we assume

$$\rho = \frac{\lambda}{\mu_1 + \mu_2} < 1$$

In exercise 4.6.4 we looked at two extreme options: make always use of the slow machine, or do not use it at all. The choice between the two options depends on the workload and the ratio of the processing speeds of the machines. In this section we consider the same model except that jobs are allowed to wait for the busy fast machine at times when the slow one is idle. The problem is to find the best operating policy, i.e. a rule which tells you when to use the slow machine, so as to minimize the mean time spent in the system.

From the point of view of an individual job the decision problem is simple. If an arriving job finds an idle fast machine, he will join it. If he finds a busy fast machine with n jobs and an idle slow machine, then he will go to the slow machine if $(n+1)/\mu_1 > 1/\mu_2$ and otherwise, he will wait for the fast machine. This rule will minimize his expected time in the system. So the *individually optimal policy* is characterized by the integer part of μ_1/μ_2, say N. Jobs wait for the fast machine as long as the number of jobs in the system is less or equal to N.

However we are interested the *overall or socially optimal policy*, i.e., in assigning jobs to the machines so as to minimize the mean time in the system over all jobs. In [30, 48] a proof is given that the overall optimal policy is of the same type as the individual optimal policy, but presumable with a different threshold N. The issue of individually versus socially optimal policies is also studied in [44] for the problem of controlling the arrivals to a single-server queue.

Below we will study the model in which jobs are processed by the slow machine only when the number of jobs in the system is more than N. It will be shown that the equilibrium probabilities for this model, and also quantities such as the mean time spent in the system, can be easily computed. Based on these results the optimal threshold level N can readily be computed.

The state of the system can be described by the pair (n, m) where n is the number of jobs in the system and m denotes the state of machine 2 which is the slower one. If $m = 1$ machine 2 is busy, if $m = 0$ it is idle. The possible states are the pairs $(n, 0)$ with n running from 0 to N and the pairs $(n, 1)$ with $n = 1, 2, \ldots$ The flow diagram of this system is shown in figure 5.5.

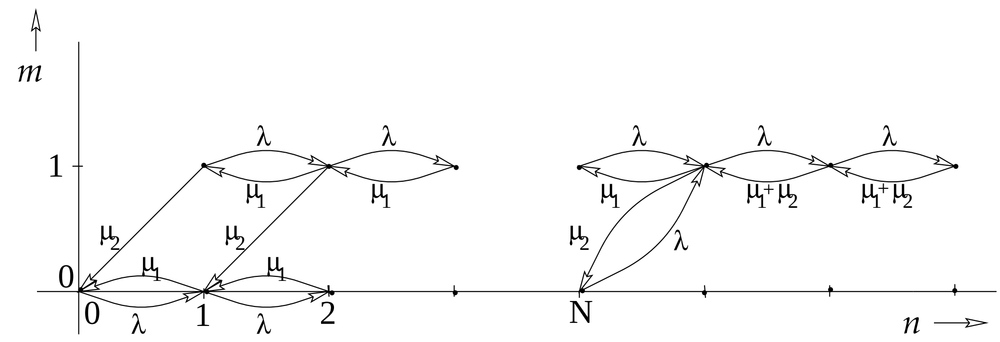

Figure 5.5: Flow diagram for the system with a fast and a slow machine

The equilibrium probabilities $p(n, m)$ satisfy the following set of equations.

$$p(0,0)\lambda = p(1,0)\mu_1 + p(1,1)\mu_2, \tag{5.20}$$

$$p(n,0)(\lambda+\mu_1) = p(n-1,0)\lambda + p(n+1,0)\mu_1 + p(n+1,1)\mu_2, \tag{5.21}$$
$$n = 1, \ldots, N-1,$$

$$
\begin{aligned}
p(N,0)(\lambda+\mu_1) &= p(N-1,0)\lambda + p(N+1,1)\mu_2, && (5.22)\\
p(1,1)(\lambda+\mu_2) &= p(2,1)\mu_1, && (5.23)\\
p(n,1)(\lambda+\mu_1+\mu_2) &= p(n-1,1)\lambda + p(n+1,1)\mu_1, && (5.24)\\
&\qquad n = 2,\dots,N, &&\\
p(N+1,1)(\lambda+\mu_1+\mu_2) &= p(N,1)\lambda + p(N+2,1)(\mu_1+\mu_2) + p(N,0)\lambda, && (5.25)\\
p(n,1)(\lambda+\mu_1+\mu_2) &= p(n-1,1)\lambda + p(n+1,1)(\mu_1+\mu_2), && (5.26)\\
&\qquad n = N+2, N+3, \dots &&
\end{aligned}
$$

