heavily on input from measurement results. Model parameters are frequently derived from measurements of earlier studies on similar systems also in the system development scenario. Conversely, measurement studies can often be better planned and executed if they are complemented and guided by a model-based evaluation.

1.3 BASICS OF PROBABILITY AND STATISTICS

We begin by giving a brief overview of the more important definitions and results of probability theory. The reader can find additional details in books such as [Alle90, Fell68, Koba78, Triv01]. We assume that the reader is familiar with the basic properties and notations of probability theory.

1.3.1 Random Variables

A random variable is a function that reflects the result of a random experiment. For example, the result of the experiment "toss a single die" can be described by a random variable that can assume the values one through six. The number of requests that arrive at an airline reservation system in one hour or the number of jobs that arrive at a computer system are also examples of a random variable. So is the time interval between the arrivals of two consecutive jobs at a computer system, or the throughput in such a system. The latter two examples can assume continuous values, whereas the first two only assume discrete values. Therefore, we have to distinguish between continuous and discrete random variables.

1.3.1.1 Discrete Random Variables A random variable that can only assume discrete values is called a *discrete random variable*, where the discrete values are often non-negative integers. The random variable is described by the possible values that it can assume and by the probabilities for each of these values. The set of these probabilities is called the *probability mass function* (pmf) of this random variable. Thus, if the possible values of a random variable X are the non-negative integers, then the pmf is given by the probabilities:

$$p_k = P(X = k), \quad \text{for} \quad k = 0, 1, 2 \ldots, \tag{1.1}$$

the probability that the random variable X assumes the value k.

The following is required:

$$P(X = k) \geq 0, \qquad \sum_{\text{all } k} P(X = k) = 1.$$

For example, the following pmf results from the experiment "toss a single die":

$$P(X = k) = \frac{1}{6}, \quad \text{for } k = 1, 2, \ldots, 6.$$

The following are other examples of discrete random variables:

- Bernoulli random variable: Consider a random experiment that has two possible outcomes, such as tossing a coin ($k = 0, 1$). The pmf of the random variable X is given by

$$P(X = 0) = 1 - p \quad \text{and} \quad P(X = 1) = p, \quad \text{with } 0 < p < 1. \tag{1.2}$$

- Binomial random variable: The experiment with two possible outcomes is carried out n times where successive trials are independent. The random variable X is now the number of times the outcome 1 occurs. The pmf of X is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n. \tag{1.3}$$

- Geometric random variable: The experiment with two possible outcomes is carried out several times, where the random variable X now represents the number of trials it takes for the outcome 1 to occur (the current trial included). The pmf of X is given by

$$P(X = k) = p(1-p)^{k-1}, \qquad k = 1, 2, \ldots. \tag{1.4}$$

- Poisson random variable: The probability of having k events (Poisson pmf) is given by

$$P(X = k) = \frac{(\alpha)^k}{k!} \cdot \mathrm{e}^{-\alpha}, \qquad k = 0, 1, 2, \ldots; \alpha > 0. \tag{1.5}$$

The Poisson and geometric random variables are very important to our topic; we will encounter them very often. Several important parameters can be derived from a pmf of a discrete random variable:

- Mean value or expected value:

$$\overline{X} = E[X] = \sum_{\text{all } k} k \cdot P(X = k). \tag{1.6}$$

The function of a random variable is another random variable with the expected value of

$$E[f(X)] = \sum_{\text{all } k} f(k) \cdot P(X = k). \tag{1.7}$$

- nth moments:

$$\overline{X^n} = E[X^n] = \sum_{\text{all } k} k^n \cdot P(X = k), \tag{1.8}$$

that is, the nth moment is the expected value of the nth power of X. The first moment of X is simply the mean of X.

- nth central moment:

$$\overline{(X-\overline{X})^n} = E[(X-E[X])^n] = \sum_{\text{all } k}(k-\overline{X})^n \cdot P(X=k)\,. \tag{1.9}$$

The nth central moment is the expected value of the nth power of the difference between X and its mean. The first central moment is equal to zero.

- The second central moment is called the variance of X:

$$\sigma_X^2 = \mathrm{var}(X) = \overline{(X-\overline{X})^2} = \overline{X^2} - \overline{X}^2\,, \tag{1.10}$$

where σ_X is called the standard deviation.

- The coefficient of variation is the normalized standard deviation:

$$c_X = \frac{\sigma_X}{\overline{X}}. \tag{1.11}$$

- The second moment can be easily obtained from the coefficient of variation:

$$\overline{X^2} = \overline{X}^2 \cdot (1+c_X^2)\,. \tag{1.12}$$

Information on the average deviation of a random variable from its expected value is provided by c_X, σ_X, and $\mathrm{var}(X)$. If $c_X = \sigma_X = \mathrm{var}(X) = 0$, then the random variable assumes a fixed value with probability one.

Table 1.1 Properties of several discrete random variables

Random Variables	Parameter	$\overline{X}$	var(X)	c_X^2
Bernoulli	p	p	$p(1-p)$	$\frac{1-p}{p}$
Binomial	n, p	np	$np(1-p)$	$\frac{1-p}{np}$
Geometric	p	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$1-p$
Poisson	α	α	α	$\frac{1}{\alpha}$

Table 1.1 gives a list of random variables, their mean values, their variances, and the squared coefficients of variation for some important discrete random variables.

1.3.1.2 Continuous Random Variables A random variable X that can assume all values in the interval $[a, b]$, where $-\infty \leq a < b \leq +\infty$, is called a *continuous random variable*. It is described by its *distribution function* (also called CDF or cumulative distribution function):

$$F_X(x) = P(X \leq x)\,, \tag{1.13}$$

which specifies the probability that the random variable X takes values less than or equal to x, for every x.

From Eq. (1.13) we get for $x < y$:

$$F_X(x) \leq F_X(y)\,, \qquad P(x < X \leq y) = F_X(y) - F_X(x)\,.$$

The *probability density function* (pdf) $f_X(x)$ can be used instead of the distribution function, provided the latter is differentiable:

$$f_X(x) = \frac{\mathrm{d}\,F_X(x)}{\mathrm{d}\,x}. \tag{1.14}$$

Some properties of the pdf are

$$f_X(x) \geq 0 \quad \text{for all } x, \qquad \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}\,x = 1\,,$$

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)\,\mathrm{d}\,x\,, \qquad P(X = x) = \int_{x}^{x} f_X(x)\,\mathrm{d}\,x = 0\,,$$

$$P(X > x_3) = \int_{x_3}^{\infty} f_X(x)\,\mathrm{d}\,x\,.$$

Note that for so-called *defective* random variables [STP96] we have

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}\,x < 1\,.$$

The density function of a continuous random variable is analogous to the pmf of a discrete random variable. The formulae for the mean value and moments of continuous random variables can be derived from the formulae for discrete random variables by substituting the pmf by the pdf and the summation by an integral:

- Mean value or expected value:

$$\overline{X} = E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}\,x \tag{1.15}$$

and

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\,\mathrm{d}\,x. \tag{1.16}$$

- nth moment:

$$\overline{X^n} = E[X^n] = \int_{-\infty}^{\infty} x^n \cdot f_X(x)\,\mathrm{d}\,x. \tag{1.17}$$

- nth central moment:

$$\overline{(X - \overline{X})^n} = E[(X - E[X])^n] = \int_{-\infty}^{\infty} (x - \overline{X})^n f_X(x)\,\mathrm{d}\,x. \tag{1.18}$$

- Variance:

$$\sigma_X^2 = \mathrm{var}(X) = \overline{(X - \overline{X})^2} = \overline{X^2} - \overline{X}^2, \tag{1.19}$$

with σ_X as the standard deviation.

- Coefficient of variation:

$$c_X = \frac{\sigma_X}{\overline{X}}. \tag{1.20}$$

A very well-known and important continuous distribution function is the *normal distribution*. The CDF of a normally distributed random variable X is given by

$$F_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{x} \exp\left(-\frac{(u - \overline{X})^2}{2\sigma_X^2}\right) \mathrm{d}\,u, \tag{1.21}$$

and the pdf by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(x - \overline{X})^2}{2\sigma_X^2}\right).$$

The standard normal distribution is defined by setting $\overline{X} = 0$ and $\sigma_X = 1$:

$$\text{CDF: } \Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}\,u, \tag{1.22}$$

$$\text{pdf: } \quad \phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right).$$

A plot of the preceding pdf is shown in Fig. 1.4.

For an arbitrary normal distribution we have

$$F_X(x) = \Phi\left(\frac{x - \overline{X}}{\sigma_X}\right) \quad \text{and} \quad f_X(x) = \phi\left(\frac{x - \overline{X}}{\sigma_X}\right),$$

respectively.

Other important continuous random variables are described as follows.

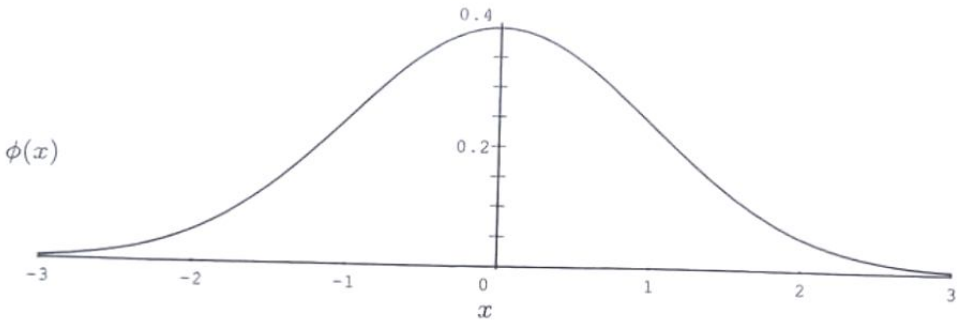

Fig. 1.4 pdf of the standard normal random variable.

(a) Exponential Distribution

The exponential distribution is the most important and also the easiest to use distribution in queueing theory. Interarrival times and service times can often be represented exactly or approximately using the exponential distribution. The CDF of an exponentially distributed random variable X is given by Eq. (1.23):

$$F_X(x) = \begin{cases} 1 - \exp\left(-\dfrac{x}{\overline{X}}\right), & 0 \le x < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{with } \overline{X} = \begin{cases} \dfrac{1}{\lambda}, & \text{if } X \text{ represents interarrival times,} \\ \dfrac{1}{\mu}, & \text{if } X \text{ represents service times.} \end{cases} \tag{1.23}$$

Here λ or μ denote the parameter of the random variable. In addition, for an exponentially distributed random variable with parameter λ the following relations hold:

$$\begin{aligned} \text{pdf:} \quad & f_X(x) = \lambda \mathrm{e}^{-\lambda x}, \\ \text{mean:} \quad & \overline{X} = \frac{1}{\lambda}, \\ \text{variance:} \quad & \mathrm{var}(X) = \frac{1}{\lambda^2}, \\ \text{coefficient of variation:} \quad & c_X = 1. \end{aligned}$$

Thus, the exponential distribution is completely determined by its mean value.

The importance of the exponential distribution is based on the fact that it is the only continuous distribution that possesses the memoryless property:

$$P(X \leq u + t \mid X > u) = 1 - \exp\left(-\frac{t}{\overline{X}}\right) = P(X \leq t). \tag{1.24}$$

As an example for an application of Eq. (1.24), consider a bus stop with the following schedule: Buses arrive with exponentially distributed interarrival times and identical mean $\overline{X}$. Now if you have already been waiting in vain for u units of time for the bus to come, the probability of a bus arrival within the next t units of time is the same as if you had just shown up at the bus stop, that is, you can forget about the past or about the time already spent waiting.

Another important property of the exponential distribution is its relation to the discrete Poisson random variable. If the interarrival times are exponentially distributed and successive interarrival times are independent with identical mean $\overline{X}$, then the random variable that represents the number of buses that arrive in a fixed interval of time $[0, t)$ has a Poisson distribution with parameter $\alpha = t/\overline{X}$.

Two additional properties of the exponential distribution can be derived from the Poisson property:

1. If we merge n Poisson processes with distributions for the interarrival times $1 - e^{-\lambda_i t}$, $1 \leq i \leq n$, into one single process, then the result is a Poisson process for which the interarrival times have the distribution $1 - e^{-\lambda t}$ with $\lambda = \sum_{i=1}^{n} \lambda_i$ (see Fig. 1.5).

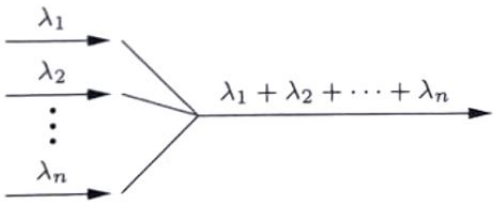

Fig. 1.5 Merging of Poisson processes.

2. If a Poisson process with interarrival time distribution $1 - e^{-\lambda t}$ is split into n processes so that the probability that the arriving job is assigned to the ith process is q_i, $1 \leq i \leq n$, then the ith subprocess has an interarrival time distribution of $1 - e^{-q_i \lambda t}$, i.e., n Poisson processes have been created, as shown in Fig 1.6.

The exponential distribution has many useful properties with analytic tractability, but is not always a good approximation to the observed distribution. Experiments have shown deviations. For example, the coefficient of variation of the service time of a processor is often greater than one, and for a peripheral device it is usually less than one. This observed behavior leads directly to the need to consider the following other distributions:

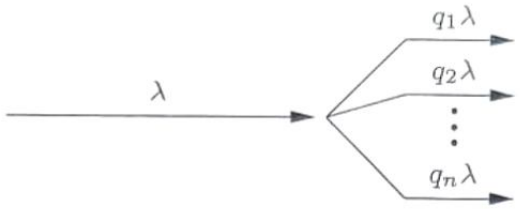

Fig. 1.6 Splitting of a Poisson process.

(b) Hyperexponential Distribution, H_k

This distribution can be used to approximate empirical distributions with a coefficient of variation larger than one. Here k is the number of phases.

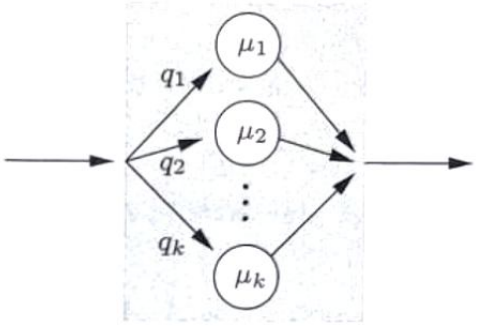

Fig. 1.7 A random variable with H_k distribution.

