László Lakatos
László Szeidl
Miklós Telek

Introduction to Queueing Systems with Telecommunication Applications

Springer

Introduction to Queueing Systems with Telecommunication Applications

László Lakatos • László Szeidl • Miklós Telek

Introduction to Queueing Systems with Telecommunication Applications

Springer

László Lakatos
Eötvös Loránd University
Budapest, Hungary

Miklós Telek
Budapest University of Technology
and Economics
Budapest, Hungary

László Szeidl
Óbuda University
Budapest, Hungary

Széchenyi István University
Győr, Hungary

ISBN 978-1-4614-5316-1 ISBN 978-1-4614-5317-8 (eBook)
DOI 10.1007/978-1-4614-5317-8
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012951675

Mathematics Subject Classification (2010): 60K25, 68M20, 90B22

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The development of queueing theory dates back more than a century. Originally the concept was examined for the purpose of maximizing performance of telephone operation centers; however, it was realized soon enough that issues in that field that were solvable using mathematical models might arise in other areas of everyday life as well. Mathematical models, which serve to describe certain phenomena, quite often correspond with each other, regardless of the specific field for which they were originally developed, be that telephone operation centers, planning and management of emergency medical services, description of computer operation, banking services, transportation systems, or other areas. The common feature in these areas is that demands and services occur (also at an abstract level) with various contents depending on the given questions. In the course of modeling, irrespective of the meaning of demand and service in the modeled system, one is dealing with only moments and time intervals. Thus it can be concluded that, despite the diversity of problems, a common theoretical background and a mathematical toolkit can be relied upon that ensures the effective and multiple application of a theory. It is worth noting as an interesting aspect that the beginning of the development of queueing theory is closely connected to the appearance of telephone operation centers more than a century ago, as described previously; nevertheless, it still plays a significant role in the planning, modeling, and analyzing of telecommunication networks supplemented by up-to-date simulation methods and procedures.

The authors of this book have been conducting research and modeling in the theoretical and practical field of queueing theory for several decades and teaching in both bachelor's, master's, and doctoral programs in the Faculty of Informatics, Eötvös Loránd University, Faculty of Engineering Sciences, Széchenyi István University, John von Neuman Faculty of Informatics, Óbuda University and the Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics (all located in Hungary).

The various scientific backgrounds of the authors complement each other; therefore, both mathematical and engineering approaches are reflected in this book. The writing of this book was partly triggered by requests from undergraduate and Ph.D. students and by the suggestions of supportive colleagues, all of whom expressed the

necessity to write a book that could be directly applied to informatics, mathematics, and applied mathematics education as well as other fields. In considering the structure of the book, the authors tried to briefly summarize the necessary theoretical basis of probability theory and stochastic processes, which provide a uniform system of symbols and conventions to study and master the material presented here. At the end of Part I, the book provides a systematic and detailed treatment of Markov chains, renewal and regenerative processes, Markov chains, and Markov chains with special structures. Following the introductory chapters on probability theory and stochastic processes, we will disregard the various possible interpretations concerning the examples to emphasize terms, methodology, and analytical skills; therefore, we will provide the proofs for each of the given examples. We think that this structure will help readers to study the material more effectively since they may have different backgrounds and knowledge concerning this area. Regarding the basics of probability theory, we refer the interested reader to the books [21, 31, 38, 84]. With respect to the general results of stochastic processes and Markov chains, we refer the reader to the following comprehensive literature: [22, 26, 35, 36, 48, 49, 54, 71].

In Part II, the book introduces and considers the classic results of Markov and non-Markov queueing systems. Then queueing networks and applied queueing systems (analysis of ATM switches, conflict resolution methods of random access protocols, queueing systems with priorities, and repeated orders queueing systems) are analyzed. For more on the classic results of queueing theory, we refer the reader to [8, 20, 39, 51, 55, 69, 82], whereas in connection with the modern theory of queueing and telecommunication systems the following books may be consulted: [6, 7, 14–16, 34, 41, 47, 83], as well as results published mainly in journals and conference papers. The numerous exercises at the end of the chapters ensure a better understanding of the material.

A short appendix appears at the end of the book that sums up those special concepts and ideas that are used in the book and that help the reader to understand the material better.

This work was supported by the European Union and cofinanced by the European Social Fund under Grant TÁMOP 4.2.1/B-09/1/KMR-2010-0003 and by the OTKA Grant No. K-101150. The authors are indebted to the Publisher for the encouragement and the efficient editorial support.

The book is recommended for students and researchers studying and working in the field of queueing theory and its applications.

Budapest, Hungary

László Lakatos
László Szeidl
Miklós Telek

Contents

Part I
Introduction to Probability Theory and Stochastic Processes

Chapter 1
Introduction to Probability Theory

1.1 Summary of Basic Notions of Probability Theory

In this chapter we summarize the most important notions and facts of probability theory that are necessary for an elaboration of our topic. In the present summary, we will apply the more specific mathematical concepts and facts – mainly measure theory and analysis – only to the necessary extent while, however, maintaining mathematical precision.

Random Event We consider experiments whose outcomes are uncertain, where the totality of the circumstances that are or can be considered does not determine the outcome of the experiment. A set consisting of all possible outcomes is called a **sample space**. We define **random events** (**events** for short) as certain sets of outcomes (subsets of the sample space). It is assumed that the set of events is closed under countable set operations, and we assign probability to events only; they characterize the quantitative measure of the degree of uncertainty. Henceforth countable means finite or countably infinite.

Denote the sample space by $\Omega = \{\omega\}$. If Ω is countable, then the space Ω is called **discrete**. In a mathematical approach, events can be defined as subsets $A \subset \Omega$ of the possible outcomes Ω having the properties (σ-algebra properties) defined subsequently.

A given event A occurs in the course of an experiment if the outcome of the experiment belongs to the given event, that is, if an outcome $\omega \in A$ exists. An event is called simple if it contains only one outcome ω. It is always assumed that the whole set Ω and the empty set $\varnothing$ are events that are called a **certain** event and an **impossible event**, respectively.

Operation with Events; Notion of σ-Algebra Let A and B be two events. The **union** $A \cup B$ of A and B is defined as an event consisting of all elements $\omega \in \Omega$ belonging to either event A or B, i.e., $A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$.

The **intersection** (**product**) $A \cap B$ (AB) of events A and B is defined as an event consisting of all elements $\omega \in \Omega$ belonging to both A and B, i.e.,

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, DOI 10.1007/978-1-4614-5317-8_1, © Springer Science+Business Media, LLC 2013

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}.$$

The **difference** $A \backslash B$, which is not a symmetric operation, is defined as the set of all elements $\omega \in \Omega$ belonging to event A but not to event B, i.e.,

$$A \backslash B = \{\omega : \omega \in A \text{ and } \omega \notin B\}.$$

A **complementary event** $\overline{A}$ of A is defined as a set of all elements $\omega \in \Omega$ that does not belong to A, i.e.,

$$\overline{A} = \Omega \backslash A.$$

If $A \cap B = \oslash$, then sets A and B are said to be **disjoint** or **mutually exclusive**.

Note that the operations $\cup$ and $\cap$ satisfy the associative, commutative, and distributive properties

$$(A \cup B) \cup C = A \cup (B \cup C), \quad \text{and} \quad (A \cap B) \cap C = A \cap (B \cap C),$$
$$A \cup B = B \cup A, \quad \text{and} \quad A \cap B = B \cap A,$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad \text{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

DeMorgan identities are valid also for the operations union, intersection, and complementarity of events as follows:

$$\overline{A \cup B} = \overline{A} \cap \overline{B}, \qquad \overline{A \cap B} = \overline{A} \cup \overline{B}.$$

With the use of the preceding definitions introduced, we can define the notion of σ-algebra of events.

Definition 1.1. Let Ω be a nonempty (abstract) set, and let $\mathcal{A}$ be a certain family of subsets of the set Ω satisfying the following conditions:

(1) $\Omega \in \mathcal{A}$.
(2) If $A \in \mathcal{A}$, then $\overline{A} \in \mathcal{A}$.
(3) If $A_1, A_2, \ldots \in \mathcal{A}$ is a countable sequence of elements, then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}.$$

The family $\mathcal{A}$ of subsets of the set Ω satisfying conditions (1)–(3) is called a **σ-algebra**. The elements of $\mathcal{A}$ are called **random events**, or simply **events**.

Comment 1.2. *The pair* $(\Omega, \mathcal{A})$ *is usually called a* ***measurable space****, which forms the general mathematical basis of the notion of probability.*

Probability Space, Kolmogorov Axioms of Probability Theory Let Ω be a nonempty sample set, and let $\mathcal{A}$ be a given σ-algebra of subsets of Ω, i.e., the pair

$(\Omega, \mathcal{A})$ is a measurable space. A nonnegative number $\mathbf{P}(A)$ is assigned to all events A of σ-algebra satisfying the axioms as follows.