To solve these balance equations we first consider the equations (5.26). These equations form a second order recurrence relation the general solution of which is given by

$$p(n,1) = A\rho^n + B, \qquad n = N+1, N+2, \dots \tag{5.27}$$

Since the probabilities $p(n,m)$ have to add up to one, we must set $B = 0$. The constant A is still free and it will be determined later on from the normalization condition. It now remains to solve the (finitely many) balance equations in the states (n,m) with $n \leq N$. Let us first look at the equations (5.24). The general solution of these equations is

$$p(n,1) = C_1 x_1^n + C_2 x_2^n, \qquad n = 1,2,\dots,N+1, \tag{5.28}$$

where x_1 and x_2 are the two roots of

$$x(\lambda+\mu_1+\mu_2) = \lambda + x^2\mu_1.$$

To determine C_1, C_2 note that the forms (5.27) and (5.28) have to agree for $n = N+1$, so

$$C_1 x_1^{N+1} + C_2 x_2^{N+1} = A\rho^{N+1}. \tag{5.29}$$

Further we have to satisfy (5.23) yielding

$$(\lambda - \mu_1 x_1)C_1 + (\lambda - \mu_1 x_2)C_2 = 0. \tag{5.30}$$

Hence, from (5.29) and (5.30) we find

$$
\begin{aligned}
C_1 &= \frac{A\rho^{N+1}(\lambda-\mu_1 x_2)}{(\lambda-\mu_1 x_2)x_1^{N+1} - (\lambda-\mu_1 x_1)x_2^{N+1}}, && (5.31)\\
C_2 &= \frac{A\rho^{N+1}(\lambda-\mu_1 x_1)}{(\lambda-\mu_1 x_1)x_2^{N+1} - (\lambda-\mu_1 x_2)x_1^{N+1}}. &&
\end{aligned}
$$

The equations (5.21) form an second order recurrence relation in $p(n,0)$ with the probabilities $p(n+1,1)$ as inhomogeneous terms. The general solution is

$$p(n,0) = D_1\rho_1^n + D_2 - C_1 x_1^{n+1} - C_2 x_2^{n+1}, \qquad n = 0,1,\dots,N, \tag{5.32}$$

where $\rho_1 = \lambda/\mu_1$. The first two terms are the general solution of the homogeneous relation, the other ones are a particular solution of the inhomogeneous relation. The constants D_1

and D_1 follow form the boundary conditions (5.20) and (5.22). The first condition yields $D_2 = 0$ and the other one

$$D_1\rho_1^{N+1} = C_1 x_1^{N+2} + C_2 x_2^{N+2},$$

whence, by (5.31),

$$D_1 = A\,\frac{(\lambda - \mu_1 x_2)x_1^{N+2} - (\lambda - \mu_1 x_1)x_2^{N+2}}{(\lambda - \mu_1 x_2)x_1^{N+1} - (\lambda - \mu_1 x_1)x_2^{N+1}}\left(\frac{\rho}{\rho_1}\right)^{N+1}. \tag{5.33}$$

To complete the solution we have to solve A from the normalization condition, stating that all probabilities have to add up to one. Substituting the forms (5.27), (5.28) and (5.32) we get

$$A\,\frac{\rho^{N+2}}{1-\rho} + D_1\frac{1-\rho_1^{N+1}}{1-\rho_1} = 1.$$

Inserting the expression for D_1 it follows that

$$A = \left\{\frac{\rho^{N+2}}{1-\rho} + \frac{1-\rho_1^{N+1}}{1-\rho_1}\,\frac{(\lambda - \mu_1 x_2)x_1^{N+2} - (\lambda - \mu_1 x_1)x_2^{N+2}}{(\lambda - \mu_1 x_2)x_1^{N+1} - (\lambda - \mu_1 x_1)x_2^{N+1}}\left(\frac{\rho}{\rho_1}\right)^{N+1}\right\}^{-1}. \tag{5.34}$$

Summarizing we can conclude the equilibrium probabilities are given by

$$\begin{aligned}
p(n,0) &= D_1\rho_1^n - C_1 x_1^{n+1} - C_2 x_2^{n+1}, \qquad n = 0,1,\ldots,N,\\
p(n,1) &= C_1 x_1^n + C_2 x_2^n, \qquad n = 1,2,\ldots,N,\\
p(n,1) &= A\rho^n, \qquad n = N+1, N+2,\ldots
\end{aligned}$$

where the constants A, C_1, C_2, D_1 satisfy (5.34), (5.31) and (5.33), respectively. From these expressions for the equilibrium probabilities it is straightforward to derive an expression for, e.g., the mean number of jobs in the system and the mean time spent in the system.