Figure 1.7 shows a model with hyperexponentially distributed time. The model is obtained by arranging k phases with exponentially distributed times and rates $\mu_1, \mu_2, \ldots, \mu_k$ in parallel. The probability that the time span is given by the jth phase is q_j, where $\sum_{j=1}^{k} q_j = 1$. However, only one phase can be occupied at any time. The resulting CDF is given by

$$F_X(x) = \sum_{j=1}^{k} q_j(1 - \mathrm{e}^{-\mu_j x}), \qquad x \geq 0. \tag{1.25}$$

In addition, the following relations hold:

$$\text{pdf:} \quad f_X(x) = \sum_{j=1}^{k} q_j \mu_j \, \mathrm{e}^{-\mu_j x}, \quad x > 0,$$

$$\text{mean:} \quad \overline{X} = \sum_{j=1}^{k} \frac{q_j}{\mu_j} = \frac{1}{\mu}, \quad x > 0,$$

$$\text{variance:} \quad \mathrm{var}(X) = 2\sum_{j=1}^{k} \frac{q_j}{\mu_j^2} - \frac{1}{\mu^2},$$

time-scale; for example, the histogram of packets per time unit has the same shape if a histogram is built over 1000 seconds or 10 seconds if the time-step for building the classes in the histogram is changed in the same scale. In contrast to this, if a Poisson process is assumed, peaks are flattened out when the time unit is enlarged. This implies, that a long-range dependence may play an important role when Internet traffic is analyzed. However, there are some recent papers [Down01] that question the self-similarity in specific situations.

Heavy-tailed distributions can be used to model self-similar processes and the Pareto distribution is one of the most common heavy-tailed distributions. For details on self-similarity the reader is referred to [LTWW94] and for more information about the effect of heavy-tails to [BBQH03].

(j) Lognormal Distribution

The lognormal distribution is given by

$$F_X(x) = \Phi\left(\frac{\ln(x) - \lambda}{\alpha}\right), \quad x > 0\,. \tag{1.42}$$

Additionally, the following relations hold:

$$\text{pdf:} \quad f_X(x) = \frac{1}{\alpha x\sqrt{2\pi}} \exp(-\{\ln(x) - \lambda\}^2/2\alpha^2)\,, \quad x > 0\,,$$

$$\text{mean:} \quad \overline{X} = \exp(\lambda + \alpha^2/2)\,,$$

$$\text{squared coefficient of variation:} \quad c_X^2 = \exp(\alpha^2) - 1\,,$$

where the shape parameter α is positive and the scale parameter λ may assume any real value. As the Pareto distribution, the lognormal distribution is also a heavy-tailed distribution.

The parameters α and λ can easily be calculated from c_X^2 and $\overline{X}$:

$$\alpha = \sqrt{\ln(c_X^2 + 1)}\,, \qquad \lambda = \ln \overline{X} - \frac{\alpha^2}{2}\,. \tag{1.43}$$

The importance of this distribution arises from the fact that the product of n mutually independent random variables has a lognormal distribution in the limit $n \to \infty$. In Table 1.2 the formulae for the expectation $E[X]$, the variance $var(X)$, and the coefficient of variation c_X for some important distribution functions are summarized. Furthermore, in Table 1.3 formulae for estimating the parameters of these distributions are given.

1.3.2 Multiple Random Variables

In some cases, the result of one random experiment determines the values of several random variables, where these values may also affect each other. The *joint probability mass function* of the discrete random variables

Table 1.2 Expectation $E[X]$, variance $\mathrm{var}(X)$, and coefficient of variation c_X of important distributions

Distribution	Parameter	$E[X]$	$\mathrm{var}(X)$	c_X
Exponential	μ	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$	1
Erlang	μ, k $k=1,2,\ldots$	$\frac{1}{\mu}$	$\frac{1}{k\mu^2}$	$\frac{1}{\sqrt{k}} \leq 1$
Gamma	μ, α $(0<\alpha<\infty)$	$\frac{1}{\mu}$	$\frac{1}{\alpha\mu^2}$	$0 < \frac{1}{\sqrt{\alpha}} < \infty$
Hypoexponential	μ_1, μ_2	$\frac{1}{\mu_1} + \frac{1}{\mu_2}$	$\frac{1}{\mu_1^2} + \frac{1}{\mu_2^2}$	$\frac{\sqrt{\mu_1^2+\mu_2^2}}{\mu_1+\mu_2} < 1$
Hyperexponential	k, μ_i, q_i	$\sum_{i=1}^{k} \frac{q_i}{\mu_i} = \frac{1}{\mu}$	$2\sum_{i=1}^{k} \frac{q_i}{\mu_i^2} - \frac{1}{\mu^2}$	$\sqrt{2\mu^2 \sum_{i=1}^{k} \frac{q_i}{\mu_i^2} - 1} > 1$

$X_1, X_2, \ldots, X_n$ is given by

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \tag{1.44}$$

and represents the probability that $X_1 = x_1$, $X_2 = x_2$, $\ldots X_n = x_n$. In the continuous case, the *joint distribution function*:

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \tag{1.45}$$

represents the probability that $X_1 \leq x_1$, $X_2 \leq x_2$, $\ldots X_n \leq x_n$, where $\mathbf{X} = (X_1, \ldots, X_n)$ is the n-dimensional random variable and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

A simple example of an experiment with multiple discrete random variables is tossing a pair of dice. The following random variables might be determined:

X_1 number that shows on the first die,
X_2 number that shows on the second die,
$X_3 = X_1 + X_2$, sum of the numbers of both dice.

1.3.2.1 Independence The random variables $X_1, X_2, \ldots, X_n$ are called (*statistically*) *independent* if, for the continuous case:

$$\begin{aligned} &P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \\ &\quad = P(X_1 \leq x_1) \cdot P(X_2 \leq x_2) \cdot \ldots \cdot P(X_n \leq x_n), \end{aligned} \tag{1.46}$$

or the discrete case:

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &\quad = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdot \ldots \cdot P(X_n = x_n). \end{aligned} \tag{1.47}$$

Table 1.3 Formulae for estimating the parameters of important distributions

Distribution	Parameter	Calculation of the Parameters
Exponential	μ	$\mu = 1/\overline{X}$
Erlang	μ, k $k=1,2,\ldots$	$k = \text{ceil}(1/c_X^2)$ $\mu = 1/(c_X^2 \cdot k\overline{X})$
Gamma	μ, α $0<\alpha<\infty$	$\alpha = 1/c_X^2$ $\mu = 1/\overline{X}$
Hypoexponential	μ_1, μ_2	$\mu_{1/2} = \frac{2}{\overline{X}}\left[1 \pm \sqrt{1 + 2(c_X^2 - 1)}\right]^{-1}$
Hyperexponential (H_2)	μ_1, μ_2, q_1, q_2	$\mu_1 = \frac{1}{\overline{X}}\left[1 - \sqrt{\frac{q_2}{q_1}\frac{c_X^2 - 1}{2}}\right]^{-1}$ $\mu_2 = \frac{1}{\overline{X}}\left[1 + \sqrt{\frac{q_1}{q_2}\frac{c_X^2 - 1}{2}}\right]^{-1}$ $q_1 + q_2 = 1, \mu_2 > 0$
Cox ($c_X \leq 1$)	k, b_i, μ_i	$k = \text{ceil}(1/c_X^2)$ $b_1 = \frac{2kc_X^2 + k - 2 - \sqrt{k^2 + 4 - 4kc_X^2}}{2(c_X^2 + 1)(k-1)}$ $b_2 = b_3 = \ldots = b_{k-1} = 0, \quad b_k = 1$ $\mu_1 = \mu_2 = \ldots = \mu_k = \frac{k - b_1 \cdot (k-1)}{\overline{X}}$
Cox ($c_X > 1$)	k, b, μ_1, μ_2	$k = 2$ $b = c_X^2\left[1 - \sqrt{1 - \frac{2}{1 + c_X^2}}\right]$ $\mu_{1/2} = \frac{1}{\overline{X}}\left[1 \pm \sqrt{1 - \frac{2}{1 + c_X^2}}\right]$

Otherwise, they are (*statistically*) *dependent*. In the preceding example of tossing a pair of dice, the random variables X_1 and X_2 are independent, whereas X_1 and X_3 are dependent on each other. For example:

$$P(X_1 = i, X_2 = j) = \frac{1}{36}, \qquad i, j = 1, \ldots, 6,$$

since there are 36 different possible results altogether that are all equally probable. Because the following is also valid:

$$P(X_1 = i) \cdot P(X_2 = j) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36},$$

$P(X_1 = i, X_2 = j) = P(X_1 = i) \cdot P(X_2 = j)$ is true, and therefore X_1 and X_2 are independent. On the other hand, if we observe the dependent variables X_1 and X_3, then

$$P(X_1 = 2, X_3 = 4) = P(X_1 = 2, X_2 = 2) = \frac{1}{36},$$

since X_1 and X_2 are independent. Having

$$\begin{aligned} & P(X_1 = 2) = \frac{1}{6} \\ \text{and} \quad & P(X_3 = 4) = P(X_1 = 1, X_2 = 3) + P(X_1 = 2, X_2 = 2) \\ & \qquad + P(X_1 = 3, X_2 = 1) = \frac{3}{36} = \frac{1}{12} \end{aligned}$$

results in $P(X_1 = 2) \cdot P(X_3 = 4) = \frac{1}{6} \cdot \frac{1}{12} = \frac{1}{72}$, which shows the dependency of the random variables X_1 and X_3:

$$P(X_1 = 2, X_3 = 4) \neq P(X_1 = 2) \cdot P(X_3 = 4).$$

1.3.2.2 Conditional Probability A *conditional probability*:

$$P(X_1 = x_1 \mid X_2 = x_2, \ldots, X_n = x_n)$$

is the probability for $X_1 = x_1$ under the conditions $X_2 = x_2$, $X_3 = x_3$, etc. Then we get

$$\begin{aligned} & P(X_1 = x_1 \mid X_2 = x_2, \ldots, X_n = x_n) \\ & \quad = \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{P(X_2 = x_2, \ldots, X_n = x_n)}. \end{aligned} \tag{1.48}$$

For continuous random variables we have

$$\begin{aligned} & P(X_1 \leq x_1 \mid X_2 \leq x_2, \ldots, X_n \leq x_n) \\ & \quad = \frac{P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)}{P(X_2 \leq x_2, \ldots, X_n \leq x_n)}. \end{aligned} \tag{1.49}$$

We demonstrate this also with the preceding example of tossing a pair of dice. The probability that $X_3 = j$ under the condition that $X_1 = i$ is given by

$$P(X_3 = j \mid X_1 = i) = \frac{P(X_3 = j, X_1 = i)}{P(X_1 = i)} .$$

For example, with $j = 4$ and $i = 2$:

$$P(X_3 = 4 \mid X_1 = 2) = \frac{1/36}{1/6} = \frac{1}{6} .$$

If we now observe both random variables X_1 and X_2, we will see that because of the independence of X_1 and X_2,

$$P(X_1 = j \mid X_2 = i) = \frac{P(X_1 = j, X_2 = i)}{P(X_2 = i)} = \frac{P(X_1 = j) \cdot P(X_2 = i)}{P(X_2 = i)}$$

$$= P(X_1 = j) .$$

1.3.2.3 Important Relations The following relations concerning multiple random variables will be used frequently in the text:

- The expected value of a sum of random variables is equal to the sum of the expected values of these random variables. If $c_1, \ldots, c_n$ are arbitrary constants and $X_1, X_2, \ldots, X_n$ are (not necessarily independent) random variables, then

$$E\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i E[X_i] . \tag{1.50}$$

- If the random variables $X_1, X_2, \ldots, X_n$ are stochastically independent, then the expected value of the product of the random variables is equal to the product of the expected values of the random variables:

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i] . \tag{1.51}$$

- The *covariance* of two random variables X and Y is a way of measuring the dependency between X and Y. It is defined by

$$\begin{aligned} \operatorname{cov}[X, Y] &= E\left[(X - E[X])(Y - E[Y])\right] = \overline{(X - \overline{X})(Y - \overline{Y})} \\ &= E[X \cdot Y] - E[X] \cdot E[Y] = \overline{XY} - \overline{X} \cdot \overline{Y} . \end{aligned} \tag{1.52}$$

If $X = Y$, then the covariance is equal to the variance:

$$\operatorname{cov}[X, X] = \overline{X^2} - \overline{X}^2 = \operatorname{var}(X) = \sigma_X^2 .$$

2

Markov Chains

2.1 MARKOV PROCESSES

Markov processes provide very flexible, powerful, and efficient means for the description and analysis of dynamic (computer) system properties. Performance and dependability measures can be easily derived. Moreover, Markov processes constitute the fundamental theory underlying the concept of queueing systems. In fact, the notation of queueing systems has been viewed sometimes as a high-level *specification* technique for (a sub-class of) Markov processes. Each queueing system can, in principle, be mapped onto an instance of a Markov process and then mathematically evaluated in terms of this process. But besides highlighting the *computational* relation between Markov processes and queueing systems, it is worthwhile pointing out also that fundamental *properties* of queueing systems are commonly *proved* in terms of the underlying Markov processes. This type of use of Markov processes is also possible even when queueing systems exhibit properties such as nonexponential distributions that cannot be represented directly by discrete-state Markov models. Markovizing methods, such as embedding techniques or supplementary variables, can be used in such cases. Here Markov processes serve as a mere *theoretical* framework to prove the correctness of computational methods applied directly to the analysis of queueing systems. For the sake of efficiency, an explicit creation of the Markov process is preferably avoided.

2.1.1 Stochastic and Markov Processes

There exist many textbooks like those from King [King90], Trivedi [Triv01], Allen [Alle90], Gross and Harris [GrHa85], Cinlar [Cinl75], Feller (his two volume classic [Fell68]), and Howard [Howa71] that provide excellent intro-

ductions into the basics of stochastic and Markov processes. Besides the theoretical background, many motivating examples are also given in those books. Consequently, we limit discussion here to the essentials of Markov processes and refer to the literature for further details.

Markov processes constitute a special, perhaps the most important, subclass of stochastic processes, while the latter can be considered as a generalization of the concept of random variables. In particular, a stochastic process provides a relation between the elements of a possibly infinite family of random variables. A series of random experiments can thus be taken into consideration and analyzed as a whole.

Definition 2.1 A *stochastic process* is defined as a family of random variables $\{X_t : t \in T\}$ where each random variable X_t is indexed by parameter $t \in T$, which is usually called the *time parameter* if $T \subseteq \mathbb{R}_+ = [0, \infty)$. The set of all possible values of X_t (for each $t \in T$) is known as the state space S of the stochastic process.

If a countable, discrete-parameter set T is encountered, the stochastic process is called a *discrete-parameter* process and T is commonly represented by (a subset of) $\mathbb{N}_0 = \{0, 1, \dots\}$; otherwise we call it a *continuous-parameter* process. The *state space* of the stochastic process may also be continuous or discrete. Generally, we restrict ourselves here to the investigation of discrete state spaces and in that case refer to the stochastic processes as *chains*, but both continuous- and discrete-parameter processes are considered.