A1. $0 \le \mathbf{P}(A) \le 1$, $A \in \mathcal{A}$.
A2. $\mathbf{P}(\Omega) = 1$.
A3. If the events $A_i \in \mathcal{A}$, $i = 1, 2, \ldots$, are disjoint (i.e., $A_i A_j = \oslash$, $i \ne j$), then

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i).$$

The number $\mathbf{P}(A)$ is called the **probability** of event A, axioms A1, A2, and A3 are called the Kolmogorov axioms, and the triplet $(\Omega, \mathcal{A}, \mathbf{P})$ is called the probability space. As usual, axiom A3 is called the σ-additivity property of the probability. The probability space characterizes completely a random experiment.

Comment 1.3. *In the measure theory context of probability theory, the function* $\boldsymbol{P}$ *defined on* $\mathcal{A}$ *is called a probability measure. Conditions A1–A3 ensure that* $\boldsymbol{P}$ *is nonnegative and that* σ *is an additive and normed [*$\boldsymbol{P}(\Omega) = 1$*] set function on* $\mathcal{A}$*, i.e., a normed measure on* $\mathcal{A}$*. Our discussion basically does not require the direct use of measure theory, but some assertions cited in this work essentially depend on this theory.*

Main Properties of Probability Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. The following properties of probability are valid for all probability spaces.

Elementary properties:

(a) The probability of an impossible event is zero, i.e.,
$\mathbf{P}(\oslash) = 0$.
(b) $\mathbf{P}\left(\overline{A}\right) = 1 - \mathbf{P}(A)$ for all $A \in \mathcal{A}$.
(c) If the relationship $A \subseteq B$ is satisfied for given events $A, B \in \mathcal{A}$, then
$\mathbf{P}(A) \le \mathbf{P}(B)$,
$\mathbf{P}(B - A) = \mathbf{P}(B) - \mathbf{P}(A)$.

Definition 1.4. A collection $\{A_i,\ i \in I\}$ of a countable set of events is called a **complete system** of events if $A_i,\ i \in I$ are disjoint (i.e., $A_i \cap A_j = \oslash$ if $i \ne j$, $i, j \in I$) and $\bigcup\limits_{i \in I} A_i = \Omega$.

Comment 1.5. *If the collection of events* $\{A_i,\ i \in I\}$ *forms a complete system of events, then*

$$\boldsymbol{P}\left(\bigcup_{i \in I} A_i\right) = 1.$$

Probability of Sum of Events, Poincaré Formula For any events A and B it is true that

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(AB).$$

Using this relation, a more general formula, called the **Poincaré formula**, can be proved. Let n be a positive integer number; then, for any events $A_1, A_2, \ldots, A_i \in \mathcal{A}$,

$$\mathbf{P}(A_1 + \ldots + A_n) = \sum_{k=1}^{n} (-1)^{k-1} S_k^{(n)},$$

where $S_k^{(n)} = \sum_{1 \le i_1 \le \ldots \le i_k \le n} \mathbf{P}\left(A_{i_1} \ldots A_{i_k}\right)$.

Subadditive Property of Probability For any countable set of events $\{A_i,\ i \in I\}$ the inequality

$$\mathbf{P}\left(\bigcup_{i \in I} A_i\right) \le \sum_{i \in I} \mathbf{P}(A_i)$$

is true.

Continuity Properties of Probability Continuity properties of probability are valid for monotonically sequences of events, each of which is equivalent to axiom A3 of probability. A sequence of events $A_1, A_2, \ldots$ is called monotonically increasing (resp. decreasing) if $A_1 \subset A_2 \subset \ldots$ (resp. $A_1 \supset A_2 \supset \ldots$).

Theorem 1.6. *If the sequence of events $A_1, A_2, \ldots$ is monotonically decreasing, then*

$$\boldsymbol{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n \to \infty} \boldsymbol{P}(A_n).$$

If the sequence of events $A_1, A_2, \ldots$ is monotonically increasing, then

$$\boldsymbol{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \to \infty} \boldsymbol{P}(A_n).$$

Conditional Probability and Its Properties, Independence of Events In practice, the following obvious question arises: if we know that event B occurs (i.e., the outcome is in $B \in \mathcal{A}$), what is the probability that the outcome is in $A \in \mathcal{A}$? In other words, how does the occurrence of an event B influence the occurrence of another event A? This effect is characterized by the notion of conditional probability $\mathbf{P}(A|B)$ as follows.

Definition 1.7. Let A and B be two events, and assume that $\mathbf{P}(B) > 0$. The quantity

$$\mathbf{P}(A|B) = \mathbf{P}(AB)/\mathbf{P}(B)$$

is called the **conditional probability of A given B**.

It is easy to verify that the conditional probability possesses the following properties:

1. $0 \le \mathbf{P}(A|B) \le 1$.
2. $\mathbf{P}(B|B) = 1$.
3. If the events $A_1, A_2, \ldots$ are disjoint, then

$$\mathbf{P}\left(\sum_{i=1}^{\infty} A_i | B\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i|B).$$

4. The definition of conditional probability $\mathbf{P}(A|B) = \mathbf{P}(AB)/\mathbf{P}(B)$ is equivalent to the so-called theorem of multiplication

$$\mathbf{P}(AB) = \mathbf{P}(A|B)\mathbf{P}(B) \text{ and } \mathbf{P}(AB) = \mathbf{P}(B|A)\mathbf{P}(A).$$

Note that these equations are valid in the cases $\mathbf{P}(B) = 0$ and $\mathbf{P}(A) = 0$ as well.

One of the most important concepts of probability theory, the independence of events, is defined as follows.

Definition 1.8. We say that events A and B are **independent** if the equation

$$\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B)$$

is satisfied.

Comment 1.9. *If A and B are independent events and $\boldsymbol{P}(B) > 0$, then the conditional probability $\boldsymbol{P}(A|B)$ does not depend on event B since*

$$\boldsymbol{P}(A|B) = \frac{\boldsymbol{P}(AB)}{\boldsymbol{P}(B)} = \frac{\boldsymbol{P}(A)\boldsymbol{P}(B)}{\boldsymbol{P}(B)} = \boldsymbol{P}(A).$$

This relation means that knowing that an event B occurs does not change the probability of another event A.

The notion of independence of an arbitrary collection A_i, $i \in I$ of events is defined as follows.

Definition 1.10. A given collection of events A_i, $i \in I$ is said to be **mutually independent (independent** for short) if, having chosen from among them any finite number of events, the probability of the product of the chosen events equals the product of the probabilities of the given events. In other words, if $\{i_1, \ldots, i_k\}$ is any subcollection of I, then one has

$$\mathbf{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right) = \mathbf{P}(A_{i_1}) \ldots \mathbf{P}\left(A_{i_k}\right).$$

This notion of independence is stricter when pairs are concerned since it is easy to create an example where pairwise independence occurs but mutual independence does not.

Example 1.11. We roll two dice and denote the pair of results by

$$(\omega_1, \omega_2) \in \Omega = \{(i, j), 1 \leq i, j \leq 6\}.$$

The number of elements of the set Ω is $|\Omega| = 36$, and we assume that the dice are standard, that is, $P\{(\omega_1, \omega_2)\} = 1/36$ for every $(\omega_1, \omega_2) \in \Omega$. Events A_1, A_2, and A_3 are defined as follows:

$$\begin{aligned} A_1 &= \{\text{the result of the first die is even}\}, \\ A_2 &= \{\text{the result of the second die is odd}\}, \\ A_3 &= \{\text{both the first and second dice are odd or both of them are even}\}. \end{aligned}$$

We check that events A_1, A_2, and A_3 are pairwise independent, but they are not (mutually) independent. It is clear that

$$\begin{aligned} A_1 &= \{(2,1), \ldots, (2,6), (4,1), \ldots, (4,6), (6,1), \ldots, (6,6)\}, \\ A_2 &= \{(1,1), \ldots, (6,1), (1,3), \ldots, (6,3), (1,5), \ldots, (6,5)\}, \\ A_3 &= \{(1,1), (1,3), (1,5), (2,2), (2,4), (2,6), (3,1), (3,3), \\ &\quad (3,5), \ldots, (6,2), (6,4)(6,6)\}, \end{aligned}$$

thus

$$|A_1| = 3 \cdot 6 = 18, \quad |A_2| = 6 \cdot 3 = 18, \quad |A_3| = 6 \cdot 3 = 18.$$

We have, then, $\mathbf{P}(A_i) = \frac{1}{2}$, $i = 1, 2, 3$, and the relations

$$\mathbf{P}\left(A_i A_j\right) = \frac{1}{4} = \mathbf{P}(A_i)\mathbf{P}\left(A_j\right), \ 1 \leq i, j \leq 3, \ i \neq j,$$

which means events A_1, A_2, and A_3 are pairwise independent. On the other hand,

$$\mathbf{P}(A_1 A_2 A_3) = 0 \neq \frac{1}{8} = \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3);$$

consequently, the mutual independence of events A_1, A_2, and A_3 does not follow from their pairwise independence.