In table 5.2 we list for fixed $\lambda = 1$ and varying ρ and μ_1/μ_2 the value of the threshold N minimizing the mean time spent in the system. The optimal threshold is found by a simple search procedure. Denote the mean time spent in the system with threshold N by $E(S(N))$. Then $E(S(N))$ is evaluated for increasing values of N until, for the first time, $E(S(N+1)) > E(S(N))$. When this happens the associated N is the optimal one and it will be denoted by N_0. The corresponding mean time spent in the system is denoted by $E(S(N_0))$. Having obtained the optimal system it is interesting to compare its performance with the two systems studied in exercise 4.6.4, i.e., the system which never uses the slow machine and the one which always uses both machines. The performance of the first system is indicated by $E(S_1)$, that of the second one by $E(S_2)$. Note that when $\rho_1 > 1$ the system with the fast machine only is unstable, and in table 5.2 this is indicated by $E(S_1) = \infty$.

We see in table 5.2 that the performance of the optimal system does not change much as μ_1/μ_2 changes while ρ is fixed. Even when μ_1/μ_2 is large, 20 say, substantial savings can be achieved by using the slow machine properly. The practical lesson may be that one

ρ	μ_1/μ_2	N_0	$E(S(N_0))$	$E(S_1)$	$E(S_2)$
0.3	20	14	0.46	0.46	0.96
	10	7	0.49	0.49	0.80
	5	3	0.55	0.56	0.68
	3	2	0.60	0.67	0.63
0.5	20	10	1.10	1.11	1.72
	10	5	1.19	1.22	1.56
	5	3	1.28	1.50	1.41
	3	2	1.32	2.00	1.33
0.8	20	6	4.45	5.25	4.83
	10	3	4.49	7.33	4.72
	5	2	4.48	24.00	4.57
	3	1	4.48	∞	4.48
0.9	20	5	9.56	17.18	9.85
	10	3	9.56	99.00	9.75
	5	2	9.54	∞	9.61
	3	1	9.52	∞	9.52

Table 5.2: Optimal thresholds and system performance where in each example $\lambda = 1$

should not throw away an old machine when the new one is much faster. Also observe that for low values of ρ the system with the fast machine only is just as good as the optimal system, while for high values of ρ the system using both machines is just as good as the optimal one. Finally remark that since the optimal threshold N_0 is (usually) smaller than (the integer part of) μ_1/μ_2 the overall optimal system makes more use of the slow machine than the individually optimal system.

5.5 Exercises

Exercise 5.5.1
Consider the $M/M/1$ model with setup times in section 5.1. Suppose that $\theta = \lambda - \mu$. Show that

$$p(n,1) = (1-\rho)^2 n\rho^n, \qquad n = 1, 2, \ldots$$

is a particular solution of (5.6).

Exercise 5.5.2
Consider the $M/M/1$ model with setup times in section 5.1. Derive expression (5.8) for the probability that the production leadtime is longer than t.

Exercise 5.5.3
Consider a single machine which is turned off when there is no work. It is turned on and restarts work when enough orders, say N, arrived to the machine. The setup times are negligible. The processing times are exponentially distributed with mean $1/\mu$ and the average number of orders arriving per unit time is λ.

(i) How can you describe the state of the system?

(ii) Formulate the balance equations for the (equilibrium) state probabilities.

(iii) Solve the balance equations.

Suppose that $\lambda = 20$ orders per hour, $1/\mu = 2$ minutes and that the setup cost is 54 dollar. In operation the machine costs 12 dollar per minute. The waiting cost is 1 dollar per minute per order.

(iv) Show how for given threshold N the average cost per unit time can be determined.

(v) Compute the threshold N minimizing the average cost per unit time.

Exercise 5.5.4
Consider a channel for the transmission of data messages. The length of the messages are exponentially distributed. Hence also the transfer times are exponentially distributed, with mean $1/\mu$ say. The packets are generated by a source (for instance a computer system) which is alternatingly on and off. The length of an on (resp. off) period is exponentially distributed with parameter θ_1 (resp. θ_0). While the source is on messages are generated with rate λ.
The state of this system can be described by the pair (n, m) where n is the number of messages in the system (waiting for transfer including the one being transferred) and m is the state of the source. If $m = 0$ then the source is off, if $m = 1$ it is on and generating messages.

(i) Describe the flow diagram and give the set of equations for the equilibrium state probabilities.

(ii) Solve these equations and give an expression for the mean delay at the transmission channel.

Exercise 5.5.5 *(Maintenance of trains)*
Consider a maintenance facility for trains. In this facility there is one track for the maintenance of small trains which are used for local distances only. The track offers space to exactly two trains. It can be entered on one side only. This implies that when the train at the front is ready, then it has to wait until the one at the back is also ready before it can leave the track.
Assume that the trains arrive according to a Poisson stream with rate λ and that maintenance of a train takes an exponentially distributed time with mean $1/\mu$.

(i) How can the state of the maintenance facility be described?

(ii) Formulate the balance equations for the state probabilities.

(iii) Under which condition is the facility capable of handling the amount of work offered per unit time?

(iv) Solve the balance equations.

(v) Give an expression for the mean number of trains in the system.