Definition 2.2 Continuous-parameter stochastic processes can be probabilistically characterized by the *joint (cumulative) distribution function* (CDF) $F_{\mathbf{X}}(\mathbf{s}; \mathbf{t})$ for a given set of random variables $\{X_{t_1}, X_{t_2}, \dots, X_{t_n}\}$, parameter vector $\mathbf{t} = (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$, and state vector $\mathbf{s} = (s_1, s_2, \dots, s_n) \in \mathbb{R}^n$, where $t_1 < t_2 < \cdots < t_n$:

$$F_{\mathbf{X}}(\mathbf{s}; \mathbf{t}) = P(X_{t_1} \leq s_1, X_{t_2} \leq s_2, \dots, X_{t_n} \leq s_n). \tag{2.1}$$

The *joint probability density function* (pdf),

$$f_{\mathbf{X}}(\mathbf{s}; \mathbf{t}) = \frac{\partial^n F_{\mathbf{X}}(\mathbf{s}; \mathbf{t})}{\partial s_1 \partial s_2 \dots \partial s_n},$$

is defined correspondingly if the partial derivatives exist. If the so-called Markov property is imposed on the *conditional CDF* of a stochastic process, a Markov process results:

Definition 2.3 A stochastic process $\{X_t : t \in T\}$ constitutes a *Markov process* if for all $0 = t_0 < t_1 < \cdots < t_n < t_{n+1}$ and all $s_i \in S$ the conditional CDF of $X_{t_{n+1}}$ depends only on the last previous value X_{t_n} and not on the earlier values $X_{t_0}, X_{t_1}, \dots, X_{t_{n-1}}$:

$$\begin{aligned} &P(X_{t_{n+1}} \leq s_{n+1} \mid X_{t_n} = s_n, X_{t_{n-1}} = s_{n-1}, \dots, X_{t_0} = s_0) \\ &\quad = P(X_{t_{n+1}} \leq s_{n+1} \mid X_{t_n} = s_n). \end{aligned} \tag{2.2}$$

This most general definition of a Markov process can be adopted to special cases. In particular, we focus here on discrete state spaces and on both discrete- and continuous-parameter Markov processes. As a result, we deal primarily with *continuous-time Markov chains* (CTMC), and with *discrete-time Markov chains* (DTMC), which are introduced in the next section. Finally, it is often sufficient to consider only systems with a time independent, i.e., *time-homogeneous*, pattern of dynamic behavior. Note that time-homogeneous system dynamics is to be discriminated from stationary system behavior, which relates to time independence in a different sense. The former refers to the stationarity of the *conditional* CDF while the latter refers to the stationarity of the CDF *itself*.

Definition 2.4 Letting $t_0 = 0$ without loss of generality, a Markov process is said to be *time-homogeneous* if the conditional CDF of $X_{t_{n+1}}$ does not depend on the observation time, that is, it is invariant with respect to time epoch t_n:

$$P(X_{t_{n+1}} \leq s_{n+1} \mid X_{t_n} = s_n) = P(X_{t_{n+1}-t_n} \leq s_{n+1} \mid X_0 = s_n). \tag{2.3}$$

2.1.2 Markov Chains

Equation (2.2) describes the well-known Markov property. Informally this can be interpreted in the sense that the whole history of a Markov chain is summarized in the current state X_{t_n}. Equivalently, given the present, the future is conditionally independent of the past. Note that the Markov property does not prevent the conditional distribution from being dependent on the time variable t_n. Such a dependence is prevented by the definition of homogeneity (see Eq. (2.3)). A unique characteristic is implied, namely, the sojourn time distribution in any state of a homogeneous Markov chain exhibits the memoryless property. An immediate, and somewhat curious, consequence is that the mean sojourn time equals the mean residual and the mean elapsed time in any state and at any time [Triv01].

If not explicitly stated otherwise, we consider Markov processes with discrete state spaces only, that is, Markov chains, in what follows. Note that in this case we are inclined to talk about probability mass functions, pmf, rather than probability density functions, pdf. Refer back to Sections 1.3.1.1 and 1.3.1.2 for details.

2.1.2.1 Discrete-Time Markov Chains We are now ready to proceed to the formal definition of Markov chains. Discrete-parameter Markov chains are considered first, that is, Markov processes restricted to a discrete, finite, or countably infinite state space, S, and a discrete-parameter space T. For the sake of convenience, we set $T \subseteq \mathbb{N}_0$. The conditional pmf reflecting the Markov property for discrete-time Markov chains, corresponding to Eq. (2.2), is summarized in the following definition:

Definition 2.5 A given stochastic process $\{X_0, X_1, \ldots, X_{n+1}, \ldots\}$ at the consecutive points of observation $0, 1, \ldots, n+1$ constitutes a *DTMC* if the following relation on the *conditional pmf*, that is, the Markov property, holds for all $n \in \mathbb{N}_0$ and all $s_i \in S$:

$$\begin{aligned} &P(X_{n+1} = s_{n+1} \mid X_n = s_n, X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) \\ &\quad = P(X_{n+1} = s_{n+1} \mid X_n = s_n)\,. \end{aligned} \tag{2.4}$$

Given an initial state s_0, the DTMC evolves over time, that is, step by step, according to *one-step transition probabilities*. The right-hand side of Eq. (2.4) reveals the conditional pmf of transitions from state s_n at time step n to state s_{n+1} at time step $(n+1)$. Without loss of generality, let $S = \{0, 1, 2, \ldots\}$ and write conveniently the following shorthand notation for the conditional pmf of the process's one-step transition from state i to state j at time n:

$$p_{ij}^{(1)}(n) = P(X_{n+1} = s_{n+1} = j \mid X_n = s_n = i)\,. \tag{2.5}$$

In the homogeneous case, when the conditional pmf is independent of epoch n, Eq. (2.5) reduces to

$$p_{ij}^{(1)} = p_{ij}^{(1)}(n) = P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i)\,, \quad \forall n \in T\,. \tag{2.6}$$

For the sake of convenience, we usually drop the superscript, so that $p_{ij} = p_{ij}^{(1)}$ refers to a one-step transition probability of a homogeneous DTMC.

Starting with state i, the DTMC will go to some state j (including the possibility of $j = i$), so that it follows that $\sum_j p_{ij} = 1$, where $0 \leq p_{ij} \leq 1$. The one-step transition probabilities p_{ij} are usually summarized in a non-negative, stochastic[1] transition matrix $\mathbf{P}$:

$$\mathbf{P} = \mathbf{P}^{(1)} = [p_{ij}] = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Graphically, a finite-state DTMC is represented by a *state transition diagram* (also referred to as *state diagram*), a finite directed graph, where state i of the chain is depicted by a vertex, and a one-step transition from state i to state j by an edge marked with one-step transition probability p_{ij}. As an example, consider the one-step transition probability matrix in Eq. (2.7) with state space $S = \{0, 1\}$ and the corresponding graphical representation in Fig. 2.1.

[1] The elements in each row of the matrix sum up to 1.

Example 2.1 The one-step transition probability matrix of the two-state DTMC in Fig. 2.1 is given by

$$\mathbf{P}^{(1)} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \end{pmatrix} . \tag{2.7}$$

Conditioned on the current DTMC state, a transition is made from state 0 to state 1 with probability 0.25, and with probability 0.75, the DTMC remains in state 0 at the next time step. Correspondingly, a transition occurs from state 1 to state 0 with probability 0.5, and with probability 0.5 the chain remains in state 1 at the next time step.

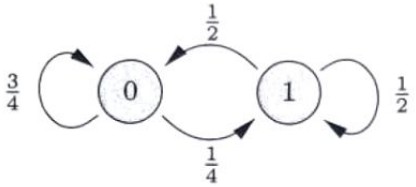

Fig. 2.1 Example of a discrete-time Markov chain referring to Eq. (2.7).

Repeatedly applying one-step transitions generalizes immediately to *n-step transition probabilities*. More precisely, let $p_{ij}^{(n)}(k,l)$ denote the probability that the Markov chain transits from state i at time k to state j at time l in exactly $n = l - k$ steps:

$$p_{ij}^{(n)}(k,l) = P(X_l = j \mid X_k = i) , \quad 0 \le k \le l . \tag{2.8}$$

Again, the theorem of total probability applies for any given state i and any given time values k and l such that $\sum_j p_{ij}^{(n)}(k,l) = 1$, where $0 \le p_{ij}^{(n)}(k,l) \le 1$. This fact, together with the Markov property, immediately leads us to a procedure for computing the n-step transition probabilities recursively from the one-step transition probabilities: The transition of the process from state i at time k to state j at time l can be split into sub-transitions from state i at time k to an intermediate state[2] h, say, at time m and from there, *independently of the history* that led to that state, from state h at time m to state j at time l, where $k < m < l$ and $n = l - k$. This condition leads to the well-known system of *Chapman–Kolmogorov equations*:

$$p_{ij}^{(n)}(k,l) = \sum_{h \in S} p_{ih}^{(m-k)}(k,m) p_{hj}^{(l-m)}(m,l) , \quad 0 \le k < m < l . \tag{2.9}$$

Note that the conditional independence assumption, i.e., the Markov property, is reflected by the product of terms on the right-hand side of Eq. (2.9).

[2]The Markov chain must simply traverse *some* state at any time.

Similar to the one-step case, the n-step transition probabilities can be simplified for homogeneous DTMC such that $p_{ij}^{(n)} = p_{ij}^{(n)}(k,l)$ depend only on the difference $n = l - k$ and not on the actual values of k and l:

$$p_{ij}^{(n)} = P(X_{k+n} = j \mid X_k = i) = P(X_n = j \mid X_0 = i)\,, \quad \forall k \in T\,. \tag{2.10}$$

Under this condition, the Chapman–Kolmogorov Eq. (2.9) for *homogeneous* DTMC simplifies to

$$p_{ij}^{(n)} = \sum_{h \in S} p_{ih}^{(m)} p_{hj}^{(n-m)}\,, \quad 0 < m < n\,. \tag{2.11}$$

Because Eq. (2.11) holds for all $m < n$, let $m = 1$ and get

$$p_{ij}^{(n)} = \sum_{h \in S} p_{ih}^{(1)} p_{hj}^{(n-1)}\,.$$

With $\mathbf{P}^{(n)}$ as the matrix of n-step transition probabilities $p_{ij}^{(n)}$, Eq. (2.11) can be rewritten in matrix form for the particular case of $m = 1$ as $\mathbf{P}^{(n)} = \mathbf{P}^{(1)}\mathbf{P}^{(n-1)} = \mathbf{P}\mathbf{P}^{(n-1)}$. Applying this procedure recursively results in the following equation[3]:

$$\mathbf{P}^{(n)} = \mathbf{P}\mathbf{P}^{(n-1)} = \mathbf{P}^n\,. \tag{2.12}$$

The n-step transition probability matrix can be computed by the $(n-1)$-fold multiplication of the one-step transition matrix by itself.

Example 2.2 Referring back to the example in Fig. 2.1 and Eq. (2.7), the four-step transition probability matrix, for instance, can be derived according to Eq. (2.12):

$$\begin{aligned}
\mathbf{P}^{(4)} &= \mathbf{P}\mathbf{P}^{(3)} = \mathbf{P}^2\mathbf{P}^{(2)} \\
&= \begin{pmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \end{pmatrix}^2 \mathbf{P}^{(2)} = \begin{pmatrix} 0.6875 & 0.3125 \\ 0.625 & 0.375 \end{pmatrix} \mathbf{P}\mathbf{P}^{(1)} \\
&= \begin{pmatrix} 0.67188 & 0.32813 \\ 0.65625 & 0.34375 \end{pmatrix} \mathbf{P}^{(1)} = \begin{pmatrix} 0.66797 & 0.33203 \\ 0.66406 & 0.33594 \end{pmatrix}.
\end{aligned} \tag{2.13}$$

Ultimately, we wish to compute the pmf of the random variable X_n, that is, the probabilities $\nu_i(n) = P(X_n = i)$ that the DTMC is in state i, at time step n. These probabilities, called *transient state probabilities* at time n, will then allow us to derive the desired performance measures. Given the n-step transition probability matrix $\mathbf{P}^{(n)}$, the vector of the state probabilities at time n, $\boldsymbol{\nu}(n) = (\nu_0(n), \nu_1(n), \nu_2(n), \dots)$ can be obtained by un-conditioning $\mathbf{P}^{(n)}$ on the *initial probability vector* $\boldsymbol{\nu}(0) = (\nu_0(0), \nu_1(0), \nu_2(0), \dots)$:

$$\boldsymbol{\nu}(n) = \boldsymbol{\nu}(0)\mathbf{P}^{(n)} = \boldsymbol{\nu}(0)\mathbf{P}^n = \boldsymbol{\nu}(n-1)\mathbf{P}. \tag{2.14}$$

[3]It is important to keep in mind that $\mathbf{P}^{(n)} = \mathbf{P}^n$ holds only for homogeneous DTMC.

Note that both $\boldsymbol{\nu}(n)$ and $\boldsymbol{\nu}(0)$ are represented as row vectors in Eq. (2.14).

Example 2.3 Assume that the DTMC from Eq. (2.7) under investigation is initiated in state 1, then the initial probability vector $\boldsymbol{\nu}^{(1)}(0)=(0,1)$ is to be applied in the un-conditioning according to Eq. (2.14). With the already computed four-step transition probabilities in Eq. (2.13) the corresponding pmf $\boldsymbol{\nu}^{(1)}(4)$ can be derived:

$$\boldsymbol{\nu}^{(1)}(4) = (0,1)\begin{pmatrix} 0.66797 & 0.33203 \\ 0.66406 & 0.33594 \end{pmatrix} = (0.66406, 0.33594)\ .$$

Example 2.4 Alternatively, with another initial probability vector:

$$\boldsymbol{\nu}^{(2)}(0) = \left(\tfrac{2}{3}, \tfrac{1}{3}\right) = (0.6\bar{6}, 0.3\bar{3})\ ,$$

according to Eq. (2.14), we would get

$$\boldsymbol{\nu}^{(2)}(4) = (0.6\bar{6}, 0.3\bar{3})$$

as the probabilities at time step 4.

Of particular importance are homogeneous DTMC on which a so-called *stationary probability vector* can be imposed in a suitable way:

Definition 2.6 State probabilities $\boldsymbol{\nu} = (\nu_0, \nu_1, \ldots \nu_i, \ldots)$ of a discrete-time Markov chain are said to be *stationary*, if any transitions of the underlying DTMC according to the given one-step transition probabilities $\mathbf{P} = [p_{ij}]$ have no effect on these state probabilities, that is, $\nu_j = \sum_{i\in S} \nu_i p_{ij}$ holds for all states $j \in S$. This relation can also be expressed in matrix form:

$$\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{P}\,, \quad \sum_{i\in S} \nu_i = 1\,. \tag{2.15}$$

Note that according to the preceding definition, more than one stationary pmf can exist for a given, unrestricted, DTMC.