Formula of Total Probability, Bayes' Rule Using the theorem of multiplication for conditional probability we can easily derive the following two theorems. Despite the fact that the two theorems are not complicated, they represent quite effective tools in the course of the various considerations.

Theorem 1.12 (*Formula of total probability*). *Let the sequence* $\{A_i,\ i \in I\}$ *be a complete system of events with* $\boldsymbol{P}(A_i > 0),\ i \in I$*; then for all events* B

$$\boldsymbol{P}(B) = \sum_{i \in I} \boldsymbol{P}(B|A_i)\boldsymbol{P}(A_i)$$

is true.

Theorem 1.13 (*Bayes' rule*). *Under the conditions of the preceding theorem, the following relation holds for all indices* $n \in I$*:*

$$\mathbf{P}(A_n|B) = \frac{\mathbf{P}(B|A_n)\mathbf{P}(A_n)}{\sum\limits_{i \in I} \mathbf{P}(B|A_i)\mathbf{P}(A_i)}.$$

Concept of Random Variables Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space that is to be fixed later on. In the course of random experiments, the experiments usually result in some kind of value. This means that the occurrence of a simple event ω results in a random $X(\omega)$ value. Different values might belong to different simple events; however, the function $X(\omega)$, depending on the simple event ω, will have a specific property. We must answer such basic questions as, for example, what is the probability that the result of the experiment will be smaller than a certain given value x? We have only determined probabilities of events (only for elements of the set $\mathcal{A}$) in connection with the definition of probability space; therefore, it has the immediate consequence that we may only consider the probability of the set if the set $\{\omega : X(\omega) \leq x\}$ is an event, which means that the set belongs to σ-algebra $\mathcal{A}$:

$$\{\omega : X(\omega) \leq x\} \in \mathcal{A}.$$

This fact led to one of the most important notions of probability theory.

Definition 1.14. The real-valued function $X : \Omega \to \mathbb{R}$ is called a **random variable** if the relationship

$$\{\omega : X(\omega) \leq x\} \in \mathcal{A}$$

is valid for all real numbers $x \in \mathbb{R}$. A function satisfying this condition is called $\mathcal{A}$ **measurable**.

A property of random variables should be mentioned here. Define by $\mathcal{B} = \mathcal{B}_1$ the σ-algebra of Borel sets of $\mathbb{R}$ as the minimal σ-algebra containing all intervals of $\mathbb{R}$; the elements of $\mathcal{B}$ are called the Borel sets of $\mathbb{R}$. If X is $\mathcal{A}$ measurable, then for all Borel sets D of $\mathbb{R}$ the set $\{\omega : X(\omega) \in D\}$ is also an element of $\mathcal{A}$, i.e., $\{\omega : X(\omega) \in D\}$ is an event. Thus the probability $\mathbf{P}_X[D] = \mathbf{P}(\{\omega : X(\omega) \in D\})$, and so $\mathbf{P}(\{\omega : X(\omega) \leq x\})$ are well defined. An important special case of random variables are the so-called **indicator variables** defined as follows. Let $A \in \mathcal{A}$ be an event, and let us introduce the random variable $\mathcal{I}_{\{A\}}$, $A \in \mathcal{A}$:

$$\mathcal{I}_{\{A\}} = \mathcal{I}_{\{A\}}(\omega) = \begin{cases} 1, \text{ if } \omega \in A, \\ 0, \text{ if } \omega \notin A. \end{cases}$$

Distribution Function Let $X = X(\omega)$ be a random variable; then the probability $\mathbf{P}(X \leq x)$, $x \in \mathbb{R}$, is well defined.

Definition 1.15. The function $F_X(x) = \mathbf{P}(X \leq x)$ for all real numbers $x \in \mathbb{R}$ is called a **cumulative distribution function**(CDF) of random variable X .

Note that the CDFs F_X and function $\mathbf{P}_X$ determine each other mutually and unambiguously. It is also clear that if the real line $\mathbb{R}$ is chosen as a new sample space, and $\mathcal{B}$ is a σ-algebra of Borel sets as the σ-algebra of events, then the triplet $(\mathbb{R}, \mathcal{B}, \mathbf{P}_X)$ determines a new probability space, where $\mathbf{P}_X$ is referred to as a probability measure induced by the random variable X.

The CDF F_X has the following properties.

(1) In all points of a real line $-\infty < x_0 < \infty$ the function $F_X(x)$ is continuous from the right, that is,

$$\lim_{x \to x_0 + 0} F_X(x) = F_X(x_0).$$

(2) The function $F_X(x)$, $-\infty < x < \infty$ is a monotonically increasing function of the variable x, that is, for all $-\infty < x < y < \infty$ the inequality $F_X(x) \leq F_X(y)$ holds.

(3) The limiting values of the function $F_X(x)$ exist under the conditions $x \to -\infty$ and $x \to \infty$ as follows:

$$\lim_{x \to -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F_X(x) = 1.$$

(4) The set of discontinuity points of the function $F_X(x)$, that is, the set of points $x \in \mathbb{R}$ for which $F_X(x) \neq F_X(x - 0)$, is countable.

Comment 1.16. *It should be noted in connection with the definition of the CDF that the literature is not consistent. The use of $F_X(x) = \boldsymbol{P}(X < x)$, $-\infty < x < \infty$ as a CDF is also widely applied. The only difference between the two definitions lies within property (1) (see preceding discussion), which means that in the latter case the CDF is continuous from the left and not from the right, but all the other properties remain the same. It is also clear that if the CDF is continuous in all $x \in \mathbb{R}$, then there is no difference between the two definitions.*

Comment 1.17. *From a practical point of view, it is sometimes useful to allow that property (3) (see preceding discussion) does not satisfy the CDF F_X of random variable X, which means that, instead, one or both of the following relations hold: In this case $\boldsymbol{P}(|X| < \infty) < 1$, and the CDF of random variable X has a **defective distribution function**.*

Let a and b be two arbitrary real numbers for which $-\infty < a < b < \infty$; then we can determine the probability of some frequently occurring events with the use of the CDF of X as follows:

$$\begin{aligned}
\mathbf{P}(X = a) &= F_X(a) - F_X(a - 0),\\
\mathbf{P}(a < X < b) &= F_X(b - 0) - F_X(a),\\
\mathbf{P}(a \le X < b) &= F_X(b - 0) - F_X(a - 0),\\
\mathbf{P}(a < X \le b) &= F_X(b) - F_X(a),\\
\mathbf{P}(a \le X \le b) &= F_X(b) - F_X(a - 0).
\end{aligned}$$

These equations also determine the connection between the CDF F_X and the distribution $\mathbf{P}_X$ for special Borel sets of a real line.

Discrete and Continuous Distribution, Density Function We distinguish two important types of distributions in practice, the so-called discrete and continuous distributions. There is also a third type of distribution, the so-called singular distribution, in which case the CDF is continuous everywhere and its derivative (with respect to the Lebesgue measure) equals 0 almost everywhere; however, we will not consider this type. This classification follows from the Jordan decomposition theorem of monotonically functions, that is, an arbitrary CDF F can always be decomposed into the sum of three functions – the monotonically increasing absolutely continuous function, the step function with finite or countably infinite sets of jumps (this part corresponds to a discrete distribution), and the singular function.

Definition 1.18. Random variable X is **discrete** or has a **discrete distribution** if there is a finite or countably infinite set of values $\{x_k,\ k \in I\}$ such that $\sum\limits_{k\in I} p_k = 1$, where $p_k = \mathbf{P}(X = x_k),\ k \in I$. The associated function

$$f_X(x) = \begin{cases} p_k, & \text{if } x = x_k,\ k \in I,\\ 0, & \text{if } x \ne x_k,\ k \in I, \end{cases} \qquad x \in \mathbb{R},$$

is termed a **probability density function** (PDF) or **probability mass function** (PMF).

It is easy to see that if random variable X is discrete with possible values $\{x_k, k = 0, 1, \ldots\}$ and with distribution $\{p_k, k = 0, 1, \ldots\}$, then the relationship between the CDF F_X and the PMF can be given as

$$F_X(x) = \sum_{x_k < x} p_k,\ -\infty < x < \infty.$$

Definition 1.19. A random variable X is **continuous** or has a **continuous distribution** if there exists a nonnegative integrable function $f_X(x)$, $-\infty < x < \infty$ such that for all real numbers a and b, $-\infty < a < b < \infty$,

$$F_X(b) - F_X(a) = \int_a^b f_X(x)\mathrm{d}x$$

holds. The function $f_X(x)$ is called the PDF of random variable X, or just the **density function** of X.

Comment 1.20. *It is clear that*

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u, \ -\infty < x < \infty,$$

and it is also true that the PDF is not uniquely defined since if we take instead of $f_X(u)$ *the function* $f_X(u)+g(u)$*, where the function* $g(u)$ *is nonnegative, integrable, and* $\int_{-\infty}^{x} g(u)\mathrm{d}u = 0$*, then the function* $f_X(u) + g(u)$ *is also a PDF of random variable* X*, which can naturally differ from the original* f_X*.*

An arbitrary PDF $f_X(x)$ is nonnegative and integrable,

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

and almost everywhere in $\mathbb{R}$ (with respect to the Lebesgue measure) the equation $F'_X(x) = f_X(x)$ is true.