Chapter 6

Priorities

In this chapter we analyse queueing models with different types of customers, where one or more types of customers have priority over other types. More precisely we consider an $M/G/1$ queue with r types of customers. The type i customers arrive according to a Poisson stream with rate λ_i, $i = 1, \dots, r$. The service time and residual service of a type i customer is denoted by B_i and R_i, respectively. The type 1 customers have the highest priority, type 2 customers the second highest priority and so on. We consider two kinds of priorities. For one priority rule higher priority customers may not interrupt the service time of a lower priority customer, but they have to wait till the service time of the low priority customer has been completed. For the other rule interruptions are allowed and after the interruption the service time of the lower priority customer resumes at the point where it was interrupted. The first priority rule is called non-preemptive, the second one preemptive-resume. In the following two sections we show how the mean waiting times can be found for the two kinds of priorities.

6.1 Non-preemptive priority

The mean waiting time of a type i customer is denoted by $E(W_i)$ and $E(L_i^q)$ is the number of type i customers waiting in the queue. Further define $\rho_i = \lambda_i E(B_i)$. For the highest priority customers it holds that

$$E(W_1) = E(L_1^q)E(B_1) + \sum_{j=1}^{r} \rho_j E(R_j).$$

According to Little's law we have

$$E(L_1^q) = \lambda_1 E(W_1)$$

Combining the two equations yields

$$E(W_1) = \frac{\sum_{j=1}^{r} \rho_j E(R_j)}{1 - \rho_1}. \tag{6.1}$$

The determination of the mean waiting time for the lower priority customers is more complicated. Consider type i customers with $i > 1$. The waiting time of a type i customer can be divided in a number of portions. The first portion for which the customer has to wait is the amount of work on arrival. Call this portion X_1. The second portion, say X_2, is the amount of higher priority work arriving during X_1. Subsequently the third portion X_3 is the amount of higher priority work arriving during X_2, and so on. A realization of the waiting time for a type 2 customer is shown in figure 6.1. The increments of the amount of work are to the service times of the arriving type 1 customers.

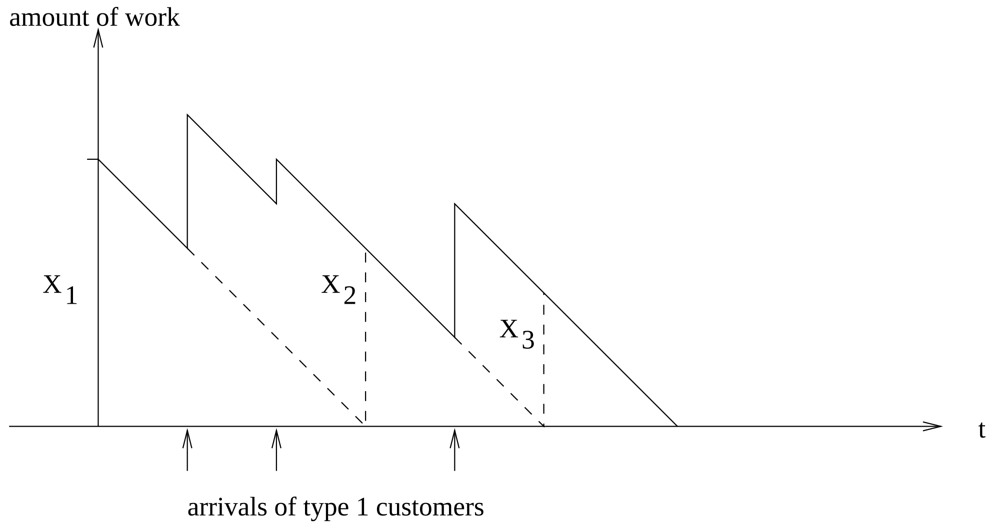

Figure 6.1: Realization of the waiting time of a type 2 customer

Hence the mean waiting time is given by

$$E(W_i) = E(X_1 + X_2 + X_3 + \cdots) = \sum_{k=1}^{\infty} E(X_k).$$

On arrival a type i customer has to wait for all jobs in the queue with the same or higher priority and for the amount of work at the server. So

$$E(X_1) = \sum_{j=1}^{i} E(L_j^q)E(B_j) + \sum_{j=1}^{r} \rho_j E(R_j).$$

Note that X_{k+1} depends on X_k. To determine $E(X_{k+1})$ we therefore condition on the length of X_k. Denote the density of X_k by $f_k(x)$. Then it follows that

$$\begin{aligned} E(X_{k+1}) &= \int_{x=0}^{\infty} E(X_{k+1}|X_k = x) f_k(x)dx \\ &= \int_{x=0}^{\infty} (\lambda_1 x E(B_1) + \cdots + \lambda_{i-1} x E(B_{i-1})) f_k(x)dx \\ &= (\rho_1 + \cdots + \rho_{i-1})E(X_k). \end{aligned}$$