Example 2.5 By substituting the one-step transition matrix from Eq. (2.7) in Eq. (2.15), it can easily be checked that

$$\boldsymbol{\nu}^{(2)} = \left(\tfrac{2}{3}, \tfrac{1}{3}\right) = (0.6\bar{6}, 0.3\bar{3})$$

is a stationary probability vector while $\boldsymbol{\nu}^{(1)} = (0,1)$ is not.

Definition 2.7 For an efficient analysis, we are interested in the *limiting state probabilities* $\tilde{\boldsymbol{\nu}}$ as a particular kind of stationary state probabilities, which are defined by

$$\tilde{\boldsymbol{\nu}} = \lim_{n\to\infty} \boldsymbol{\nu}(n) = \lim_{n\to\infty} \boldsymbol{\nu}(0)\mathbf{P}^{(n)} = \boldsymbol{\nu}(0) \lim_{n\to\infty} \mathbf{P}^{(n)} = \boldsymbol{\nu}(0)\tilde{\mathbf{P}}\,. \tag{2.16}$$

Definition 2.8 As $n \to \infty$, we may require both the n-step transition probability matrix $\mathbf{P}^{(n)}$ and the state probability vector $\boldsymbol{\nu}(n)$ to *converge independently* of the initial probability vector $\boldsymbol{\nu}(0)$ to $\tilde{\mathbf{P}}$ and $\tilde{\boldsymbol{\nu}}$, respectively. Also, we may only be interested in the case where the state probabilities $\tilde{\nu}_i > 0, \forall i \in S$, are *strictly positive* and $\sum_i \tilde{\nu}_i = 1$, that is, $\tilde{\boldsymbol{\nu}}$ constitutes a pmf.

If all these restrictions apply to a given probability vector, it is said to be the *unique steady-state probability vector* of the DTMC.

Example 2.6 Returning to Example 2.1, we note that the n-step transition probabilities converge as $n \to \infty$:

$$\tilde{\mathbf{P}} = \lim_{n\to\infty} \mathbf{P}^{(n)} = \lim_{n\to\infty} \begin{pmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \end{pmatrix}^n = \begin{pmatrix} 0.6\bar{6} & 0.3\bar{3} \\ 0.6\bar{6} & 0.3\bar{3} \end{pmatrix} .$$

Example 2.7 With this result, the limiting state probability vector $\tilde{\boldsymbol{\nu}}$, which is independent of any initial probability vector $\boldsymbol{\nu}(0)$, can be derived according to Eq. (2.16):

$$\tilde{\boldsymbol{\nu}} = (0.6\bar{6}, 0.3\bar{3}) .$$

Example 2.8 Since all probabilities in the vector $(0.6\bar{6}, 0.3\bar{3})$ are strictly positive, this vector constitutes the unique steady-state probability vector of the DTMC.

Eventually, the limiting state probabilities become *independent of time steps*, such that once the limiting probability vector is reached, further transitions of the DTMC do not change this vector, i.e., it is stationary. Note that such a probability vector does not necessarily exist for all DTMCs.

If Eq. (2.16) holds and $\tilde{\boldsymbol{\nu}}$ is independent of $\boldsymbol{\nu}(0)$, it follows that the limit $\mathbf{P}^{(n)} = [p_{ij}^{(n)}]$ is independent of time n and of index i. All rows of $\tilde{\mathbf{P}}$ would be identical, that is, the rows would match element by element. Furthermore, the jth element $\tilde{p}_{ij}$ of row i equals $\tilde{\nu}_j$ for all $i \in S$:

$$\tilde{\mathbf{P}} = \lim_{n\to\infty} \mathbf{P}^{(n)} = \lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \tilde{\nu}_0 & \tilde{\nu}_1 & \tilde{\nu}_2 & \cdots \\ \tilde{\nu}_0 & \tilde{\nu}_1 & \tilde{\nu}_2 & \cdots \\ \tilde{\nu}_0 & \tilde{\nu}_1 & \tilde{\nu}_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} . \tag{2.17}$$

If the unique steady-state probability vector of a DTMC *exists*, it can be determined by the solution of the system of linear Eqs. (2.15), so that $\tilde{\mathbf{P}}$ need not be determined explicitly. From Eq. (2.14), we have $\boldsymbol{\nu}(n) = \boldsymbol{\nu}(n-1)\mathbf{P}$. If the limit exists, we can take it on both sides of the equation and get

$$\lim_{n\to\infty} \boldsymbol{\nu}(n) = \tilde{\boldsymbol{\nu}} = \lim_{n\to\infty} \boldsymbol{\nu}(n-1)\mathbf{P} = \tilde{\boldsymbol{\nu}}\mathbf{P} . \tag{2.18}$$

In the steady-state case no ambiguity can arise, so that, for the sake of convenience, we may drop the annotation and refer to steady-state probability

vector by using the notation $\boldsymbol{\nu}$ instead of $\tilde{\boldsymbol{\nu}}$. Steady-state and stationarity coincide in that case, i.e., there is only a unique stationary probability vector.

The computation of the steady-state probability vector $\boldsymbol{\nu}$ of a DTMC is usually significantly simpler and less expensive than a time-dependent computation of $\boldsymbol{\nu}(n)$. It is therefore the steady-state probability vector of a DTMC that is preferably taken advantage of in modeling endeavors. But a steady-state probability vector does not exist for all DTMCs.[4] Additionally, it is not always appropriate to restrict the analysis to the steady-state case, even if it does exist. Under some circumstances, time-dependent (i.e., transient) analysis would result in more meaningful information with respect to an application. Transient analysis has special relevance if short-term behavior is of more importance than long-term behavior. In modeling terms, "short term" means that the influence of the initial state probability vector $\boldsymbol{\nu}(0)$ on $\boldsymbol{\nu}(n)$ has not yet disappeared by time step n.

Before continuing, some simple example DTMCs are presented to clarify the definitions of this section. The following four one-step transition matrices are examined for the conditions under which stationary, limiting state, and steady-state probabilities exist for them.

Example 2.9 Consider the DTMC shown in Fig. 2.2 with the one-step transition probability matrix (TPM):

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} . \tag{2.19}$$

Fig. 2.2 Example of a discrete-time Markov chain referring to Eq. (2.19).

- For this one-step TPM, an infinite number of stationary probability vectors exists: Any arbitrary probability vector is stationary in this case according to Eq. (2.15).
- The n-step TPM $\mathbf{P}^{(n)}$ converges in the limit to

$$\lim_{n\to\infty} \mathbf{P}^{(n)} = \tilde{\mathbf{P}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} .$$

Furthermore, all n-step TPMs are identical:

$$\mathbf{P} = \tilde{\mathbf{P}} = \mathbf{P}^{(n)} , \quad \forall n \in T .$$

[4]The conditions under which DTMCs converge to steady-state is precisely stated in the following section.

The limiting state probabilities $\tilde{\nu}$ do exist and are identical to the initial probability vector $\nu(0)$ in all cases:

$$\tilde{\nu} = \nu(0)\tilde{\mathbf{P}} = \nu(0) .$$

- A unique steady-state probability vector does not exist for this example.

Example 2.10 Next, consider the DTMC shown in Fig. 2.3 with the TPM:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} . \tag{2.20}$$

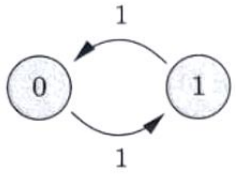

Fig. 2.3 Example of a discrete-time Markov chain referring to Eq. (2.20).

- For this one-step TPM, a stationary probability vector, which is unique in this case, does exist according to Eq. (2.15):

$$\nu = (0.5, 0.5) .$$

- The n-step transition matrix $\mathbf{P}^{(n)}$ does not converge in the limit to any $\tilde{\mathbf{P}}$. Therefore, the limiting state probabilities $\tilde{\nu}$ do not exist.
- Consequently, a unique steady-state probability vector does not exist.

Example 2.11 Consider the DTMC shown in Fig. 2.4 with the TPM:

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} . \tag{2.21}$$

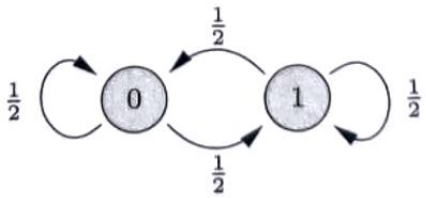

Fig. 2.4 Example of a discrete-time Markov chain referring to Eq. (2.21).

- For this one-step TPM a unique stationary probability vector does exis according to Eq. (2.15):

$$\nu = (0.5, 0.5) .$$

Note that this is the same unique stationary probability vector as for the different DTMC in Eq. (2.20).

- The n-step TPM $\mathbf{P}^{(n)}$ converges in the limit to

$$\lim_{n\to\infty} \mathbf{P}^{(n)} = \tilde{\mathbf{P}} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} .$$

Furthermore, all n-step TPMs are identical:

$$\mathbf{P} = \tilde{\mathbf{P}} = \mathbf{P}^{(n)} , \quad \forall n \in T .$$

The limiting state probabilities $\tilde{\boldsymbol{\nu}}$ do exist, are independent of the initial probability vector, and are unique:

$$\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu}(0)\tilde{\mathbf{P}} = (0.5, 0.5) .$$

- A unique steady-state probability vector does exist. All probabilities are strictly positive and identical to the stationary probabilities, which can be derived from the solution of Eq. (2.15):

$$\boldsymbol{\nu} = \tilde{\boldsymbol{\nu}} = (0.5, 0.5) .$$

Example 2.12 Now consider the DTMC shown in Fig. 2.5 with the TPM:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} . \tag{2.22}$$

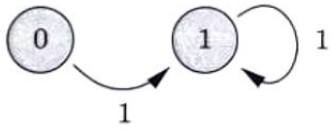

Fig. 2.5 Example of a discrete-time Markov chain referring to Eq. (2.22).

- For this one-step TPM a unique stationary probability vector does exist according to Eq. (2.15):

$$\boldsymbol{\nu} = (0, 1) .$$

- The n-step TPM $\mathbf{P}^{(n)}$ converges in the limit to

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

Furthermore, all n-step TPMs are identical:

$$\mathbf{P} = \tilde{\mathbf{P}} = \mathbf{P}^{(n)} , \quad \forall n \in T .$$

The limiting state probability vector $\tilde{\nu}$ does exist, is independent of the initial probability vector, and is identical to the unique stationary probability vector ν:

$$\tilde{\nu} = \nu = (0, 1) .$$

- A unique steady-state probability vector does not exist for this example. The elements of the unique stationary probability vector are not strictly positive.

We proceed now to identify necessary and sufficient conditions for the existence of a steady-state probability vector of a DTMC. The conditions can be given immediately in terms of properties of the DTMC.

2.1.2.1.1 Classifications of DTMC DTMCs are categorized based on the classifications of their constituent states.

Definition 2.9 Any state j is said to be *reachable* from any other state i, where $i, j \in S$, if it is possible to transit from state i to state j in a finite number of steps according to the given transition probability matrix. For some integer $n \geq 1$, the following relation must hold for the n-step transition probability:

$$p_{ij}^{(n)} > 0 , \quad \exists n, n \geq 1 . \tag{2.23}$$

A DTMC is called *irreducible* if all states in the chain can be reached pairwise from each other, i.e., $\forall i, j \in S, \exists n, n \geq 1 : p_{ij}^{(n)} > 0$.

A state $i \in S$ is said to be an *absorbing state*[5] if and only if no other state of the DTMC can be reached from it, i.e., $p_{ii} = 1$.

Note that a DTMC containing at least one absorbing state cannot be irreducible. If countably infinite state models are encountered, we have to discriminate more accurately how states are reachable from each other. The recurrence time and the probability of recurrence must also be taken into account.

Definition 2.10 Let $f_i^{(n)}$, called the *n-step recurrence probability*, denote the conditional probability of the first return to state $i \in S$ in exactly $n \geq 1$ steps after leaving state i. Then, the probability f_i of ever returning to state i is given by

$$f_i = \sum_{n=1}^{\infty} f_i^{(n)} . \tag{2.24}$$

Any state $i \in S$ to which the DTMC will return with probability $f_i = 1$, is called a *recurrent state*; otherwise, if $f_i < 1$, i is called a *transient state*.

[5] Absorbing states play an important role in the modeling of dependable systems where transient analysis is of primary interest.

Given a recurrent state i, the *mean recurrence time* m_i of state i of a DTMC is given by:

$$m_i = \sum_{n=1}^{\infty} n f_i^{(n)} . \tag{2.25}$$

If the mean recurrence time is finite, that is, $m_i < \infty$, i is called *positive recurrent* or *recurrent non-null*; otherwise, if $m_i = \infty$, state i is said to be *recurrent null*. For any recurrent state $i \in S$, let d_i denote the *period* of state i, then d_i is the greatest common divisor of the set of positive integers n such that $p_{ii}^{(n)} > 0$. A recurrent state i is called *aperiodic* if its period $d_i = 1$, and *periodic* with period d_i if $d_i > 1$.

It has been shown by Feller [Fell68] that the states of an *irreducible* DTMC are all of the same type. Hence, all states are periodic, aperiodic, transient, recurrent null, or recurrent non-null.

Definition 2.11 If one of the states i of an irreducible DTMC is aperiodic then so are all the other states $j \in S$, that is, $d_j = 1, \forall j \in S$, and the DTMC itself is called *aperiodic*; otherwise it is said to be *periodic* with unique period d.
An irreducible, aperiodic, discrete-time Markov chain with all states i being recurrent non-null with finite mean recurrence time m_i is called an *ergodic* Markov chain.

We are now ready to summarize the main results for the classification of discrete-time Markov chains:

- The states of a *finite-state*, *irreducible* Markov chain are all recurrent non-null.
- Given an *aperiodic* DTMC, the limits $\tilde{\boldsymbol{\nu}} = \lim_{n\to\infty} \boldsymbol{\nu}(n)$ do exist.
- For any *irreducible* and *aperiodic* DTMC, the limit $\tilde{\boldsymbol{\nu}}$ exists and is independent of the initial probability vector $\boldsymbol{\nu}(0)$.
- For an *ergodic* DTMC, the limit $\tilde{\boldsymbol{\nu}} = (\tilde{\nu}_0, \tilde{\nu}_1, \tilde{\nu}_2, \dots)$ exists and comprises the unique steady-state probability vector $\boldsymbol{\nu}$.
- The steady-state probabilities $\nu_i > 0$, $i \in S$, of an *ergodic* Markov chain can be obtained by solving the system of linear Eq. (2.15) or, if the (finite) mean recurrence times m_i are known, by exploiting the relation

$$\nu_i = \frac{1}{m_i}, \quad \forall i \in S . \tag{2.26}$$

If *finite-state* DTMCs are investigated, the solution of Eq. (2.15) can be obtained by applying standard methods for the solution of linear systems. In the *infinite* case, either *generating functions* can be applied or special *structure*

of the one-step transition probability matrix $\mathbf{P}^{(1)} = \mathbf{P}$ may be exploited to find the solution in closed form. An example of the latter technique is given in Section 3.1 where we investigate the important class of birth–death processes. The special (tridiagonal) structure of the matrix will allow us to derive closed-form solutions for the state probabilities that are not restricted to any fixed matrix size so that the limiting state probabilities of infinite state DTMCs are captured by the closed-form formulae as well.