Distribution of a Function of a Random Variable Let $X = X(\omega)$ be a random variable. Let $h(x)$, $x \in \mathbb{R}$ be a real-valued function, and let us define it as $Y = h(X)$. The equation $Y = h(X)$ determines a random variable if for all $y \in \mathbb{R}$ the set $\{\omega : Y(\omega) = h(X(\omega)) \le y\}$ is an event that is an element of σ-algebra $\mathcal{A}$. If h is a continuous function or, more generally, is a Borel-measurable function (h is Borel measurable if for all x the relationship $\{u : h(u) \le x\} \in \mathcal{B}$ is true), then Y, which is determined by the equation $Y = h(X)$, is a random variable. The question is how the CDF and the density function (if the latter exists) of random variable Y can be determined. It is usually true that

$$F_X(y) = \mathbf{P}(Y \le y) = \mathbf{P}(h(X) \le y) = \mathbf{P}_X[\{x : h(x) \le y\}], \ -\infty < y < \infty.$$

If h is a strictly monotonically increasing function, then this formula can be given in a simpler form. Let us denote by h^{-1} the inverse function of h, which in this case must exist. Then

$$F_X(y) = \mathbf{P}(h(X) \le y) = \mathbf{P}\left(X \le h^{-1}(y)\right) = F_X(h^{-1}(y)),\ -\infty < y < \infty.$$

If h is a strictly monotonically decreasing function, then

$$F_X(y) = \mathbf{P}(h(X) \le y) = \mathbf{P}\left(X \ge h^{-1}(y)\right) = 1 - F_X(h^{-1}(y) - 0),\ -\infty < y < \infty.$$

With these relations, a formula can be given for the PDF of Y in special cases.

Theorem 1.21. *Let us suppose that random variable X has a PDF f_X and h is a strictly monotonically, differentiable real function. Then*

$$f_Y(y) = f_X(h^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} h^{-1}(y) \right|,\ -\infty < y < \infty.$$

Comment 1.22. *If h is a linear function, that is, $h(y) = ay + b$, $a \neq 0$, and X has a PDF f_X, then the random variable $Y = h(X)$ also has a PDF and the formula*

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-a}{b}\right),\ -\infty < y < \infty,$$

is true.

Joint Distribution and Density Function of Random Variables, Marginal Distributions In the majority of problems arising in practice, we have not one but several random variables, and we examine the probability of events where random variables simultaneously satisfy certain conditions.

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, and let there be two random variables X and Y on that space. The joint statistical behavior of the two random variables can be determined by a **joint CDF**. We should note that the joint analysis of the random variables X and Y corresponds to the examination of two-dimensional random vector variables such as (X, Y) that have random variable coordinates.

Definition 1.23. The function

$$F_{XY}(x, y) = \mathbf{P}(X \le x, Y \le y),\ -\infty < x, y < \infty,$$

is called the **joint CDF** of random variables X and Y.

From a practical point of view, the two most important types of distributions are the discrete and the continuous ones, as in the one-dimensional case.

Definition 1.24. The joint distribution function of random variables X and Y is called **discrete**; in other words, the random vector (X, Y) has a **discrete distribution** if random variables X and Y are discrete. If we denote the values

of random variables X and Y by $\{x_i,\ i \in I\}$ and $\{y_j,\ j \in J\}$, respectively, then the function

$$f_{X,Y}(x,y) = \begin{cases} p_{i,j}, & \text{if } x = x_i,\ y = y_j,\ i \in I,\ j \in J, \\ 0, & \text{if } x \neq x_i,\ y \neq y_j,\ i \in I,\ j \in J, \end{cases} \quad x \in \mathbb{R},$$

is called a **joint PMF** or **joint PDF**.

It is clear that in the discrete case the joint distribution function is

$$F_{XY}(x,y) = \sum_{x_i \leq x,\ y_j \leq y} p_{ij}.$$

The case of a joint continuous distribution is analogous to the discrete one.

Definition 1.25. The joint distribution of random variables X and Y is called **continuous**; in other words, the random vector (X,Y) has a **continuous distribution** if there exists a nonnegative, real-valued integrable function on the plane $f_{XY}(x,y)$, $-\infty < x, y < \infty$, for which the relation

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u,v)\mathrm{d}u\mathrm{d}v$$

holds for all $-\infty < x, y < \infty$.

Definition 1.26. If F_{XY} denotes the joint CDF of random variables X and Y, then the CDFs

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x,y),$$
$$F_Y(y) = \lim_{x \to \infty} F_{XY}(x,y)$$

are called **marginal distribution functions**.

It is not difficult to see that marginal distribution functions do not determine the joint CDF. It is also clear that if a joint PDF $f_{XY}(x,y)$ of random variables X and Y exists, then marginal PDFs can be given in the form

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}y,\ -\infty < x < \infty,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x\ -\infty < y < \infty.$$

If there are more than two random variables $X_1, \ldots, X_n$, $n \geq 3$, i.e., in the case of an n-dimensional random vector $(X_1, \ldots, X_n)$, then the definitions of joint

distribution function and density functions can be given analogously to the case of two random variables, so there is no essential difference. We will return to this question when we introduce the concept of stochastic processes.

Conditional Distributions Let A be an arbitrary event, with $P(A) > 0$, and X an arbitrary random variable. Using the notion of conditional probability, we can define the **conditional distribution** of random variable X given event A as the function

$$F_X(x|A) = \mathbf{P}(X \leq x|A),\ x \in \mathbb{R}.$$

The function $F_X(x|A)$ has all the properties of a distribution function mentioned previously.

The function $f_X(x|A_i)$ is called a **conditional density function** of random variable X given event A if a nonnegative integrable function $f_X(x|A)$ exists for which the equation

$$F_X(x|A) = \int_{-\infty}^{x} f_X(u|A)\mathrm{d}u,\ -\infty < x < \infty,$$

holds.

The result for the distribution function $F_X(x)$ can be easily proved in the same way as the theorem of full events. If the sequence of events $A_1, A_2, \ldots$ is a complete system of events with the property $\mathbf{P}(A_i) > 0,\ i = 1, 2, \ldots$, then

$$F_X(x) = \sum_{i=1}^{\infty} F_X(x|A_i)\mathbf{P}(A_i),\ -\infty < x < \infty.$$

A similar relation holds for the conditional PDFs $f_X(x|A_i),\ i \geq 1$, if they exist:

$$f_X(x) = \sum_{i=1}^{\infty} f_X(x|A_i)\mathbf{P}(A_i),\ -\infty < x < \infty.$$

A different approach is required to define the conditional distribution function $F_{X|Y}(x|y)$ of random variable X given $Y = y$, where Y is another random variable. The difficulty is that if a random variable Y has a continuous distribution function, then the probability of the event $\{Y = y\}$ equals zero, and therefore the conditional distribution function $F_{X|Y}(x|y)$ cannot be defined with the help of the notion of conditional probability. In this case the conditional distribution function $F_{X|Y}(x|y)$ is defined as follows:

$$F_{X|Y}(x|y) = \lim_{\Delta y \to +0} \mathbf{P}(X \leq x|y \leq Y < y + \Delta y)$$

if the limit exists.

Let us assume that the joint density function $f_{XY}(x, y)$ of random variables X and Y exists. In such a case random variable X has the conditional CDF $F_{X|Y}(x|y)$ and conditional PDF $f_{X|Y}(x|y)$ given $Y = y$. If a joint PDF exists and $f_X(y) > 0$, then it is not difficult to see that the following relation holds:

$$\begin{aligned} F_{X|Y}(x|y) &= \lim_{\Delta y \to +0} \mathbf{P}(X \le x | y \le Y < y + \Delta y) \\ &= \lim_{\Delta y \to +0} \frac{\mathbf{P}(X \le x, y \le Y < y + \Delta y)}{\mathbf{P}(y \le Y < y + \Delta y)} \\ &= \lim_{\Delta y \to +0} \frac{\frac{F_{XY}(x, y+\Delta y) - F_{XY}(x,y)}{\Delta y}}{\frac{F_Y(y+\Delta y) - F_Y(y)}{\Delta y}} = \frac{1}{f_Y(y)} \frac{\partial}{\partial y} F_{XY}(x, y). \end{aligned}$$

From this relation we get the conditional PDF $f_{X|Y}(x|y)$ as follows:

$$f_{X|Y}(x|y) = \frac{\partial}{\partial x} F_{X|Y}(x|y) = \frac{1}{f_Y(y)} \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) = \frac{f_{XY}(x, y)}{f_Y(y)}. \tag{1.1}$$

Independence of Random Variables Let X and Y be two random variables. Let $F_{XY}(x, y)$ be the joint distribution function of X and Y, and let $F_X(x)$ and $F_Y(y)$ be the marginal distribution functions.