Repeated application of the relation above yields

$$E(X_{k+1}) = (\rho_1 + \cdots + \rho_{i-1})^k E(X_1), \qquad k = 0, 1, 2, \ldots$$

Hence we find for $i = 2, \ldots, r$

$$E(W_i) = \frac{E(X_1)}{1 - (\rho_1 + \cdots + \rho_{i-1})} = \frac{\sum_{j=1}^{i} E(L_j^q)E(B_j) + \sum_{j=1}^{r} \rho_j E(R_j)}{1 - (\rho_1 + \cdots + \rho_{i-1})}, \tag{6.2}$$

An intuitive argument (which can be made rigorous) to directly obtain the above equation is by observing that the waiting time is equal to the amount of work on arrival plus all the higher priority work arriving during the waiting time. So

$$E(W_i) = E(X_1) + \sum_{j=1}^{i-1} \lambda_j E(W_i) E(B_j),$$

from which equation (6.2) immediately follows. Substitution of Little's law

$$E(L_i^q) = \lambda_i E(W_i)$$

into equation (6.2) yields

$$\begin{aligned} (1 - (\rho_1 + \ldots + \rho_i))E(W_i) &= \sum_{j=1}^{i-1} E(L_j^q)E(B_j) + \sum_{j=1}^{r} \rho_j E(R_j) \\ &= (1 - (\rho_1 + \ldots + \rho_{i-2}))E(W_{i-1}). \end{aligned}$$

Repeatedly applying this equation and using (6.1) finally leads to

$$E(W_i) = \frac{\sum_{j=1}^{r} \rho_j E(R_j)}{(1 - (\rho_1 + \ldots + \rho_i))(1 - (\rho_1 + \ldots + \rho_{i-1}))}, \qquad i = 1, \cdots, r. \tag{6.3}$$

The mean sojourn time $E(S_i)$ of a type i customer follows from

$$E(S_i) = E(W_i) + E(B_i).$$

6.2 Preemptive-resume priority

We will show that the results in case the service times may be interrupted easily follow from the ones in the previous section.

Consider a type i customer. For a type i customer there do not exist lower priority customers (they can be considered as a hobby of the server). So we henceforth assume that $\lambda_{i+1} = \cdots = \lambda_r = 0$.

The waiting time can again be divided into portions $X_1, X_2, \ldots$. Now observe that the amount of work in the system does not depend on the order in which the customers are served. Hence it is exactly the same as in the system with non-preemptive priorities.

So $X_1, X_2, \ldots$, and thus also W_i are stochastically the same as in the system with non-preemptive priorities (and with $\lambda_{i+1} = \cdots = \lambda_r = 0$). From (6.3) we obtain

$$E(W_i) = \frac{\sum_{j=1}^{i} \rho_j E(R_j)}{(1-(\rho_1+\ldots+\rho_i))(1-(\rho_1+\ldots+\rho_{i-1}))}, \qquad i = 1, \cdots, r.$$

For the mean sojourn time we have to add the service time plus all the interruptions of higher priority customers during the service time. The mean of such a generalized service time can be found along the same lines as (6.2), yielding

$$\frac{E(B_i)}{1-(\rho_1+\cdots+\rho_{i-1})}.$$

So the mean sojourn time of a type i customer is given by

$$E(S_i) = \frac{\sum_{j=1}^{i} \rho_j E(R_j)}{(1-(\rho_1+\ldots+\rho_i))(1-(\rho_1+\ldots+\rho_{i-1}))} + \frac{E(B_i)}{1-(\rho_1+\cdots+\rho_{i-1})},$$

for $i = 1, \cdots, r$.

6.3 Shortest processing time first

In production systems one often processes jobs according to the shortest processing time first rule (SPTF). The mean production lead time in a single machine system can be found using the results in section 6.1.

Consider an $M/G/1$ queue with arrival rate λ and service times B with density $f_B(x)$. Assume that $\rho = \lambda E(B) < 1$. The server works according to the SPTF rule. That is, after a service completion, the next customer to be served is the one with the shortest service time. So in this system the small ones have priority.

Define type x customers as the ones with a service time between x and $x + dx$. The mean waiting time of a type x customer is denoted by $E(W_x)$ and ρ_x is the fraction of time the server helps type x customers, so

$$\rho_x = (\lambda f_B(x)dx)x = \lambda x f_B(x)dx. \tag{6.4}$$

From (6.3) we obtain

$$\begin{aligned} E(W_x) &= \frac{\rho E(R)}{(1-\int_{y=0}^{y=x} \rho_y)^2} \\ &= \frac{\rho E(R)}{(1-\lambda\int_{y=0}^{y=x} y f_B(y)dy)^2}. \end{aligned}$$

Hence the mean overall waiting time is given by

$$\begin{aligned} E(W) &= \int_{x=0}^{\infty} E(W_x) f_B(x)dx \\ &= \rho E(R) \int_{x=0}^{\infty} \frac{f_B(x)dx}{(1-\lambda\int_{y=0}^{y=x} y f_B(y)dy)^2}. \end{aligned} \tag{6.5}$$

In table 6.1 we compare the SPTF rule with the usual first-come first-served (FCFS) rule for an $M/M/1$ system with mean service time 1. The results show that considerable reductions in the mean waiting time are possible.