2.1.2.1.2 DTMC State Sojourn Times The state *sojourn times* – the time between state changes – play an important role in the characterization of DTMCs. Only homogeneous DTMCs are considered here. We have already pointed out that the transition behavior reflects the memoryless property, that is, it only depends on the current state and neither on the history that led to the state nor on the time already spent in the current state. At every instant of time, the probability of leaving current state i is independently given by $(1 - p_{ii}) = \sum_{i \neq j} p_{ij}$. Applying this repeatedly leads to a description of a random experiment in form of a sequence of Bernoulli trials with probability of success $(1 - p_{ii})$, where "success" denotes the event of leaving current state i. Hence, the *sojourn time* R_i during a single visit to state i is a geometrically distributed random variable[6] with pmf:

$$P(R_i = k) = (1 - p_{ii}) p_{ii}^{k-1}, \quad \forall k \in \mathbb{N}^+ . \tag{2.27}$$

We can therefore immediately conclude that the expected sojourn time $E[R_i]$, that is, the mean number of time steps the process spends in state i per visit, is

$$E[R_i] = \frac{1}{1 - p_{ii}} . \tag{2.28}$$

Accordingly, the variance $\mathrm{var}[R_i]$ of the sojourn time per visit in state i is given by

$$\mathrm{var}[R_i] = \frac{p_{ii}}{(1 - p_{ii})^2} . \tag{2.29}$$

2.1.2.2 Continuous-Time Markov Chains Continuous- and discrete-time Markov chains provide different yet related modeling paradigms, each of them having their own domain of applications. For the definition of CTMCs we refer back to the definition of general Markov processes in Eq. (2.2) and specialize it to the continuous parameter, discrete state-space case. CTMCs are distinct from DTMCs in the sense that state transitions may occur at arbitrary instants of time and not merely at fixed, discrete time points, as is the case with DTMCs. Therefore, we use a subset of the set of non-negative real numbers $\mathbb{R}_0^+$ to refer to the parameter set T of a CTMC, as opposed to $\mathbb{N}_0$ for DTMCs:

[6] Note that the geometric distribution is the discrete-time equivalent of the exponential distribution; i.e., it is the only discrete distribution with the memoryless property.

Definition 2.12 A given stochastic process $\{X_t : t \in T\}$ constitutes a CTMC if for arbitrary $t_i \in \mathbb{R}_0^+$, with $0 = t_0 < t_1 < \cdots < t_n < t_{n+1}, \forall n \in \mathbb{N}$, and $\forall s_i \in S = \mathbb{N}_0$ for the conditional pmf, the following relation holds:

$$\begin{aligned} &P(X_{t_{n+1}} = s_{n+1} \mid X_{t_n} = s_n, X_{t_{n-1}} = s_{n-1}, \ldots, X_{t_0} = s_0) \\ &\quad = P(X_{t_{n+1}} = s_{n+1} \mid X_{t_n} = s_n)\,. \end{aligned} \tag{2.30}$$

Similar to Eq. (2.4) for DTMCs, Eq. (2.30) expresses the *Markov* property of continuous-time Markov chains. If we further impose homogeneity, then because the exponential distribution is the only continuous-time distribution that provides the memoryless property, the state sojourn times of a CTMC are necessarily exponentially distributed.

Again, the right-hand side of Eq. (2.30) is referred to as the *transition probability* [7] $p_{ij}(u, v)$ of the CTMC to travel from state i to state j during the period of time $[u, v)$, with $u, v \in T$ and $u \leq v$:

$$p_{ij}(u, v) = P(X_v = j \mid X_u = i)\,. \tag{2.31}$$

For $u = v$ we define

$$p_{ij}(u, u) = \begin{cases} 1\,, & i = j\,, \\ 0\,, & \text{otherwise.} \end{cases} \tag{2.32}$$

If the transition probabilities $p_{ij}(u, v)$ depend only on the time difference $t = v - u$ and not on the actual values of u and v, the simplified transition probabilities for *time-homogeneous* CTMC result:

$$p_{ij}(t) = p_{ij}(0, t) = P(X_{u+t} = j \mid X_u = i) = P(X_t = j \mid X_0 = i)\,, \quad \forall u \in T\,. \tag{2.33}$$

Given the transition probabilities $p_{ij}(u, v)$ and the probabilities $\pi_i(u)$ of the CTMC at time u, the unconditional state probabilities $\pi_j(v), j \in S$ of the process at time v can be derived:

$$\pi_j(v) = \sum_{i \in S} p_{ij}(u, v)\pi_i(u)\,, \quad \forall u, v \in T \;\; (u \leq v). \tag{2.34}$$

With $\mathbf{P}(u, v) = [p_{ij}(u, v)]$ as the matrix of the transition probabilities, for any pair of states $i, j \in S$ and any time interval $[u, v)$, $u, v \in T$, from the parameter domain, and the vector $\boldsymbol{\pi}(u) = (\pi_0(u), \pi_1(u), \pi_2(u), \ldots)$ of the state probabilities at any instant of time u, Eq. (2.34) can be given in vector–matrix form:

$$\boldsymbol{\pi}(v) = \boldsymbol{\pi}(u)\mathbf{P}(u, v)\,, \quad \forall u, v \in T \;\; (u \leq v)\,. \tag{2.35}$$

[7] Note that, as opposed to the discrete-time case, there is no fixed, discrete number of transition steps considered here.

Definition 2.12 A given stochastic process $\{X_t : t \in T\}$ constitutes a CTMC if for arbitrary $t_i \in \mathbb{R}_0^+$, with $0 = t_0 < t_1 < \cdots < t_n < t_{n+1}, \forall n \in \mathbb{N}$, and $\forall s_i \in S = \mathbb{N}_0$ for the conditional pmf, the following relation holds:

$$\begin{aligned} &P(X_{t_{n+1}} = s_{n+1} \mid X_{t_n} = s_n, X_{t_{n-1}} = s_{n-1}, \ldots, X_{t_0} = s_0) \\ &\quad = P(X_{t_{n+1}} = s_{n+1} \mid X_{t_n} = s_n)\,. \end{aligned} \tag{2.30}$$

Similar to Eq. (2.4) for DTMCs, Eq. (2.30) expresses the *Markov* property of continuous-time Markov chains. If we further impose homogeneity, then because the exponential distribution is the only continuous-time distribution that provides the memoryless property, the state sojourn times of a CTMC are necessarily exponentially distributed.

Again, the right-hand side of Eq. (2.30) is referred to as the *transition probability* [7] $p_{ij}(u, v)$ of the CTMC to travel from state i to state j during the period of time $[u, v)$, with $u, v \in T$ and $u \leq v$:

$$p_{ij}(u, v) = P(X_v = j \mid X_u = i)\,. \tag{2.31}$$

For $u = v$ we define

$$p_{ij}(u, u) = \begin{cases} 1\,, & i = j\,, \\ 0\,, & \text{otherwise.} \end{cases} \tag{2.32}$$

If the transition probabilities $p_{ij}(u, v)$ depend only on the time difference $t = v - u$ and not on the actual values of u and v, the simplified transition probabilities for *time-homogeneous* CTMC result:

$$p_{ij}(t) = p_{ij}(0, t) = P(X_{u+t} = j \mid X_u = i) = P(X_t = j \mid X_0 = i)\,, \quad \forall u \in T\,. \tag{2.33}$$

Given the transition probabilities $p_{ij}(u, v)$ and the probabilities $\pi_i(u)$ of the CTMC at time u, the unconditional state probabilities $\pi_j(v), j \in S$ of the process at time v can be derived:

$$\pi_j(v) = \sum_{i \in S} p_{ij}(u, v)\pi_i(u)\,, \quad \forall u, v \in T \ \ (u \leq v). \tag{2.34}$$

With $\mathbf{P}(u, v) = [p_{ij}(u, v)]$ as the matrix of the transition probabilities, for any pair of states $i, j \in S$ and any time interval $[u, v)$, $u, v \in T$, from the parameter domain, and the vector $\boldsymbol{\pi}(u) = (\pi_0(u), \pi_1(u), \pi_2(u), \ldots)$ of the state probabilities at any instant of time u, Eq. (2.34) can be given in vector–matrix form:

$$\boldsymbol{\pi}(v) = \boldsymbol{\pi}(u)\mathbf{P}(u, v)\,, \quad \forall u, v \in T \ \ (u \leq v)\,. \tag{2.35}$$

[7] Note that, as opposed to the discrete-time case, there is no fixed, discrete number of transition steps considered here.

Note that for all $u \in T$, $\mathbf{P}(u,u) = \mathbf{I}$ is the identity matrix.

In the time-homogeneous case, Eq. (2.34) reduces to

$$\pi_j(t) = \sum_{i \in S} p_{ij}(t)\pi_i(0) = \sum_{i \in S} p_{ij}(0,t)\pi_i(0)\,, \tag{2.36}$$

or in vector–matrix notation:

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\mathbf{P}(t) = \boldsymbol{\pi}(0)\mathbf{P}(0,t)\,. \tag{2.37}$$

Similar to the discrete-time case (Eq. (2.9)), the *Chapman–Kolmogorov* equation (Eq. (2.38)) for the transition probabilities of a CTMC can be derived from Eq. (2.30) by applying again the theorem of total probability:

$$p_{ij}(u,v) = \sum_{k \in S} p_{ik}(u,w)p_{kj}(w,v)\,, \quad 0 \le u \le w < v\,. \tag{2.38}$$

But, *unlike* the discrete-time case, Eq. (2.38) cannot be solved easily and used directly for computing the state probabilities. Rather, it has to be transformed into a system of differential equations which, in turn, leads us to the required results. For this purpose, we define the instantaneous *transition rates* $q_{ij}(t)$ $(i \neq j)$ of the CTMC traveling from state i to state j. These transition rates are related to conditional transition probabilities. Consider the period of time $[t, t+\Delta t)$, where Δt is chosen such that $\sum_{j \in S} q_{ij}(t)\Delta t + o(\Delta t) = 1$.[8] The non-negative, finite, continuous functions $q_{ij}(t)$ can be shown to exist under rather general conditions. For all states i, j, $i \neq j$, we define

$$q_{ij}(t) = \lim_{\Delta t \to 0} \frac{p_{ij}(t, t+\Delta t)}{\Delta t}\,, \quad i \neq j, \tag{2.39}$$

$$q_{ii}(t) = \lim_{\Delta t \to 0} \frac{p_{ii}(t, t+\Delta t) - 1}{\Delta t}\,. \tag{2.40}$$

If the limits do exist, it is clear from Eqs. (2.39) and (2.40) that, since $\sum_{j \in S} p_{ij}(t, t+\Delta t) = 1$, at any instant of time t:

$$\sum_{j \in S} q_{ij}(t) = 0\,, \quad \forall i \in S\,. \tag{2.41}$$

The quantity $-q_{ii}(t)$ can be interpreted as the total rate at which state i is exited (to any other state) at time t. Accordingly, $q_{ij}(t)$, $(i \neq j)$, denotes the rate at which the CTMC leaves state i in order to transit to state j at time t. As an equivalent interpretation, we can regard $q_{ij}(t)\Delta t + o(\Delta t)$ as the transition probability $p_{ij}(t, t+\Delta t)$ of the Markov chain to transit from

[8] The notation $o(\Delta t)$ is defined such that $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$; that is, we might substitute any function for $o(\Delta t)$ that approaches zero faster than the linear function Δt.

state i to state j in $[t, t+\Delta t)$, where Δt is chosen appropriately. Having these definitions, we return to the Chapman–Kolmogorov equation (Eq. (2.38)). Substituting $v + \Delta t$ for v in (2.38) and subtracting both sides of the original Eq. (2.38) from the result gives us

$$p_{ij}(u, v + \Delta t) - p_{ij}(u, v) = \sum_{k \in S} p_{ik}(u, w)[p_{kj}(w, v + \Delta t) - p_{kj}(w, v)] \,. \quad (2.42)$$

Dividing both sides of Eq. (2.42) by Δt, taking $\lim_{\Delta t \to 0}$ of the resulting quotient of differences, and letting $w \to v$, we derive a differential equation, the well-known *Kolmogorov forward equation*:

$$\frac{\partial p_{ij}(u, v)}{\partial v} = \sum_{k \in S} p_{ik}(u, v) q_{kj}(v) \,, \quad 0 \le u < v \,. \quad (2.43)$$

In the homogeneous case, we let $t = v - u$ and get from Eqs. (2.39) and (2.40) *time-independent* transition rates $q_{ij} = q_{ij}(t), \forall i, j \in S$, such that simpler versions of the Kolmogorov forward differential equation for homogeneous CTMCs result:

$$\frac{\mathrm{d}\, p_{ij}(t)}{\mathrm{d}\, t} = \sum_{k \in S} p_{ik}(t) q_{kj} = \sum_{k \in S} p_{ik}(0, t) q_{kj} \,. \quad (2.44)$$

Instead of the forward equation (2.43), we can equivalently derive and use the *Kolmogorov backward equation* for further computations, both in the homogeneous and nonhomogeneous cases, by letting $w \to u$ in Eq. (2.42) and taking $\lim_{\Delta t \to 0}$ to get

$$\frac{\partial p_{ij}(u, v)}{\partial u} = \sum_{k \in S} p_{kj}(u, v) q_{ik}(u) \,, \quad 0 \le u < v \,. \quad (2.45)$$

Differentiating Eq. (2.34) on both sides gives Eq. (2.47), using the Kolmogorov (forward) equation (Eq. (2.43)) yields Eq. (2.48), and then, again, applying Eq. (2.34) to Eq. (2.49), we derive the differential equation for the *unconditional state probabilities* $\pi_j(v)$, $\forall j \in S$, at time v in Eq. (2.50):

$$\begin{aligned}
\frac{\mathrm{d}\, \pi_j(v)}{\mathrm{d}\, v} &= \frac{\partial \sum_{i \in S} p_{ij}(u, v) \pi_i(u)}{\partial v} && (2.46)\\
&= \sum_{i \in S} \frac{\partial p_{ij}(u, v)}{\partial v} \pi_i(u) && (2.47)\\
&= \sum_{i \in S} \left(\sum_{k \in S} p_{ik}(u, v) q_{kj}(v) \right) \pi_i(u) && (2.48)\\
&= \sum_{k \in S} q_{kj}(v) \sum_{i \in S} p_{ik}(u, v) \pi_i(u) && (2.49)\\
&= \sum_{k \in S} q_{kj}(v) \pi_k(v) \,. && (2.50)
\end{aligned}$$