Definition 1.27. Random variables X and Y are called independent of each other, or just independent, if the identity

$$F_{XY}(x, y) = F_X(x) F_Y(y)$$

holds for any $x, y, -\infty < x, y < \infty$.

In other words, random variables X and Y are independent if and only if the joint distribution function of X and Y equals the product of their marginal distribution functions.

The definition of independence of two random variables can be easily generalized to the case where an arbitrary collection of random variables $\{X_i,\ i \in I\}$ is given, analogously to the notion of the independence of events.

Definition 1.28. A collection of random variables $\{X_i,\ i \in I\}$ is called **mutually independent** (or just **independent**), if for any choice of a finite number of elements $X_{i_1}, \ldots, X_{i_n}$ the relation

$$F_{X_{i_1}, \ldots, X_{i_n}}(x_1, \ldots, x_n) = F_{X_{i_1}}(x_1) \cdot \ldots \cdot F_{X_{i_n}}(x_n),\ x_1, \ldots, x_n \in \mathbb{R}$$

holds.

Note that from the **pairwise independence** of random variables $\{X_i,\ i \in I\}$, which means that the condition

$$F_{X_{i_1},X_{i_2}}(x_1, x_2) = F_{X_{i_1}}(x_1)F_{X_{i_2}}(x_2),\ x_1, x_2 \in \mathbb{R},\ i_1, i_2 \in I,$$

is satisfied, mutual independence does not follow.

Example 1.29. Consider Example 1.11 given earlier and preserve the notation. Denote by $X_i = \mathcal{I}_{\{A_i\}}$ the indicator variables of the events $A_i,\ i = 1, 2, 3$. Then we can verify that random variables X_1, X_2, and X_3 are pairwise independent, but they do not satisfy mutual independence. The pairwise independence of random variables X_i can be easily proved. Since the events A_1, A_2, A_3 are independent and

$$\{X_i = 1\} = A_i \quad \text{and} \quad \{X_i = 0\} = \overline{A}_i,$$

then, using the relation proved in Example 1.11, we obtain for $i \neq j$

$$\begin{aligned}
\mathbf{P}\left(X_i = 1, X_j = 1\right) &= \mathbf{P}\left(A_i A_j\right) = \mathbf{P}\left(A_i\right)\mathbf{P}\left(A_j\right) = \frac{1}{4},\\
\mathbf{P}\left(X_i = 1, X_j = 0\right) &= \mathbf{P}\left(A_i \overline{A}_j\right) = \mathbf{P}\left(A_i\right)\mathbf{P}\left(\overline{A}_j\right) = \frac{1}{4},\\
\mathbf{P}\left(X_i = 0, X_j = 0\right) &= \mathbf{P}\left(\overline{A}_i \overline{A}_j\right) = \mathbf{P}\left(\overline{A}_i\right)\mathbf{P}\left(\overline{A}_j\right) = \frac{1}{4},
\end{aligned}$$

while, for example,

$$\begin{aligned}
&\mathbf{P}\left(X_1 = 1, X_2 = 1, X_3 = 1\right) = \mathbf{P}\left(A_1 A_2 A_3\right) = 0 \neq \frac{1}{8}\\
&\quad = \mathbf{P}\left(A_1\right)\mathbf{P}\left(A_2\right)\mathbf{P}\left(A_3\right) = \mathbf{P}\left(X_1 = 1\right)\mathbf{P}\left(X_2 = 1\right)\mathbf{P}\left(X_3 = 1\right).
\end{aligned}$$

Consider how we can characterize the notion of independence for two random variables in the discrete and continuous cases (if more than two random variables are given, then we may proceed in a similar manner).

Firstly, let us assume that the sets of values of discrete random variables X and Y are $\{x_i, i \geq 0\}$ and $\{y_j, j \geq 0\}$, respectively. If we denote the joint and marginal distributions of X and Y by

$$\begin{aligned}
&\left\{p_{ij} = \mathbf{P}\left(X = x_i, Y = y_j\right),\ i, j \geq 0\right\}, \left\{q_i = \mathbf{P}\left(X = x_i\right),\ i \geq 0\right\},\\
&\quad \text{and } \left\{r_j = \mathbf{P}\left(Y = y_j\right), j \geq 0\right\},
\end{aligned}$$

then the following assertion holds. Random variables X and Y are independent if and only if

$$p_{ij} = q_i r_j,\ i, j \geq 0.$$

Now assume that random variables X and Y have joint density $f_{XY}(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$. Thus, in this case, random variables X and Y are independent if and only if the joint PDF takes a product form, that is,

$$f_{XY}(x, y) = f_X(x) f_Y(y), \ -\infty < x, y < \infty.$$

Convolution of Distributions Let X and Y be independent random variables with distribution functions $F_X(x)$ and $F_Y(y)$, respectively, and let us consider the distribution of the random variable $Z = X + Y$.

Definition 1.30. The distribution (CDF, PDF) of the random variable $Z = X+Y$ is called the **convolution** of the distribution (CDF, PDF), and the equations expressing the relation among them are called convolution formulas.

Definition 1.31. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with the common CDF F_X. The CDF F_X^{*n} of the sum $Z_n = X_1 + \ldots + X_n$ $(n \geq 1)$ is uniquely determined by F_X and is called the n**-fold convolution** of the CDF of F_X.

Note that the CDF $F_Z(z)$ of the random variable $Z = X + Y$, which is called the **convolution** of CDFs $F_X(x)$ and $F_Y(y)$, can be given in the general form

$$F_Z(z) = \mathbf{P}(Z \leq z) = \mathbf{P}(X + Y \leq z) = \int_{-\infty}^{\infty} F_X(z - y) \, \mathrm{d}F_Y(y).$$

This formula gets a simpler form in cases where the discrete random variables X and Y take only integer numbers, or if the PDFs $f_X(x)$ and $f_Y(y)$ of X and Y exist.

Let X and Y be independent discrete random variables taking values in $\{0, \pm 1, \pm 2, \ldots\}$ with probabilities $\{q_i = \mathbf{P}(X = x_i)\}$ and $\{r_j = \mathbf{P}(Y = y_j)\}$, respectively. Then the random variable $Z = X + Y$ takes values in $\{0, \pm 1, \pm 2, \ldots\}$, and its distribution satisfies the identity

$$s_k = \sum_{n=-\infty}^{\infty} q_{k-n} r_n, \ k = 0, \pm 1, \pm 2, \ldots .$$

If the independent random variables X and Y have a continuous distribution with the PDFs $f_X(x)$ and $f_Y(y)$, respectively, then random variable Z is continuous and its PDF $f_Z(z)$ can be given in the integral form

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) \mathrm{d}y.$$

Mixture of Distributions Let $F_1(x), \ldots, F_n(x)$ be a given collection of CDFs, and let $a_1, \ldots, a_n$ be nonnegative numbers with the sum $a_1 + \ldots + a_n = 1$. The function

$$F(x) = a_1 F_1(x) + \ldots + a_n F_n(x),\ -\infty < x < \infty,$$

is called a **mixture** of CDFs $F_1(x), \ldots, F_n(x)$ with weights $a_1, ..., a_n$.

Comment 1.32. *Any CDF can be given as a mixture of discrete, continuous, and singular CDFs, where the weights can also take a value of* 0.

Clearly, the function $F(x)$ possesses all the properties of CDFs; therefore it is also a CDF. In practice, the modeling of mixture distributions plays a basic role in stochastic simulation methods. A simple way to model mixture distributions is as follows.

Let us assume that the random variables $X_1, \ldots, X_n$ with distribution functions $F_1(x), \ldots, F_n(x)$ can be modeled. Let Y be a random variable taking values in $\{1, \ldots, n\}$ and independent of $X_1, \ldots, X_n$. Assume that Y has a distribution $P(Y = i) = a_i,\ 1 \le i \le n\ (a_i \ge 0,\ a_1 + \ldots + a_n = 1)$. Let us define random variable Z as follows:

$$Z = \sum_{i=1}^{n} \mathcal{I}_{\{Y=i\}} X_i,$$

where $\mathcal{I}_{\{\}}$ denotes the indicator variable. Then the CDF of random variable Z equals $F(z)$.

Proof. Using the formula of total probability, we have the relation

$$\mathbf{P}(Z \le z) = \sum_{i=1}^{n} \mathbf{P}(Z \le z | Y = i)\mathbf{P}(Y = i) = \sum_{i=1}^{n} \mathbf{P}(X_i \le z)a_i = F(z).$$

□

Concept and Properties of Expectation A random variable can be completely characterized in a statistical sense by its CDF. To define a distribution function $F(x)$, one needs to determine its values for all $x \in \mathbb{R}$, but this is not possible in many cases. Fortunately, there is no need to do so because in many cases it suffices to give some values that characterize the CDF in a certain sense depending on concrete practical considerations. One of the most important concepts is expectation, which we define in general form, and we give the definition for discrete and continuous distributions as special cases.

Definition 1.33. Let X be a random variable, and let $F_X(x)$ be its CDF. The **expected value** (or **mean value**) of random variable X is defined as

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x \mathrm{d} F_X(x)$$

if the expectation exists.