	$E(W)$	
ρ	FCFS	SPTF
0.5	1	0.713
0.8	4	1.883
0.9	9	3.198
0.95	19	5.265

Table 6.1: The mean waiting time for FCFS and SPTF in an $M/M/1$ with mean service time 1

6.4 A conservation law

In this section we consider a single-server queue with r types of customers. The type i customers arrive according to a general arrival stream with rate λ_i, $i = 1, \ldots, r$. The mean service time and mean residual service time of a type i customer is denoted by $E(B_i)$ and $E(R_i)$, respectively. Define $\rho_i = \lambda_i E(B_i)$. We assume that

$$\sum_{i=1}^{r} \lambda_i E(B_i) < 1,$$

so that the server can handle the amount of work offered per unit of time. The customers are served according to a service discipline which does not allow interruptions of the service times. So, for example, the customers may be served according to FCFS, SPTF, random or a non-preemptive priority rule. Below we derive a conservation law for the mean waiting times of the r types of customers, which expresses that a weighted sum of these waiting times is independent of the service discipline.

Let $E(V(P))$ and $E(L_i^q(P))$ denote the mean amount of work in the system and the mean number of type i customers waiting in the queue, respectively, for discipline P. The mean amount of work in the system is given by

$$E(V(P)) = \sum_{i=1}^{r} E(L_i^q(P))E(B_i) + \sum_{i=1}^{r} \rho_i E(R_i). \tag{6.6}$$

The first sum at the right-hand side is the mean amount of work in the queue, and the second one is the mean amount of work at the server. Clearly the latter does not depend on the discipline P.

The crucial observation is that the amount of work in the system does not depend on the order in which the customers are served. The amount of work decreases with one unit per unit of time independent of the customer being served and when a new customer arrives the amount of work is increased by the service time of the new customer. Hence, the amount of work does not depend on P. Thus from equation (6.6) and Little's law

$$E(L_i^q) = \lambda_i E(W_i(P)),$$

we obtain the following conservation law for the mean waiting times,

$$\sum_{i=1}^{r} \rho_i E(W_i(P)) = \text{constant (independent of } P).$$

Below we present two examples where this law is used.

Example 6.4.1 *(M/G//1 with FCFS and SPTF)*
We check whether for the $M/G/1$ the weighted sum of the mean waiting times for SPTF is indeed the same as for FCFS.

In case the customers are served in order of arrival it holds that (see (4.43))

$$\rho E(W) = \frac{\rho^2 E(R)}{1-\rho}.$$

When the server works according to the SPTF rule we have (see (6.4) and (6.5))

$$\begin{aligned}
\int_{x=0}^{\infty} \rho_x E(W_x) &= \int_{x=0}^{\infty} \frac{\lambda x f_B(x) dx \rho E(R)}{(1-\lambda\int_{y=0}^{y=x} y f_B(y)dy)^2} \\
&= \left. \frac{\rho E(R)}{1-\lambda\int_{y=0}^{y=x} y f_B(y)dy} \right|_{x=0}^{\infty} \\
&= \frac{\rho^2 E(R)}{1-\rho},
\end{aligned}$$

which indeed is the same as for the FCFS rule.

Example 6.4.2 *(M/G/1 with non-preemptive priority)*
Consider an $M/G/1$ queue with two types of customers. The type 1 customers have non-preemptive priority over the type 2 customers. The mean waiting time for the type 1 customers is given by (6.1). We now derive the mean waiting time for the type 2 customers by using the conservation law. According to this law it holds that

$$\rho_1 E(W_1) + \rho_2 E(W_2) = C, \tag{6.7}$$

where C is some constant independent of the service discipline. To determine C consider the FCFS discipline. For FCFS it follows from (4.43) that

$$E(W_1) = E(W_2) = \frac{\rho_1 E(R_1) + \rho_2 E(R_2)}{1-\rho_1-\rho_2}.$$

Hence,

$$C = (\rho_1+\rho_2)\frac{\rho_1 E(R_1) + \rho_2 E(R_2)}{1-\rho_1-\rho_2}. \tag{6.8}$$

By substituting (6.1) and (6.8) into equation (6.7) we retrieve formula (6.3) for the mean waiting time of the type 2 customers under the non-preemptive priority rule.