In the time-homogeneous case, a simpler version of Eq. (2.50) results by assuming $t = v - u$ and using time-independent transition rates q_{ij}. So we get the system of differential Eqs. (2.51):

$$\frac{\mathrm{d}\,\pi_j(t)}{\mathrm{d}\,t} = \sum_{i \in S} q_{ij}\pi_i(t)\,, \quad \forall j \in S\,, \tag{2.51}$$

which is repeatedly used throughout this text. Usually, we prefer vector–matrix form rather than the notation used in Eq. (2.51). Therefore, for the homogeneous case, we define the *infinitesimal generator matrix* $\mathbf{Q}$ of the transition probability matrix $\mathbf{P}(t) = [p_{ij}(0,t)] = [p_{ij}(t)]$ by referring to Eqs. (2.39) and (2.40). The matrix $\mathbf{Q}$

$$\mathbf{Q} = [q_{ij}]\,, \quad \forall i,j \in S\,, \tag{2.52}$$

contains the transition rates q_{ij} from any state i to any other state j, where $i \neq j$, of a given continuous-time Markov chain. The elements q_{ii} on the main diagonal of $\mathbf{Q}$ are defined by $q_{ii} = -\sum_{j,j\neq i} q_{ij}$. With the definition in Eq. (2.52), Eq. (2.51) can be given in vector–matrix form as

$$\dot{\boldsymbol{\pi}}(t) = \frac{\mathrm{d}\,\boldsymbol{\pi}(t)}{\mathrm{d}\,t} = \boldsymbol{\pi}(t)\mathbf{Q}\,. \tag{2.53}$$

For the sake of completeness, we include also the matrix form of the Kolmogorov differential equations in the time-homogeneous case. The Kolmogorov *forward* equation (Eq. (2.44)) can be written as

$$\dot{\mathbf{P}}(t) = \frac{\mathrm{d}\,\mathbf{P}(t)}{\mathrm{d}\,t} = \mathbf{P}(t)\mathbf{Q}\,. \tag{2.54}$$

The Kolmogorov *backward* equation in the homogeneous case results in matrix form as

$$\dot{\mathbf{P}}(t) = \frac{\mathrm{d}\,\mathbf{P}(t)}{\mathrm{d}\,t} = \mathbf{Q}\mathbf{P}(t)\,. \tag{2.55}$$

As in the discrete-time case, often the *steady-state probability vector* of a CTMC is of primary interest. The required properties of the steady-state probability vector, which is also called the *equilibrium probability vector*, are equivalent to the discrete-time case. For all states $i \in S$, the steady-state probabilities π_i are:

1. Independent of time t
2. Independent of the initial state probability vector $\boldsymbol{\pi}(0)$
3. Strictly positive, $\pi_i > 0$
4. Given as the time limits, $\pi_i = \lim_{t\to\infty} \pi_i(t) = \lim_{t\to\infty} p_{ji}(t)$, of the state probabilities $\pi_i(t)$ and of the transition probabilities $p_{ji}(t)$, respectively

If existing for a given CTMC, the steady-state probabilities are independent of time, we immediately get

$$\lim_{t\to\infty} \frac{\mathrm{d}\,\boldsymbol{\pi}(t)}{\mathrm{d}\,t} = 0\,. \tag{2.56}$$

Under condition (2.56), the differential equation (2.51) for determining the unconditional state probabilities resolves to a much simpler *system of linear equations*:

$$0 = \sum_{i\in S} q_{ij}\pi_i\,, \quad \forall j \in S\,. \tag{2.57}$$

In vector–matrix form, we get accordingly

$$\mathbf{0} = \boldsymbol{\pi}\mathbf{Q}\,. \tag{2.58}$$

Definition 2.13 In analogy to the discrete-time case, a CTMC for which a unique steady-state probability vector exists is called an *ergodic* CTMC.

The strictly positive steady-state probabilities can be gained by the unique solution of Eq. (2.58), when an additional normalization condition is imposed.[9] To express it in vector form, we introduce the *unit vector* $\mathbf{1} = [1, 1, \ldots, 1]^T$ so that the following relation holds:

$$\boldsymbol{\pi}\mathbf{1} = \sum_{i\in S} \pi_i = 1\,. \tag{2.59}$$

Another possibility for determining the steady-state probabilities π_i for all states $i \in S$ of a CTMC, is to take advantage of a well-known relation between the π_i and the *mean recurrence time*[10] $M_i < \infty$, that is, the mean time elapsed between two successive visits of the CTMC to state i:

$$\pi_i = -\frac{1}{M_i q_{ii}}\,, \quad \forall i \in S\,. \tag{2.60}$$

In the time-homogeneous case, we can derive from Eq. (2.41) that for any $j \in S$ $q_{jj} = -\sum_{i,i\neq j} q_{ji}$ and from Eq. (2.57) we get $\sum_{i,i\neq j} q_{ij}\pi_i = -q_{jj}\pi_j$. Putting these together immediately yields the system of *global balance equations*:

$$\sum_{i,i\neq j} q_{ij}\pi_i = \pi_j \sum_{i,i\neq j} q_{ji}\,, \quad \forall j \in S\,. \tag{2.61}$$

[9] Note that besides the trivial solution $\pi_i = 0$, $\forall i \in S$, any vector, obtained by multiplying a solution of Eq. (2.58) by an arbitrary real-valued constant, would also yield a solution of Eq. (2.58).

[10] In contrast to DTMCs, where lowercase notation is used to refer to recurrence time, uppercase notation is used for CTMCs.

On the left-hand side of Eq. (2.61), the *total flow* from any other state $i \in S$ *into* state j is captured. On the right-hand side, the *total flow out* of state j into any other state i is summarized. The flows are *balanced* in steady state, i.e., they are in equilibrium.

The conditions under which a CTMC is called ergodic are similar to those for a DTMC. Therefore, we can briefly summarize the criteria for classifying CTMCs and for characterizing their states.

2.1.2.2.1 Classifications of CTMC

Definition 2.14 As for DTMCs, we call a CTMC *irreducible* if every state i is *reachable* from every other state j, where $i, j \in S$; that is, $\forall i, j, i \neq j, \exists t : p_{ji}(t) > 0$. In other words, no proper subset $\hat{S} \subset S$, $\hat{S} \neq S$, of state space S exists, such that $\sum_{j \in \hat{S}} \sum_{i \in S \setminus \hat{S}} q_{ji} = 0$.

Definition 2.15 An irreducible, homogeneous CTMC is called *ergodic* if and only if the unique steady-state probability vector $\boldsymbol{\pi}$ exists.

As opposed to DTMCs, CTMCs cannot be periodic. Therefore, it can be shown that for an irreducible, homogeneous CTMC:

- The limits $\tilde{\pi}_i = \lim_{t \to \infty} \pi_i(t) = lim_{t \to \infty} p_{ji}(t)$ exist $\forall i, j \in S$ and are independent of the initial probability vector $\boldsymbol{\pi}(0)$. [11]
- The steady-state probability vector $\boldsymbol{\pi}$, if existing, can be uniquely determined by the solution of the linear system of Eq. (2.58) constrained by normalization condition (2.59).
- A unique steady-state, or equilibrium, probability vector $\boldsymbol{\pi}$ exists, if the irreducible, homogeneous CTMC is finite.
- The mean recurrence times M_i are finite for all states $i \in S$, $M_i < \infty$, if the steady-state probability vector exists.

2.1.2.2.2 CTMC State Sojourn Times

We have already mentioned that the distribution of the state *sojourn times* of a homogeneous CTMC must have the memoryless property. Since the exponential distribution is the only continuous distribution with this property, the random variables denoting the sojourn times, or holding times, must be exponentially distributed. Note that the same is true for the random variable referred to as the *residual* state holding time, that is, the time remaining until the next state change occurs.[12] Furthermore, the means of the two random variables are equal to $1/(-q_{ii})$.

Let the random variable R_i denote either the sojourn time or the residual time in state i, then the CDF is given by

$$F_{Ri}(r) = 1 - \mathrm{e}^{q_{ii} r}, \quad r \geq 0. \tag{2.62}$$

[11] The limits do not necessarily constitute a steady-state probability vector.
[12] The residual time is often referred to as the forward recurrence time.

The mean value of R_i, the mean sojourn time or the mean residual time, is given by

$$E[R_i] = -\frac{1}{q_{ii}}, \tag{2.63}$$

where q_{ii} is defined in Eq. (2.40).

2.1.2.3 Recapitulation We have introduced Markov chains and indicated their modeling power. The most important feature of homogeneous Markov chains is their unique *memoryless* property that makes them remarkably attractive. Both continuous- and discrete-time Markov chains have been defined and their properties discussed.

The most important algorithms for computation of their state probabilities are discussed in following chapters. Different types of algorithms are related to different categories of Markov chains such as ergodic, absorbing, finite, or infinite chains. Furthermore, the algorithms can be divided into those applicable for computing the steady-state probabilities and those applicable for computing the time-dependent state probabilities. Others provide approximate solutions, often based on an implicit transformation of the state space. Typically, these methods fall into the categories of aggregation/disaggregation techniques. Note that this modeling approximation has to be discriminated from the mathematical properties of the core algorithms, which, in turn, can be numerically exact or approximate, independent of their relation to the underlying model. Typical examples include round-off errors in direct methods such as Gaussian elimination and convergence errors in iterative methods for the solution of linear systems.

2.2 PERFORMANCE MEASURES

We begin by introducing a simple example and then provide an introduction to Markov reward models as a means to obtain performance measures.

2.2.1 A Simple Example

As an example adapted from Heimann, Mittal, and Trivedi [HMT91], consider a multiprocessor system with n processor elements processing a given workload. Each processor is subject to failures with a *mean time to failure* (MTTF), $1/\gamma$. In case of a failure, recovery can successfully be performed with probability c. Typically, recovery takes a brief period of time with mean $1/\beta$. Sometimes, however, the system does not successfully recover from a processor failure and suffers from a more severe impact. In this case, we assume the system needs to be rebooted with longer average duration of $1/\alpha$. Probability c is called the *coverage factor* and is usually close to 1. Unsuccessful recovery is most commonly caused by error propagation when the effect of a failure is

3

Steady-State Solutions of Markov Chains

Some sections

In this chapter, we restrict ourselves to the computation of the steady-state probability vector[1] of *ergodic* Markov chains. Most of the literature on solution techniques of Markov chains assumes ergodicity of the underlying model. A comprehensive source on algorithms for steady-state solution techniques is the book by Stewart [Stew94].

From Eq. (2.15) and Eq. (2.58), we have $\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{P}$ and $\mathbf{0} = \boldsymbol{\pi}\mathbf{Q}$, respectively, as points of departure for the study of steady-state solution techniques. Equation (2.15) can be transformed so that

$$\mathbf{0} = \boldsymbol{\nu}(\mathbf{P} - \mathbf{I}). \tag{3.1}$$

Therefore, both for CTMC and DTMC, a linear system of the form

$$\mathbf{0} = \mathbf{x}\mathbf{A} \tag{3.2}$$

needs to be solved. Due to its type of entries representing the parameters of a Markov chain, matrix $\mathbf{A}$ is singular and it can be shown that $\mathbf{A}$ is of rank $n-1$ for any Markov chain of size $|S| = n$. It follows immediately that the resulting set of equations is not linearly independent and that one of the equations is redundant. To yield a unique, positive solution, we must impose a normalization condition on the solution $\mathbf{x}$ of equation $\mathbf{0} = \mathbf{x}\mathbf{A}$. One way to approach the solution of Eq. (3.2) is to directly incorporate the normalization

[1] For the sake of convenience we sometimes use the term "steady-state analysis" as a shorthand notation.

condition

$$\mathbf{x1} = 1 \tag{3.3}$$

into Eq. (3.2). This can be regarded as substituting one of the columns (say, the last column) of matrix $\mathbf{A}$ by the unit vector $\mathbf{1} = [1, 1, \ldots, 1]^T$. With a slight abuse of notation, we denote the new matrix also by $\mathbf{A}$. The resulting linear system of nonhomogeneous equations is

$$\mathbf{b} = \mathbf{xA}, \quad \mathbf{b} = [0, 0, \ldots, 0, 1]. \tag{3.4}$$

An alternative to solving Eq. (3.2) is to separately consider normalization Eq. (3.3) as an additional step in numerical computations. We demonstrate both ways when example studies are presented. It is worthwhile pointing out that for any given ergodic CTMC, a DTMC can be constructed that yields an identical steady-state probability vector as the CTMC, and vice versa. Given the generator matrix $\mathbf{Q} = [q_{ij}]$ of a CTMC, we can define

$$\mathbf{P} = \mathbf{Q}/q + \mathbf{I}, \tag{3.5}$$

where q is chosen such that $q > \max_{i,j \in S} |q_{ij}|$. Setting $q = \max_{i,j \in S} |q_{ij}|$ should be avoided in order to assure *aperiodicity* of the resulting DTMC [GLT87]. The resulting matrix $\mathbf{P}$ can be used to determine the steady-state probability vector $\boldsymbol{\pi} = \boldsymbol{\nu}$, by solving $\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{P}$ and $\boldsymbol{\nu}\mathbf{1} = 1$. This method, used to reduce a CTMC to a DTMC, is called *randomization* or sometimes *uniformization* in the literature. If, on the other hand, a transition probability matrix $\mathbf{P}$ of an ergodic DTMC were given, a generator matrix $\mathbf{Q}$ of a CTMC can be defined by

$$\mathbf{Q} = \mathbf{P} - \mathbf{I}. \tag{3.6}$$

By solving $\mathbf{0} = \boldsymbol{\pi}\mathbf{Q}$ under the condition $\boldsymbol{\pi}\mathbf{1} = 1$, the desired steady-state probability vector $\boldsymbol{\pi} = \boldsymbol{\nu}$ can be obtained.

To determine the steady-state probabilities of finite Markov chains, three different approaches for the solution of a linear system of the form $\mathbf{0} = \mathbf{xA}$ are commonly used: *direct* or *iterative numerical* methods and techniques that yield *closed-form* results. Both types of numerical methods have merits of their own. Whereas direct methods yield exact results,[2] iterative methods are generally more efficient, both in time and space. Disadvantages of iterative methods are that for some of these methods no guarantee of convergence can be given in general and that determination of suitable error bounds for termination of the iterations is not always easy. Since iterative methods are considerably more efficient in solving Markov chains, they are commonly used for larger models. For smaller models with fewer than a few thousand states, direct methods are reliable and accurate. Though closed-form results are highly desirable, they can be obtained for only a small class of models that have some structure in their matrix.