Note that the finite expected value $\mathbf{E}(X)$ exists if and only if $\int_{-\infty}^{\infty} |x| \mathrm{d}F_X(x) < \infty$. It is conventional to denote the expected value of the random variable X by μ_X.

Expected Value of Discrete and Continuous Random Variables Let X be a discrete valued random variable with countable values $\{x_i, i \in I\}$ and with probabilities $\{p_i = \mathbf{P}(X = x_i), i \in I\}$. The finite expected value $\mathbf{E}(X)$ of random variable X exists and equals

$$\mathbf{E}(X) = \sum_{i \in I} p_i x_i$$

if and only if the sum is absolutely convergent, that is, $\sum_{i \in I} p_i |x_i| < \infty$. In the case of continuous random variables, the expected value can also be given in a simple form. Let $f_X(x)$ be the PDF of a random variable X. If the condition $\int_{-\infty}^{\infty} |x| f_X(x) \mathrm{d}x < \infty$ holds (i.e., the integral is absolutely convergent), then the finite expected value of X exists and can be given as

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x.$$

From a practical point of view, it is generally enough to give two special, discrete, and continuous cases. Let X be a random variable that has a mixed CDF with discrete and continuous components $F_1(x)$ and $F_2(x)$, respectively, and with weights a_1 and a_2, that is,

$$F(x) = a_1 F_1(x) + a_2 F_2(x),\ a_1, a_2 \geq 0,\ a_1 + a_2 = 1.$$

Assume that the set of discontinuities of $F_1(x)$ is $\{x_i,\ i \in I\}$ and denote $p_i = F_1(x_i) - F_1(x_i -), i \in I$. In addition, we assume that the continuous CDF $F_2(x)$ has the PDF $f(x)$. Then the expected value of random variable X is determined as follows:

$$\mathbf{E}(X) = a_1 \sum_{i \in I} p_i x_i + a_2 \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

if the series and the integral on the right-hand side of the last formula are absolutely convergent. The expected values related to special and different CDFs will be given later in this chapter.

The operation of expectation can be interpreted as a functional

$$\mathbf{E} : X \to \mathbf{E}(X)$$

that assigns a real value to the given random variable. We enumerate the basic properties of this functional as follows.

1. If random variable X is finite, i.e., if there are constants x_1 and x_2 for which the inequality $x_1 \le X \le x_2$ holds, then

$$x_1 \le \mathbf{E}(X) \le x_2.$$

If random variable X is nonnegative and the expected value $\mathbf{E}(X)$ exists, then

$$\mathbf{E}(X) \ge 0.$$

2. Let us assume that the expected value $\mathbf{E}(X)$ exists; then the expected value of random variable cX exists for an arbitrary given constant c, and the identity

$$\mathbf{E}(cX) = c\mathbf{E}(X)$$

is true.
3. If random variable X satisfies the condition $\mathbf{P}(X = c) = 1$, then

$$\mathbf{E}(X) = c.$$

4. If the expected values of random variables X and Y exist, then the sum $X + Y$ has an expected value, and the equality

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y)$$

holds. This relation can usually be interpreted in such a way that the operation of expectation on the space of random variables is an additive functional.
5. The preceding properties can be expressed in a more general form. If there are finite expected values of random variables $X_1, \ldots, X_n$ and $c_1, \ldots, c_n$ are constants, then the equality

$$\mathbf{E}(c_1X_1 + \ldots + c_nX_n) = c_1\mathbf{E}(X_1) + \ldots + c_n\mathbf{E}(X_n)$$

holds. This property means that the functional $\mathbf{E}()$ is a linear one.
6. Let X and Y be independent random variables with finite expected value. Then the expected value of the product of random variables $X \cdot Y$ exists and equals the product of expected values, i.e., the equality

$$\mathbf{E}(XY) = \mathbf{E}(X) \cdot \mathbf{E}(Y)$$

is true.

Expectation of Functions of Random Variables, Moments and Properties Let X be a discrete random variable with finite or countable values $\{x_i, i \in I\}$ and with distribution $\{p_i, i \in I\}$. Let $h(x),\ x \in \mathbb{R}$ be a real-valued function for which the expected value of the random variable $Y = h(X)$ exists; then the equality

$$\mathbf{E}(Y) = \mathbf{E}(h(X)) = \sum_{i \in I} p_i h(x_i)$$

holds.

If the continuous random variable X has a PDF $f_X(x)$ and the expected value of the random variable $Y = h(X)$ exists, then the expected value of Y can be given in the form

$$\mathbf{E}(Y) = \int_{-\infty}^{\infty} h(x) f_X(x) \mathrm{d}x.$$

In cases where the expected value of functions of random variables (functions of random vectors) are investigated, analogous results to the one-dimensional case can be obtained. We give the formulas in connection with the two-dimensional case only. Let X and Y be two random variables, and let us assume that the expected value of the random variable $Z = h(X, Y)$ exists. With the appropriate notation, used earlier, for the cases of discrete and continuous distributions, the expected value of random variable Z can be given in the forms

$$\mathbf{E}(Z) = \sum_{i \in I} \sum_{j \in J} h(x_i, y_j) \mathbf{P}\left(X = x_i, Y = y_j\right),$$

$$\mathbf{E}(Z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f_{XY}(x, y) \mathrm{d}x \mathrm{d}y.$$

Consider the important case where h is a power function, i.e., for a given positive integer number k, $h(x) = x^k$. Assume that the expected value of X^k exists. Then the quantity

$$\mu_k = \mathbf{E}\left(X^k\right),\ k = 1, 2, \ldots,$$

is called the kth moment of random variable X. It stands to reason that the **first moment** $\mu = \mu_1 = \mathbf{E}\left(X^1\right)$ is the expected value of X and the frequently used **second moment** is $\mu_2 = \mathbf{E}\left(X^2\right)$.

Theorem 1.34. *Let j and k be integer numbers for which $1 \leq j \leq k$. If the kth moment of random variable X exists, then the jth moment also exists.*

Proof. From the existence of the kth moment it follows that $\mathbf{E}\left(|X|^k\right) < \infty$. Since $k/j \geq 1$, the function $x^{k/j},\ x \geq 0$, is convex, and by the use of Jensen's inequality we get the relation

$$\left[\mathbf{E}\left(|X|^j\right)\right]^{k/j} \leq \mathbf{E}\left(\left(|X|^j\right)^{k/j}\right) = \mathbf{E}\left(|X|^k\right) < \infty.$$

□

The kth central moment $\mathbf{E}\left((X-\mathbf{E}(X))^k\right)$ is also used in practice; it is defined as the kth moment of the random variable centered at the first moment (expected value). The kth central moment $\mathbf{E}\left((X-\mathbf{E}(X))^k\right)$ can be expressed by the noncentral moments $\mu_i, 1 \leq i \leq k$ of random variable X as follows:

$$\mathbf{E}\left((X-\mathbf{E}(X))^k\right) = \mathbf{E}\left(\sum_{i=0}^{k}\binom{k}{i}X^i(-\mathbf{E}(X))^{k-i}\right)$$
$$= \sum_{i=0}^{k}\binom{k}{i}\mathbf{E}\left(X^i\right)(-\mathbf{E}(X))^{k-i}.$$

In the course of a random experiment, the observed values fluctuate around the expected value. One of the most significant characteristics of the quantity of fluctuations is the variance. Assume that the second moment of random variable X is finite. Then the quantities

$$\mathbf{Var}(X) = \mathbf{E}\left((X-\mathbf{E}(X))^2\right)$$

are called the **variance** of random variable X. The **standard deviation** of a random variable X is the square root of its variance:

$$\mathbf{D}(X) = \sqrt{\mathbf{E}((X-\mathbf{E}(X))^2)}.$$

It is clear that the variance of X can be given with the help of the first and second moments as follows:

$$\mathbf{D}^2(X) = \mathbf{Var}(X) = \mathbf{E}\left((X-\mathbf{E}(X))^2\right) = \mathbf{E}\left(X^2\right) - 2\mathbf{E}(X)\cdot\mathbf{E}(X) + (\mathbf{E}(X)^2)$$
$$= \mathbf{E}\left(X^2\right) - (\mathbf{E}(X))^2 = \mu_2 - \mu^2.$$

It is conventional to denote the variance of the random variable X by $\sigma_X^2 = \mathbf{D}^2(X)$.

It should be noted that the variance of a random variable exists if and only if its second moment is finite. In addition, from the last inequality it follows that an upper estimation can be given for the variance as

$$\mathbf{D}^2(X) \leq \mathbf{E}\left(X^2\right).$$

It can also be seen that for every constant c the relation

$$\mathbf{E}\left((X-c)^2\right) = \mathbf{E}\left([(X-\mathbf{E}(X)) + (\mathbf{E}(X)-c)]^2\right) = \mathbf{D}^2(X) + (\mathbf{E}(X)-c)^2$$

holds, which is analogous to the Steiner formula, well known in the field of mechanics.