6.5 Exercises

Exercise 6.5.1
Consider a machine where jobs arrive according to a Poisson stream with a rate of 4 jobs per hour. Half of the jobs has a processing time of exactly 10 minutes, a quarter of the jobs has a processing time of exactly 15 minutes and the remaining quarter has a processing time of 20 minutes. The jobs with a processing time of 10 minutes are called type 1 jobs, the ones with a processing time of 15 minutes type 2 jobs and the rest type 3 jobs. The jobs are processed in order of arrival.

(i) Determine the mean sojourn time (waiting time plus processing time) of a type 1, 2 and 3 job and also of an arbitrary job.

One decides to process smaller jobs with priority. So type 1 orders have highest priority, type 2 orders second highest priority and type 3 orders lowest priority.
Answer question (i) for the following two cases:

(ii) Jobs processed at the machine may not be interrupted.

(iii) Type 1 and type 2 jobs may interrupt the processing of a type 3 job. Type 1 jobs may not interrupt the processing of a type 2 job.

Bibliography

[1] I.J.B.F. Adan, *A compensation approach for queueing problems*, CWI Tract 104, Stichting Mathematisch Centrum, Amsterdam, 1994.

[2] I.J.B.F. Adan, J. Wessels, W.H.M. Zijm, *Analysis of the symmetric shortest queue problem*, Stochastic Models, 6 (1990), pp. 691–713.

[3] I.J.B.F. Adan, J. Wessels, W.H.M. Zijm, *A compensation approach for two-dimensional Markov processes*, Adv. Appl. Prob., 25 (1993), pp. 783–817.

[4] I.Adan, Y.Zhao, *Analyzing $GI/E_r/1$ queues*, Opns. Res. Lett., 19 (1996), pp. 183–190.

[5] I.J.B.F. Adan, W.A. van de Waarsenburg, J. Wessels, *Analyzing $E_k|E_r|c$ queues*, EJOR, 92 (1996), pp. 112–124.

[6] D. Anick, D. Mitra, M.M. Sondhi, *Stochastic theory of a data handling system*, Bell System Tech. J., 61 (1982), pp. 1871–1894.

[7] D. Bertsimas, *An analytic approach to a general class of $G/G/s$ queueing systems*, Opns. Res., 38 (1990), pp. 139–155.

[8] D. Bertsimas, *An exact FCFS waiting time analysis for a class of $G/G/s$ queueing systems*, QUESTA, 3 (1988), pp. 305–320.

[9] D. Bertsimas, X.A. Papaconstantinou *Analysis of the stationary $E_k/C_2/s$ queueing system*, EJOR, 37 (1988), pp. 272–287.

[10] O.J. Boxma, G.J. van Houtum, *The compensation approach applied to a 2×2 switch*, Prob. Engineer. Inform. Sci., 7 (1993), pp. 471–493.

[11] J.A. Buzacott, J.G. Shanthikumar, *Stochastic models of manufacturing systems*, Prentice Hall, 1993.

[12] A.S. Carr, A.R. Güllü, P.L. Jackson, J.A. Muckstadt, *An exact analysis of a production-inventory strategy for industrial suppliers*, Working paper, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1993.

[13] E.A. Coddington, N. Levinson, *Theory of ordinary differential equations*, McGraw-Hill, London, 1955.

[14] J.W. Cohen, O.J. Boxma, *Boundary value problems in queueing system analysis*, North-Holland, Amsterdam, 1983.

[15] M.J.A. van Eenige, *Queueing systems with periodic service*, Thesis, Eindhoven University of Technology, Eindhoven, 1996.

[16] G. Fayolle, *Méthodes analytiques pour les files d'attente coupleés*, Thesis, Univ. de Paris VI, Paris ,1979.

[17] G. Fayolle, R. Iasnogorodski, *Two coupled processors: the reduction to a Riemann–Hilbert problem*, Z. Wahrscheinlichkeitsth. Verw. Gebiete, 47 (1979), pp. 325–351.

[18] A. Federgruen, Z. Katalan, *Make-to-stock or make-to-order: that is the question; novel answers to an ancient debate*, Working paper, Graduate School of Business, Columbia University, New York, 1994.

[19] L. Flatto, H.P. McKean, *Two queues in parallel*, Comm. Pure Appl. Math., 30 (1977), pp. 255–263.

[20] P.R. Halmos, *Finite-dimensional vector spaces*, Van Nostrand, London, 1958.

[21] M.C. van der Heijden, *Performance analysis for reliability and inventory models*, Thesis, Vrije Universiteit, Amsterdam, 1993.

[22] A. Hordijk, G. Koole, *On the optimality of the generalized shortest queue policy*, Prob. Engineer. Inform. Sci., 4 (1990), pp. 477–487.

[23] R. Iasnogorodski, *Problèmes-frontières dans les files d'attente*, Thesis, Univ. de Paris IV, Paris, 1979.

[24] S. Jaffe, *The equilibrium distribution for a clocked buffered switch*, Prob. Engineer. Inform. Sci., 6 (1992), pp. 425–438.