[2]Modulo round-off errors resulting from finite precision arithmetic.

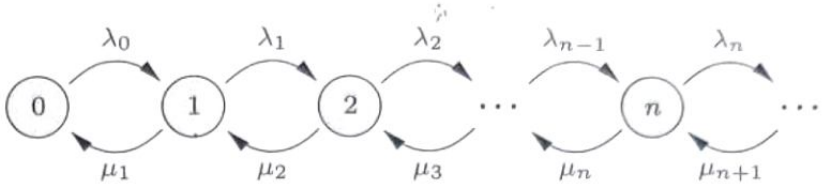

Fig. 3.1 Birth–death process.

Problem 3.1 Show that $\mathbf{P} - \mathbf{I}$ has the properties of a CTMC generator matrix.

Problem 3.2 Show that $\mathbf{Q}/q+\mathbf{I}$ has the properties of a stochastic (DTMC) matrix.

Problem 3.3 Define a CTMC and its generator matrix $\mathbf{Q}$ so that the corresponding DTMC would be periodic if randomization were applied with $q = \max_{i,j \in S} |q_{ij}|$ in Eq. (3.5).

3.1 SOLUTION FOR A BIRTH–DEATH PROCESS

Birth–death processes are Markov chains where transitions are allowed only between neighboring states. We treat the continuous-time case here, but analogous results for the discrete-time case are easily obtained (see [Triv01] for details).

A one-dimensional birth–death process is shown in Fig. 3.1 and its generator matrix is shown as Eq. (3.7):

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} . \tag{3.7}$$

The transition rates $\lambda_k, k \geq 0$ are state-dependent *birth rates* and $\mu_l, l \geq 1$, are referred to as state dependent *death rates*. Assuming ergodicity, the steady-state probabilities of CTMCs of the form depicted in Fig. 3.1 can be uniquely determined from the solution of Eq. (2.58):

$$0 = -\pi_0\lambda_0 + \pi_1\mu_1 , \tag{3.8}$$

$$0 = -\pi_k(\lambda_k + \mu_k) + \pi_{k-1}\lambda_{k-1} + \pi_{k+1}\mu_{k+1} , \quad k \geq 1 . \tag{3.9}$$

Solving Eq. (3.8) for π_1, and then using this result for substitution with $k = 1$ in Eq. (3.9) and solving it for π_2 yields

$$\pi_1 = \frac{\lambda_0}{\mu_1}\pi_0 , \qquad \pi_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2}\pi_0 . \tag{3.10}$$

Equation (3.10) together with Eq. (3.9) suggest a general solution of the following form:

$$\pi_k = \pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}, \quad k \geq 1. \tag{3.11}$$

Indeed, Eq. (3.11) provides the unique solution of a one-dimensional birth–death process. Since it is not difficult to prove this hypothesis by induction, we leave this as an exercise to the reader. From the law of total probability, $\sum_i \pi_i = 1$, we get for the probability π_0 of the CTMC being in State 0:

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} = \frac{1}{\sum_{k=0}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}. \tag{3.12}$$

The condition for convergence of the series in the denominator of Eq. (3.12), which is also the condition for the ergodicity of the birth–death CTMC, is

$$\exists k_0, \quad \forall k > k_0 : \quad \frac{\lambda_k}{\mu_k} < 1. \tag{3.13}$$

Equations (3.11) and (3.12) are used extensively in Chapter 6 to determine the probabilities π_k for many different queueing systems. These probabilities are then used to calculate performance measures such as mean queue length, or mean waiting time for these queueing systems. We deal with multidimensional birth–death processes in Section 3.2 and in Chapter 7.

A special case of a birth–death process arises from M/M/1 queueing systems that will be treated in Chapter 7. These systems can be modeled as birth–death processes with state-independent birth and death rates:

$$\lambda_k = \lambda, \quad k \geq 0, \tag{3.14}$$

$$\mu_l = \mu, \quad l \geq 1. \tag{3.15}$$

Equation (3.11) then boils down to the geometric pmf:

$$\pi_k = \pi_0 \left(\frac{\lambda}{\mu}\right)^k, \quad k \geq 0, \tag{3.16}$$

Problem 3.4 Consider a discrete-time birth–death process with birth probability b_i, the death probability d_i, and no state change probability $1 - b_i - d_i$ in state i. Derive expressions for the steady-state probabilities and conditions for convergence [Triv01].

6

Single Station Queueing Systems

A single station queueing system, as shown in Fig. 6.1, consists of a queueing buffer of finite or infinite size and one or more identical servers. Such an elementary queueing system is also referred to as a service station or, simply, as a node. A server can only serve one customer at a time and hence, it is

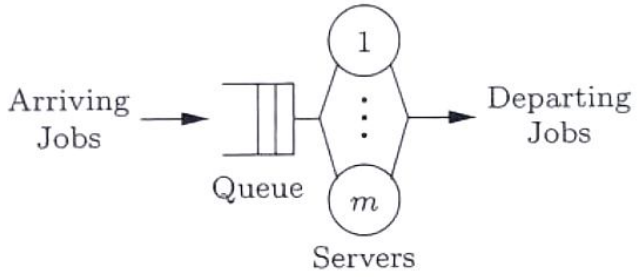

Fig. 6.1 Service station with m servers (a multiple server station).

either in a "busy" or an "idle" state. If all servers are busy upon the arrival of a customer, the newly arriving customer is buffered, assuming that buffer space is available, and waits for its turn. When the customer currently in service departs, one of the waiting customers is selected for service according to a *queueing (or scheduling) discipline.* An elementary queueing system is further described by an arrival process, which can be characterized by its sequence of interarrival time random variables $\{A_1, A_2, \cdots\}$. It is common to assume that the sequence of interarrival times is independent and identically distributed, leading to an arrival process that is known as a renewal process [Triv01]. Special classes of arrival processes are introduced in Section 6.8. The distribution function of interarrival times can be continuous or discrete.

We deal only with the former case in this book. For information related to discrete interarrival time distributions, the reader may consult [Dadu96].

The average interarrival time is denoted by $E[A] = \overline{T}_A$ and its reciprocal by the average arrival rate λ:

$$\lambda = \frac{1}{\overline{T}_A}. \tag{6.1}$$

The most common interarrival time distribution is the exponential, in which case the arrival process is Poisson. The sequence $\{B_1, B_2, \ldots\}$ of service times of successive jobs also needs to be specified. We assume that this sequence is also a set of independent random variables with a common distribution function. The mean service time $E[B]$ is denoted by $\overline{T}_B$ and its reciprocal by the service rate μ:

$$\mu = \frac{1}{\overline{T}_B}. \tag{6.2}$$

6.1 NOTATION

6.1.1 Kendall's Notation

The following notation, known as Kendall's notation, is widely used to describe elementary queueing systems:

A/B/m - queueing discipline,

where A indicates the distribution of the interarrival times, B denotes the distribution of the service times, and m is the number of servers ($m \geq 1$). The following symbols are normally used for A and B:

M	Exponential distribution (memoryless property)
E_k	Erlang distribution with k phases
H_k	Hyperexponential distribution with k phases
C_k	Cox distribution with k phases
D	Deterministic distribution, i.e., the interarrival time or service time is constant
G	General distribution
GI	General distribution with independent interarrival times

Due to proliferation of high-speed networks, there is considerable interest in traffic arrival processes where successive arrivals are correlated. Such non-GI arrival processes include the Markov modulated Poisson process (MMPP) (see Section 6.8.2), Markovian arrival process (see Section 6.8.3), or batch Markovian arrival process (BMAP) (see Section 6.8.4).

The queueing discipline or service strategy determines which job is selected from the queue for processing when a server becomes available. Some commonly used queueing disciplines are:

FCFS (First-Come-First-Served): If no queueing discipline is given in the Kendall notation, then the default is assumed to be the FCFS discipline. The jobs are served in the order of their arrival.

LCFS (Last-Come-First-Served): The job that arrived last is served next.

SIRO (Service-In-Random-Order): The job to be served next is selected at random.

RR (Round Robin): If the servicing of a job is not completed at the end of a time slice of specified length, the job is preempted and returns to the queue, which is served according to FCFS. This action is repeated until the job service is completed.

PS (Processor Sharing): This strategy corresponds to round robin with infinitesimally small time slices. It is as if all jobs are served simultaneously and the service time is increased correspondingly.

IS (Infinite Server): There is an ample number of servers so that no queue ever forms.

Static Priorities: The selection depends on priorities that are permanently assigned to the job. Within a class of jobs with the same priority, FCFS is used to select the next job to be processed.

Dynamic Priorities: The selection depends on dynamic priorities that alter with the passing of time.

Preemption: If priority or LCFS discipline is used, then the job currently being processed is interrupted and preempted if there is a job in the queue with a higher priority.

As an example of Kendall's notation, the expression

M/G/1-LCFS preemptive resume (PR)

describes an elementary queueing system with exponentially distributed inter arrival times, arbitrarily distributed service times, and a single server. The queueing discipline is LCFS where a newly arriving job interrupts the job currently being processed and replaces it in the server. The servicing of the job that was interrupted is resumed only after all jobs that arrived after it have completed service.

Kendall's notation can be extended in various ways. An additional parameter is often introduced to represent the number of places in the queue (if the queue is finite) and we get the extended notation

A/B/m/K-queueing discipline,

where K is the capacity of the station (queue + server). This means that if the number of jobs at server and queue is K, a newly arriving job is lost.

6.1.2 Performance Measures

The different types of queueing systems are analyzed mathematically to determine performance measures from the description of the system. Because a queueing model represents a dynamic system, the values of the performance measures vary with time. Normally, however, we are content with the results in the steady state. The system is said to be in steady state when all transient behavior has ended, the system has settled down, and the values of the performance measures are independent of time. The system is then said to be in statistical equilibrium; i.e., the rate at which jobs enter the system is equal to the rate at which jobs leave the system. Such a system is also called a *stable system*. Transient solutions of simple queueing systems are available in closed form, but for more general cases, we need to resort to Markov chain techniques as described in Chapter 5. Recall that the generation and the solution of large Markov chains can be automated via stochastic reward nets [MuTr92].

The most important performance measures are:

Probability of the Number of Jobs in the System π_k: It is often possible to describe the behavior of a queueing system by means of the probability vector of the number of jobs in the system π_k. The mean values of most of the other interesting performance measures can be deduced from π_k:

$$\pi_k = P[\text{there are } k \text{ jobs in the system}] .$$

Utilization ρ: If the queueing system consists of a single server, then the utilization ρ is the fraction of the time in which the server is busy, i.e., occupied. In case there is no limit on the number of jobs in the single server queue, the server utilization is given by

$$\rho = \frac{\text{mean service time}}{\text{mean interarrival time}} = \frac{\text{arrival rate}}{\text{service rate}} = \frac{\lambda}{\mu} . \tag{6.3}$$

The utilization of a service station with multiple servers is the mean fraction of active servers. Since $m\mu$ is the overall service rate, we have:

$$\rho = \frac{\lambda}{m\mu} , \tag{6.4}$$

and ρ can be used to formulate the condition for stationary behavior mentioned previously. The condition for stability is

$$\rho < 1 ; \tag{6.5}$$

i.e., on average the number of jobs that arrive in a unit of time must be less than the number of jobs that can be processed. All the results given in Chapters 6 – 10 apply only to stable systems.

Throughput λ: The throughput of an elementary queueing system is defined as the mean number of jobs whose processing is completed in a single unit of time, i.e., the departure rate. Since the departure rate is equal to the arrival rate λ for a queueing system in statistical equilibrium, the throughput is given by

$$\lambda = m \cdot \rho \cdot \mu \tag{6.6}$$

in accordance with Eq. (6.4). We note that in the case of finite buffer queueing system, throughput can be different from the external arrival rate.

Response Time T: The response time, also known as the sojourn time, is the total time that a job spends in the queueing system.

Waiting Time W: The waiting time is the time that a job spends in a queue waiting to be serviced. Therefore, we have

$$\text{Response time} = \text{waiting time} + \text{service time}\,.$$

Since W and T are usually random numbers, their mean is calculated. Then

$$\overline{T} = \overline{W} + \frac{1}{\mu}\,. \tag{6.7}$$

The distribution functions of the waiting time, $F_W(x)$, and the response time, $F_T(x)$, are also sometimes required.

Queue Length Q: The queue length, Q, is the number of jobs in the queue.

Number of Jobs in the System K: The number of jobs in the queueing system is represented by K. Then

$$\overline{K} = \sum_{k=1}^{\infty} k \cdot \pi_k\,. \tag{6.8}$$

The mean number of jobs in the queueing system $\overline{K}$ and the mean queue length $\overline{Q}$ can be calculated using one of the most important theorems of queueing theory, *Little's theorem*:

$$\overline{K} = \lambda \overline{T}\,, \tag{6.9}$$

$$\text{and} \quad \overline{Q} = \lambda \overline{W}\,. \tag{6.10}$$

Little's theorem is valid for all queueing disciplines and arbitrary GI/G/m queue. The proof is given in [Litt61].

6.2 MARKOVIAN QUEUES

6.2.1 The M/M/1 Queue

Recall that in this case, the arrival process is Poisson, the service times are exponentially distributed, and there is a single server. The system can be modeled as a birth–death process with birth rate (arrival rate) λ and a constant death rate (service rate) μ. We assume that $\lambda < \mu$ so the underlying CTMC is ergodic and hence the queueing system is stable. Then using Eq. (3.12), we obtain the steady-state probability of the system being empty:

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda}{\mu}} = \frac{1}{1 + \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k},$$

which can be simplified to

$$\pi_0 = \frac{1}{1 + \frac{\lambda/\mu}{1-\lambda/\mu}} = 1 - \frac{\lambda}{\mu}.$$

From Eq. (3.11), for the steady-state probability that there are k jobs in the system we get:

$$\pi_k = \pi_0 \left(\frac{\lambda}{\mu}\right)^k, \quad k \geq 0,$$

$$\pi_k = \left(1 - \frac{\lambda}{\mu}\right) \cdot \left(\frac{\lambda}{\mu}\right)^k,$$

or with the utilization $\rho = \lambda/\mu$ we obtain:

$$\pi_0 = 1 - \rho \tag{6.11}$$

and

$$\pi_k = (1-\rho)\rho^k, \tag{6.12}$$

the probability mass function (pmf) of the modified geometric random variable. In Fig. 6.2, we plot this pmf for $\rho = 1/2$. The mean number of jobs is obtained using Eqs. (6.12) and (6.8):

$$\overline{K} = \frac{\rho}{1-\rho}. \tag{6.13}$$

In Fig. 6.3, the mean number of jobs is plotted as a function of the utilization ρ. This is the typical behavior of all queueing systems.