As an important consequence of this identity, we have the following result: the second moment $\mathbf{E}\left((X-c)^2\right)$ takes the minimal value for the constant $c = \mathbf{E}(X)$.

We will now mention some frequently used properties of variance.

1. If the variance of random variable X exists, then for all constants a and b the identity
$$\mathbf{D}^2(aX+b) = a^2\mathbf{D}^2(X)$$
is true.
2. Let $X_1, \ldots, X_n$ be independent random variables with finite variance; then
$$\mathbf{D}^2(X_1 + \ldots + X_n) = \mathbf{D}^2(X_1) + \ldots + \mathbf{D}^2(X_n). \tag{1.2}$$

The independence of random variables that play a role in formula (1.2) is not required for the last identity, and it is also true if instead of assuming the independence of the random variables $X_1, \ldots, X_n$ we assume that they are uncorrelated. The notion of correlation is to be defined later. If $X_1, \ldots, X_n$ are independent and identically distributed random variables with finite variance σ, then

$$\mathbf{D}^2(X_1 + \ldots + X_n) = \mathbf{D}^2(X_1) + \ldots + \mathbf{D}^2(X_n) = n\sigma^2,$$

from which
$$\mathbf{D}(X_1 + \ldots + X_n) = \sigma\sqrt{n}$$
follows.

In the literature on queueing theory, the notion of **relative variance** $\mathbf{CV}(X)^2$ is applied, which is defined as

$$\mathbf{CV}(X)^2 = \frac{\mathbf{D}^2(X)}{\mathbf{E}(|X|)^2} \ .$$

Its square root $\mathbf{CV}(X) = \mathbf{D}(X)/\mathbf{E}(|X|)$ is called the **coefficient of variation**, which serves as a normalized measure of variance of a distribution. The following inequalities hold:

Exponential distribution:	$CV = 1,$
Hyperexponential distribution:	$CV > 1,$
Erlang distribution:	$CV < 1.$

Markov and Chebyshev Inequalities The role of the Markov and Chebyshev inequalities is significant, not only because they provide information concerning distributions with the help of expected value and variance but because they are also effective tools for proving certain results.

Theorem 1.35 (*Markov inequality*). *If the expected value of a nonnegative random variable X exists, then the following inequality is true for any constant $\varepsilon > 0$,:*

$$\boldsymbol{P}(X \geq \varepsilon) \leq \frac{\boldsymbol{E}(X)}{\varepsilon}.$$

Proof. For an arbitrary positive constant $\varepsilon > 0$ we have the relation

$$\mathbf{E}(X) \geq \mathbf{E}\left(X\mathcal{I}_{\{X \geq \varepsilon\}}\right) \geq \varepsilon \mathbf{E}\left(\mathcal{I}_{\{X \geq \varepsilon\}}\right) = \varepsilon\, \mathbf{P}(X \geq \varepsilon),$$

from which the Markov inequality immediately follows. □

Theorem 1.36 (*Chebyshev inequality*). *If the variance of random variable X is finite, then for any constant $\varepsilon > 0$ the inequality*

$$\boldsymbol{P}(|X - \boldsymbol{E}(X)| \geq \varepsilon) \leq \frac{\boldsymbol{D}^2(X)}{\varepsilon^2}$$

holds.

Proof. Using the Markov inequality for a constant $\varepsilon > 0$ and for the random variable $(X - \mathbf{E}(X))^2$ we find that

$$\mathbf{P}(|X - \mathbf{E}(X)| \geq \varepsilon) = \mathbf{P}\left((X - \mathbf{E}(X))^2 \geq \varepsilon^2\right) \leq \frac{\mathbf{E}(X - \mathbf{E}(X))^2}{\varepsilon^2} = \frac{\mathbf{D}^2(X)}{\varepsilon^2},$$

from which the assertion of the theorem follows. □

Comment 1.37. *Let X be a random variable. If $h(x)$ is a convex function and $\boldsymbol{E}(h(X))$ exists, then the Jensen inequality $\boldsymbol{E}(h(X)) \geq h(\boldsymbol{E}(X))$ is true. Using this inequality we can obtain some other relations, similar to the case of the Markov inequality.*

Example 1.38. As a simple application of the Chebyshev inequality, let us consider the average $(X_1 + \ldots + X_n)/n$, where the random variables $X_1, \ldots, X_n$ are independent identically distributed with finite second moment. Let us denote the joint expected value and variance by μ and σ^2, respectively. Using the property (1.2) of variance and the Chebyshev inequality and applying $(n\varepsilon)$ instead of ε, we get the inequality

$$\begin{aligned}\mathbf{P}(|X_1 + \ldots + X_n - n\mu| \geq n\varepsilon) &= \mathbf{P}\left((X_1 + \ldots + X_n - n\mu)^2 \geq n^2\varepsilon^2\right) \\ &\leq \frac{n\sigma^2}{(n\varepsilon)^2} = \frac{\sigma^2}{n\varepsilon^2};\end{aligned}$$

then

$$\mathbf{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

As a consequence of the last inequality, for every fixed positive constant ε the probability $\mathbf{P}\left(\left|\frac{X_1+\ldots+X_n}{n}-\mu\right|\geq\varepsilon\right)$ tends to 0 as n goes to infinity. This assertion is known as the **weak law of large numbers**.

Generating and Characteristic Functions So far, certain quantities characterizing the distribution of random variables have been provided. Now such transformations of distributions will be given where the distributions and the functions obtained by the transformations uniquely determine each other. The investigated transformations provide effective tools for determining, for instance, distributions and moments and for proving limit theorems.

Definition 1.39. Let X be a random variable taking values in $\{0,1,\ldots\}$, with probabilities $p_0, p_1, \ldots$. Then the power series

$$G_X(z)=\mathbf{E}\left(z^X\right)=\sum_{i=0}^{\infty}p_i z^i$$

is convergent for all $z\in[-1,1]$, and the function $G_X(z)$ is called the **probability generating function** (or just **generating function**) of the discrete random variable X.

In engineering practice, the power series defining the generating function is applied in a more general approach instead of in the interval $[-1,1]$, and the generating function is defined on the closed complex unit circle $z\in\mathbb{C}$, $|z|\leq 1$, which is usually called a z**-transform** of the distribution $\{p_i, i=0,1,\ldots\}$. This notion is also applied if, instead of a distribution, a transformation is made to an arbitrary sequence of real numbers.

It should be noted that $|G_X(z)|\leq 1$ if $z\in\mathbb{C}$ and the function $G_X(z)$ is differentiable on the open unit circle of the complex plane $z\in\mathbb{C}$, $|z|<1$ infinitely many times and the kth derivative of $G_X(z)$ equals the sum of the kth derivative of the members of the series.

It is clear that

$$p_k=G_X^{(k)}(0)/k!,\ k=0,1,\ldots\ .$$

This formula makes it possible to compute the distribution if the generating function is given. It is also true that if the first and second derivatives $G'_X(1-)$ and $G''_X(1-)$ exist on the left-hand side at $z=1$, then the first and second moments of random variable X can be computed as follows:

$$\mathbf{E}(X)=G'_X(1-)\quad\text{and}\quad\mathbf{E}\left(X^2\right)=\left(zG'_X(z)\right)'\Big|_{z=1}=G''_X(1-)+G'_X(1-).$$

From this we can obtain the variance of X as follows:

$$\mathbf{D}^2(X)=G''_X(1-)+G'_X(1-)-(G'_X(1-))^2.$$

It can also be verified that if the nth derivative of the generating function $G_X(z)$ exists on the left-hand side at $z = 1$, then

$$\begin{aligned}\mathbf{E}\left(X(X-1)\dots(X-m+1)\right) &= \sum_{k=m}^{\infty} k(k-1)\dots(k-m+1)\,p_k \\ &= G_X^{(m)}(1-),\ 1 \le m \le n.\end{aligned}$$

Computing the expected values on the left-hand side of these identities, we can obtain linear equations between the moments $\mu_k = \mathbf{E}(X^k)$, $1 \le k \le m$, and the derivatives $G_X^{(m)}(1-)$ for all $1 \le m \le n$. The moments μ_m, $m = 1, 2, \dots, n$ can be determined in succession with the help of the derivatives $G_X^{(k)}(1-)$, $1 \le k \le m$. The special cases of $k = 1, 2$ give the preceding formulas for the first and second moments.

Characteristic Function

Definition 1.40. The complex valued function

$$\varphi_X(s) = \mathbf{E}\left(\mathrm{e}^{isX}\right) = \mathbf{E}\left(\cos(sX)\right) + i\,\mathbf{E}\left(\sin(sX)\right),\ s \in \mathbb{R},$$

is called the **characteristic function** of random variable X, where $i = \sqrt{-1}$.