[25] M.A. Johnson, *An emperical study of queueing approximations based on phase-type approximations*, Stochastic Models, 9 (1993), pp. 531–561.

[26] J.F.C. Kingman, *Two similar queues in parallel*, Ann. Math. Statist., 32 (1961), pp. 1314–1323.

[27] L. Kleinrock, *Queueing Systems, Vol. I: Theory*. Wiley, New York, 1975.

[28] S. Lang, *Linear algebra*, Addison-Wesley, new York, 1978.

[29] G. Latouche, V. Ramaswami *A logarithmic reduction algorithm for quasi-birth-death processes*, J. Appl. Prob., 30 (1993), pp. 650–674.

[30] W. Lin, P.R. Kumar, *Optimal control of a queueing system with two heterogeneous servers*, Mathematics Research Report No. 82–83, Dept. of Math., University of maryland, Baltimore County, 1982.

[31] J.D. Little, *A proof of the queueing formula* $L = \lambda W$, Opns. Res., 9 (1961), pp. 383–387.

[32] R.A. Marie, *Calculating equilibrium probabilities for* $\lambda(n)/C_k/1/N$ *queue*, in: Proceedings Performance' 80, Toronto, (May 28–30, 1980), pp. 117–125.

[33] D. Mitra, *Stochastic theory of a fluid model of producers and consumers coupled by a buffer*, Adv. Appl. Prob., 20 (1988), pp. 646–676.

[34] M.F. Neuts, *Matrix-geometric solutions in stochastic models*, Johns Hopkins University Press, Baltimore, 1981.

[35] V. Ramaswami, G. Latouche *A general class of Markov processes with explicit matrix-geometric solutions*, OR Spektrum, 8 (1986), pp. 209–218.

[36] S.M. Ross, *Introduction to probability models*, 6th ed., Academic Press, London, 1997.

[37] M. Rubinovitch, *The slow server problem*, J. Appl. Prob., 22 (1985), pp. 205–213.

[38] M. Rubinovitch, *The slow server problem: a queue with stalling*, J. Appl. Prob., 22 (1985), pp. 879–892.

[39] R.S. Schassberger, *On the waiting time in the queueing system* $GI/G/1$, Ann. Math. Statist., 41 (1970), pp. 182–187.

[40] R.S. Schassberger, *Warteschlangen,* Springer-Verlag, Berlin, 1973.

[41] S. Shapiro, *The M-server queue with Poisson input and Gamma-distributed service of order two*, Opns. Res., 14 (1966), pp. 685–694.

[42] C.R. Sox, L.J. Thomas, J.O. McClain, *Jellybeans: Sorting capacity to improve service*, Working paper, Department of Industrial Engineering, Auburn University, Alabama, 1994.

[43] S. Stidham, *A last word on* $L = \lambda W$, Opns. Res., 22 (1974), pp. 417–421.

[44] S. Stidham, *Socially and individually optimal control of arrivals to a* $GI/M/1$ *queue*, Management Sci., 24 (1978), pp. 1598–1610.

[45] Y.Takahashi, *Asymptotic exponentiality of the tail of the waiting time distribution in a $Ph/Ph/1$ queue*, Adv. Appl. Prob., 13 (1981), pp. 619–630.

[46] H.C. Tijms, *Stochastic models: a algorithmic approach*, John Wiley & Sons, Chichester, 1994.

[47] H.C. Tijms, M.C.T. Van de Coevering, *A simple numerical approach for infinite-state Markov chains*, Prob. Engineer. Inform. Sci., 5 (1991), pp. 285–295.

[48] J. Walrand, *A note on 'Optimal control of a queueing system with two heterogeneous servers'*, Technical Report, Dept. of Electrical Enigeering and Computer Sciences and Electronic Research Laboratory, University of California, Berkeley, 1983.

[49] R.W. Weber, *On the optimal assignment of customers to parallel servers*, J. Appl. Prob., 15 (1978), pp. 406–413.

[50] W. Whitt *Approximating a point process by a renewal process I: two basic methods*, Opns. Res., 30 (1986), pp. 125–147.

[51] W. Whitt *Deciding which queue to join: some counterexamples*, Opns. Res., 34 (1986), pp. 55–62.

[52] T.M. Williams, *Special products and uncertainty in production/inventory systems*, EJOR, 15 (1984), pp. 46–54.

[53] W. Winston, *Optimality of the shortest line discipline*, J. Appl. Prob., 14 (1977), pp. 181–189.

[54] R.W. Wolff, *Poisson arrivals see time averages*, Opns. Res., 30 (1982), pp. 223–231.

[55] R.W. Wolff, *Stochastic modeling and the theory of queues*, Prentice-Hall, London, 1989.

[56] J. Wijngaard, *The effect of interstage buffer storage on the output of two unreliable production units in series, with different production rates*, AIIE Trans., 11 (1979), pp. 42–47.

Index