From Eq. (1.10) we obtain the variance of the number of jobs in the system:

$$\sigma_K^2 = \frac{\rho}{(1-\rho)^2} \tag{6.14}$$

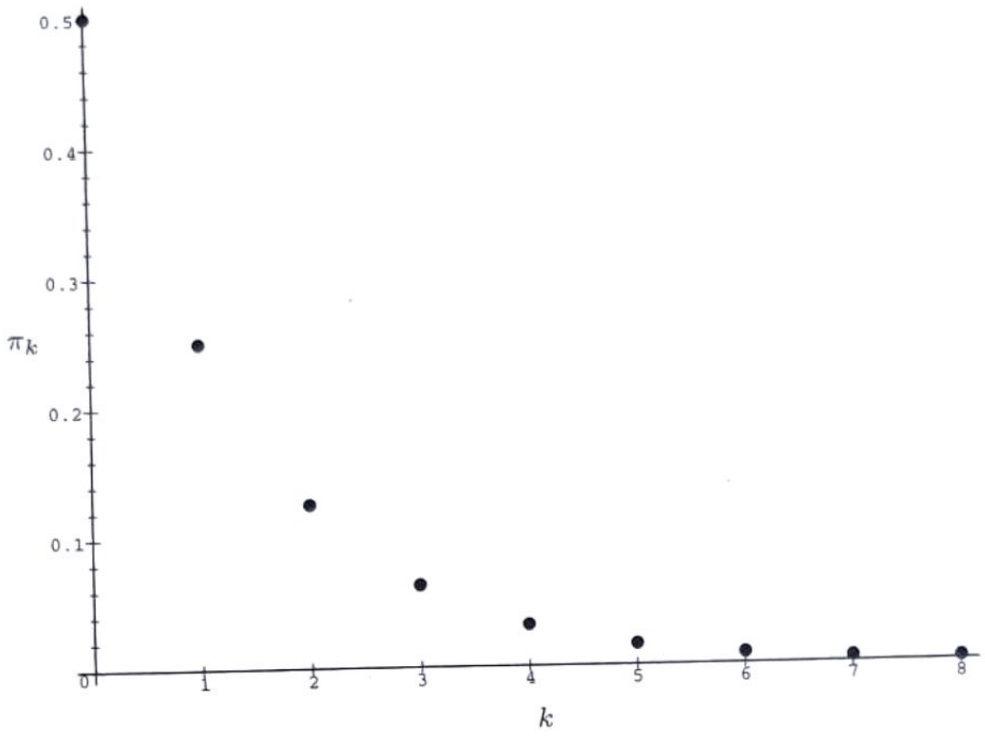

Fig. 6.2 The solution for π_k in an M/M/1 queue.

and the coefficient of variation:

$$c_K = \frac{\sigma_K}{\overline{K}} = \frac{1}{\sqrt{\rho}} .$$

With Little's theorem (Eqs. (6.9) and (6.10)) we get the following for the mean response time:

$$\overline{T} = \frac{1/\mu}{1-\rho} , \tag{6.15}$$

with Eq. (6.7) for the mean waiting time:

$$\overline{W} = \frac{\rho/\mu}{1-\rho} , \tag{6.16}$$

and with Little's theorem again for the mean queue length:

$$\overline{Q} = \frac{\rho^2}{1-\rho} . \tag{6.17}$$

The same formulae are valid for M/G/1-PS and M/G/1-LCFS preemptive resume (see [Lave83]). For M/M/1-FCFS we can get a relation for the response time distribution if we consider the response time as the sum of $k+1$ independent exponentially distributed random variables [Triv01]:

$$X + X_1 + X_2 + \cdots + X_k ,$$

where X is the service time of the tagged job, X_1 is the remaining service time of the job in service when the tagged job arrives, and $X_2, \ldots, X_k$ are the

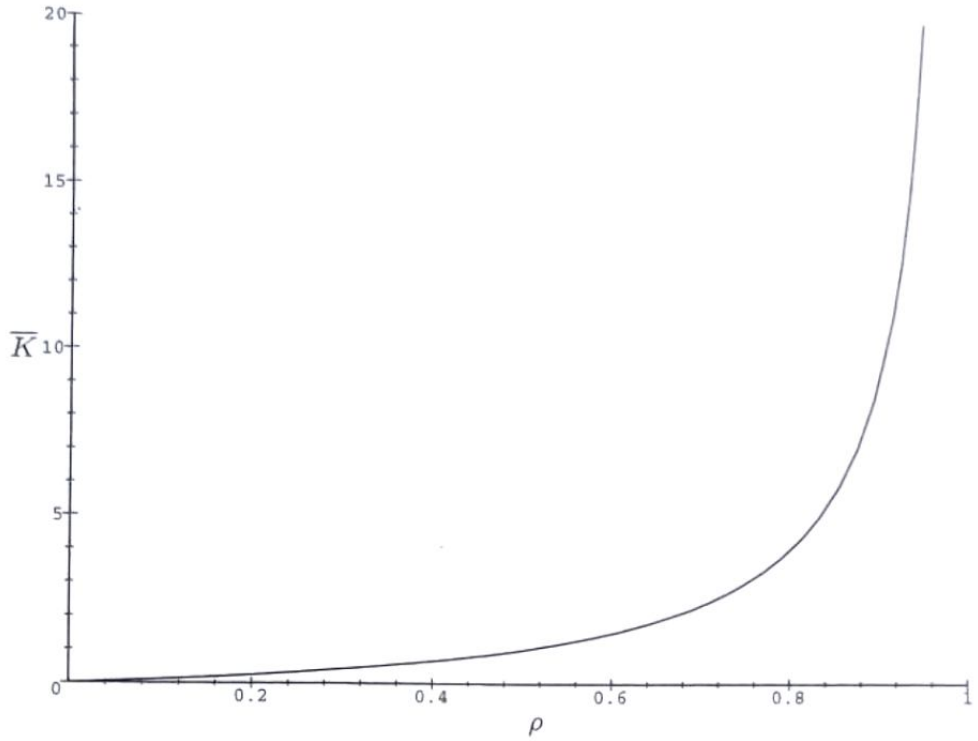

Fig. 6.3 The mean number of jobs $\overline{K}$ in an M/M/1 queue.

service times of the jobs in the queue; each of these is exponentially distributed with parameter μ. Note that X_1 is also exponentially distributed with rate μ due to the memoryless property of the exponential distribution. Noting that the Laplace–Stieltjes transform (LST) of the exponentially distributed service time (see Table 1.5) is

$$L_X(s) = \frac{\mu}{\mu + s},$$

we get for the conditional LST of the response time:

$$L_{T|K}(s|k) = \left(\frac{\mu}{\mu + s}\right)^{k+1}.$$

Unconditioning using the steady-state probability Eq. (6.12), the LST of the response time is

$$\begin{aligned} L_T(s) &= \sum_{k=0}^{\infty} \left(\frac{\mu}{\mu + s}\right)^{k+1} \cdot (1-\rho)\rho^k \\ L_T(s) &= \frac{\mu(1-\rho)}{s + \mu(1-\rho)}. \end{aligned} \tag{6.18}$$

Thus, the response time T is exponentially distributed with the parameter $\mu(1-\rho)$:

$$F_T(x) = 1 - \mathrm{e}^{-\mu(1-\rho)x} \tag{6.19}$$

and the variance:

$$\mathrm{var}(T) = \frac{1}{\mu^2(1-\rho)^2}\,. \tag{6.20}$$

Similarly we get the distribution of the waiting time:

$$F_W(x) = \begin{cases} 1-\rho\,, & x = 0\,, \\ \\ 1-\rho\cdot \mathrm{e}^{-\mu(1-\rho)x}\,, & x > 0\,. \end{cases} \tag{6.21}$$

Thus, $F_W(0) = P(W=0) = 1-\rho$ is the mass at origin, corresponding to the probability that an arriving customer does not have to wait in the queue.

6.2.2 The M/M/∞ Queue

In an M/M/∞ queueing system we have a Poisson arrival process with arrival rate λ and an infinite number of servers with service rate μ each. If there are k jobs in the system, then the overall service rate is $k\mu$ because each arriving job immediately gets a server and does not have to wait. Once again, the underlying CTMC is a birth–death process. From Eq. (3.11) we obtain the steady-state probability of k jobs in the system:

$$\pi_k = \pi_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = \pi_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}\,;$$

with Eq. (3.12), we obtain the steady-state probability of no jobs in the system:

$$\pi_0 = \frac{1}{1+\sum\limits_{k=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{k!}} = \mathrm{e}^{-\frac{\lambda}{\mu}}\,, \tag{6.22}$$

and finally:

$$\pi_k = \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!}\cdot \mathrm{e}^{-\frac{\lambda}{\mu}}\,. \tag{6.23}$$

This is the Poisson pmf, and the expected number of jobs in the system is

$$\overline{K} = \frac{\lambda}{\mu}\,. \tag{6.24}$$

With Little's theorem the mean response time as expected is:

$$\overline{T} = \frac{1}{\mu}\,. \tag{6.25}$$

6.2.3 The M/M/m Queue

An M/M/m queueing system with arrival rate λ and service rate μ for each server can also be modeled as a birth–death process with

$$\lambda_k = \lambda\,, \quad k \geq 0\,,$$
$$\mu_k = \begin{cases} k\mu\,, & 0 \leq k \leq m\,, \\ m\mu\,, & m \leq k\,. \end{cases}$$

The condition for the queueing system to be stable (underlying CTMC to be ergodic) is $\lambda < m\mu$. The steady-state probabilities are given by (from Eq. (3.11))

$$\pi_k = \begin{cases} \pi_0 \displaystyle\prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = \pi_0 \left(\frac{\lambda}{\mu}\right)^k \cdot \frac{1}{k!}\,, & 0 \leq k \leq m\,, \\ \pi_0 \displaystyle\prod_{i=0}^{m-1} \frac{\lambda}{(i+1)\mu} \cdot \prod_{i=m}^{k-1} \frac{\lambda}{m\mu}\,, & k \geq m\,. \end{cases}$$

With an individual server utilization, $\rho = \lambda/(m\mu)$, we obtain

$$\pi_k = \begin{cases} \pi_0 \dfrac{(m\rho)^k}{k!}\,, & 0 \leq k \leq m\,, \\ \pi_0 \dfrac{\rho^k m^m}{m!}\,, & k \geq m\,, \end{cases} \tag{6.26}$$

and from Eq. (3.12) we obtain:

$$\pi_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \frac{1}{1-\rho}\right]^{-1}. \tag{6.27}$$

The steady-state probability that an arriving customer has to wait in the queue is given by

$$\begin{aligned} P_m = P(K \geq m) &= \sum_{k=m}^{\infty} \pi_k \\ &= \frac{(m\rho)^m}{m!(1-\rho)} \cdot \pi_0 \end{aligned} \tag{6.28}$$

Using Eqs. (6.26) and (6.8), for the mean number of jobs in the system we obtain:

$$\overline{K} = m\rho + \frac{\rho}{1-\rho} \cdot P_m\,, \tag{6.29}$$

and for the mean queue length we obtain:

$$\overline{Q} = \frac{\rho}{1-\rho} \cdot P_m \,. \tag{6.30}$$

From this the mean response time $\overline{T}$ and mean waiting time $\overline{W}$ by Little's theorem (Eqs. (6.9) and (6.10)) can be easily derived. A formula for the distribution of the waiting time is given in [GrHa85]:

$$F_W(x) = \begin{cases} 1 - P_m \,, & x = 0 \,, \\ 1 - P_m \cdot \mathrm{e}^{-m\mu(1-\rho)x} \,, & x > 0 \,. \end{cases} \tag{6.31}$$

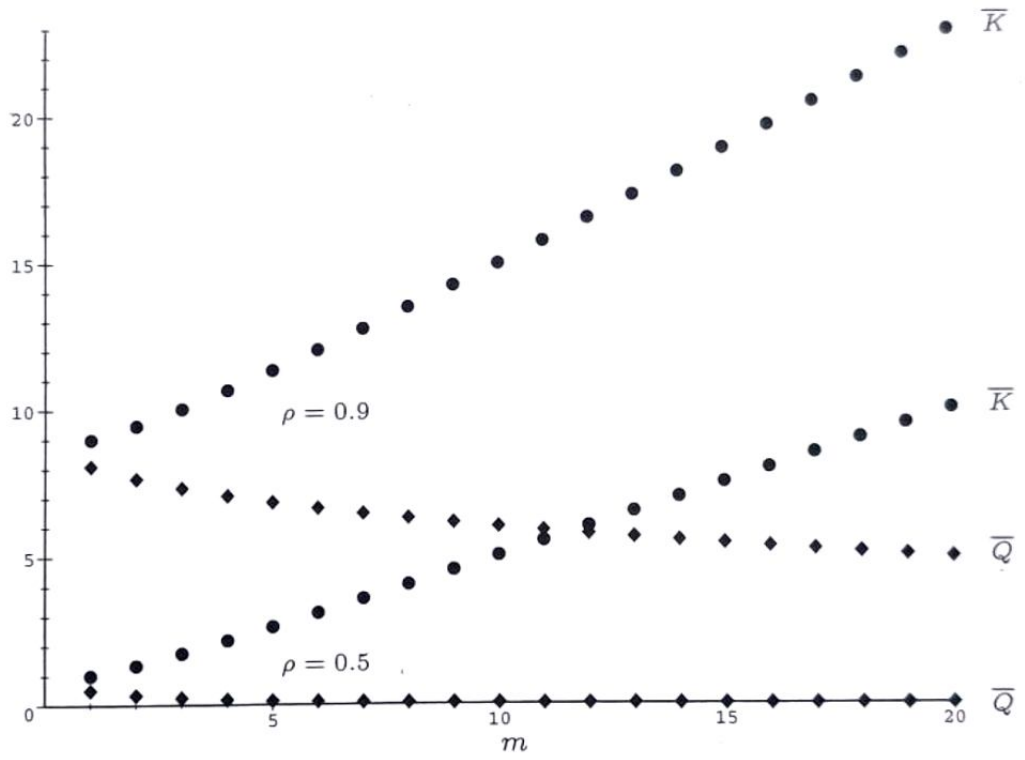

Fig. 6.4 Mean queue length $\overline{Q}$ and mean number of jobs in the system $\overline{K}$ as functions of the number of servers m.

Figure 6.4 shows the interesting and important property of an M/M/m queue that the mean number of jobs in the system $\overline{K}$ increases with the number of servers m if the server utilization is constant but the mean queue length $\overline{Q}$ decreases.

6.2.4 The M/M/1/K Finite Capacity Queue

In an M/M/1/K queueing system, the maximum number of jobs in the system is K, which implies a maximum queue length of $K-1$. An arriving job enters the queue if it finds fewer than K jobs in the system and is lost otherwise.