Note that a characteristic function can be rewritten in the form

$$\varphi_X(s) = \int_{-\infty}^{\infty} \mathrm{e}^{isx}\,\mathrm{d}F_X(x),$$

which is the well-known Fourier–Stieltjes transform of the CDF $F_X(x)$.
Using conventional notation, in discrete and continuous cases we have

$$\varphi_X(s) = \sum_{k=0}^{\infty} p_k \mathrm{e}^{isx_k}, \quad \text{and} \quad \varphi_X(s) = \int_{-\infty}^{\infty} \mathrm{e}^{isx} f_X(x)\mathrm{d}x.$$

The characteristic function and the CDFs determine each other uniquely. Now some important properties of characteristic functions will be enumerated.

1. The characteristic function is real valued if and only if the distribution is symmetric.
2. If the kth moment $\mathbf{E}\left(X^k\right)$ exists at point 0, then

$$\mathbf{E}\left(X^k\right) = \frac{\varphi_X^{(k)}(0)}{i^k}.$$

3. If the derivative $\varphi_X^{(2k)}(0)$ is finite for a positive integer k, then the moment $\mathbf{E}\left(X^{2k}\right)$ exists. Note that from the existence of the finite derivative $\varphi_X^{(2k+1)}(0)$ only the existence of the finite moment $\mathbf{E}\left(X^{2k}\right)$ follows.
4. Let $X_1, \ldots, X_n$ be independent random variables; then the characteristic function of the sum $X_1 + \ldots + X_n$ equals the product of the characteristic functions of the random variables X_i, that is,

$$\begin{aligned}\varphi_{X_1+\ldots+X_n}(s) &= \mathbf{E}\left(\mathrm{e}^{is(X_1+\ldots+X_n)}\right) = \mathbf{E}\left(\mathrm{e}^{isX_1}\ldots\mathrm{e}^{isX_n}\right)\\ &= \mathbf{E}\left(\mathrm{e}^{isX_1}\right)\cdot\ldots\cdot\mathbf{E}\left(\mathrm{e}^{isX_n}\right) = \varphi_{X_1}(s)\ldots\varphi_{X_1}(s).\end{aligned}$$

Note that property 4 plays an important role in the limit theorems of probability theory.

Laplace–Stieltjes and Laplace Transforms If, instead of the CDFs, the Laplace–Stieltjes and Laplace transforms were used, the problem could be solved much easier in many practical cases and the results could additionally often be given in more compact form. Let X be a nonnegative random variable with the CDF $F(x)$ ($F(0) = 0$). Then the real or, in general, complex varying function

$$F^{\sim}(s) = \mathbf{E}\left(\mathrm{e}^{-sX}\right) = \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}F(x),\ \mathrm{Re}s \geq 0,\ F^{\sim}(0) = 1$$

is called the **Laplace–Stieltjes transform** of the CDF F. Since $\left|\mathrm{e}^{-sX}\right| \leq 1$ if $\mathrm{Re}s \geq 0$, then the function $F^{\sim}(s)$ is well defined. If f is a PDF, then the function

$$f^*(s) = \int_0^\infty \mathrm{e}^{-sx} f(x)\mathrm{d}x,\ \mathrm{Re}s \geq 0,$$

is called the **Laplace transform** of the function f. These notations will be used even if the functions F and f do not possess the necessary properties of distribution and PDFs but $F^{\sim}(s)$ and $f^*(s)$ are well defined. If f is a PDF related to the CDF F, then the equality

$$F^{\sim}(s) = f^*(s) = sF^*(s) \tag{1.3}$$

holds.

Proof. It is clear that

$$F^{\sim}(s) = \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}F(x) = \int_0^\infty \mathrm{e}^{-sx} f(x)\mathrm{d}x = f^*(s),$$

and integrating by parts we have

$$F^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx} \mathrm{d}F(x) = \int_0^{\infty} s\mathrm{e}^{-sx} F(x)\mathrm{d}x = sF^*(s).$$

□

Since the preceding equation is true between the two introduced transforms, it is enough to consider the Laplace–Stieltjes transform only and to enumerate its main properties.

(a) $F^{\sim}(s)$, $\mathrm{Re}s \geq 0$ is a continuous function and $0 \leq |F^{\sim}(s)| \leq 1$, $\mathrm{Re}s \geq 0$.
(b) $F^{\sim}_{aX+b}(s) = \mathrm{e}^{-bs} F^{\sim}(as)$.
(c) For all positive integers k

$$(-1)^k F^{\sim(k)}(s) = \int_0^{\infty} x^k \mathrm{e}^{-sx} \mathrm{d}F(x), \quad \mathrm{Re}s > 0.$$

If the kth moment $\mu_k = \mathbf{E}\left(X^k\right)$ exists, then $\mu_k = (-1)^k F^{\sim(k)}(0)$.
(d) If the nonnegative random variables X and Y are independent, then

$$F^{\sim}_{X+Y}(s) = F^{\sim}_X(s) F^{\sim}_Y(s).$$

(e) For all continuity points of the CDF F the inversion formula

$$F(x) = \lim_{a\to\infty} \sum_{n\leq ax} (-1)^n (F^{\sim}(a))^{(n)} \frac{a^n}{n!}$$

is true.

Covariance and Correlation Let X and Y be two random variables with finite variances σ_X^2 and σ_Y^2, respectively. The **covariance** between the pair of random variables (X, Y) is defined as

$$\mathrm{cov}(X, Y) = \mathbf{E}\left((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))\right).$$

The covariance can be rewritten in the simple computational form

$$\mathrm{cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y).$$

If the variances σ_X^2 and σ_Y^2 satisfy the conditions $\mathbf{D}(X) > 0$, $\mathbf{D}(Y) > 0$, then the quantity

$$\mathrm{corr}(X, Y) = \mathrm{cov}\left(\frac{X - \mathbf{E}(X)}{\mathbf{D}(X)}, \frac{Y - \mathbf{E}(Y)}{\mathbf{D}(Y)}\right) = \frac{\mathrm{cov}(X, Y)}{\mathbf{D}(X)\mathbf{D}(Y)}$$

is called the **correlation** between the pair of random variables (X, Y).

Correlation can be used as a measure of the dependence between random variables. It is always true that

$$-1 \leq \operatorname{corr}(X, Y) \leq 1,$$

provided that the variances of random variables X and Y are finite and nonzero.

Proof. Since by the Cauchy–Schwartz inequality for all random variables U and V with finite second moments

$$(\mathbf{E}(UV))^2 \leq \mathbf{E}\left(U^2\right)\mathbf{E}\left(V^2\right),$$

therefore

$$(\operatorname{cov}(X, Y))^2 \leq \mathbf{E}\left((X - \mathbf{E}(X))^2\right)\mathbf{E}\left((Y - \mathbf{E}(Y))^2\right) = \mathbf{D}^2(X)\mathbf{D}^2(Y),$$

from which the inequality $|\operatorname{corr}(X, Y)| \leq 1$ immediately follows. □

It can also be proved that the equality $|\operatorname{corr}(X, Y)| = 1$ holds if and only if a linear relation exists between random variables X and Y with probability 1, that is, there are two constants a and b for which $\mathbf{P}(Y = aX + b) = 1$.

Both covariance and correlation play essential roles in multivariate statistical analysis. Let $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ be a column vector whose n elements $X_1, \ldots, X_n$ are random variables. Here it should be noted that in probability theory and statistics usually column vectors are applied, but in queueing theory row vectors are used if Markov processes are considered. We define

$$\mathbf{E}(\mathbf{X}) = (\mathbf{E}(X_1), \ldots, \mathbf{E}(X_n))^{\mathrm{T}},$$

provided that the expected values of components exist. The upper index T denotes the transpose of vectors or matrices. Similarly, if a matrix $W = \left(W_{ij}\right) \in \mathbb{R}^{k \times m}$ is given whose elements W_{ij} are random variables of finite expected values, then we define

$$\mathbf{E}(W) = \left(\mathbf{E}\left(W_{ij}\right)\right),\ 1 \leq i \leq k,\ 1 \leq j \leq m).$$

If the variances of components of a random vector $X = (X_1, \ldots, X_k)^{\mathrm{T}}$ are finite, then the matrix

$$R = \mathbf{E}\left((X - \mathbf{E}(X))(X - \mathbf{E}(X))^{\mathrm{T}}\right) \tag{1.4}$$

is called a **covariance matrix** of X. It can be seen that the (i, j) entries of matrix R are $R_{ij} = \operatorname{cov}(X_i, X_j)$, which are the covariances between the random variables X_i and X_j.

The covariance matrix can be defined in cases where the components of X are complex valued random variables replacing in definition (1.4) $(X - \mathbf{E}(X))^{\mathrm{T}}$ by $(X - \mathbf{E}(X))^{*\mathrm{T}}$ the complex composed of transpose.

An important property of a covariance matrix R is that it is nonnegative definite, i.e., for all real or complex k-dimensional column vectors $z = (z_1, \dots, z_k)^{\mathrm{T}}$ the inequality

$$zRz^{\mathrm{T}} \geq 0$$

holds.

The matrix $r = (r_{i,j})$ with components $r_{i,j} = \operatorname{corr}(X_i, X_j)$, $1 \leq i \leq k$, $1 \leq j \leq m$ is called a **correlation matrix** of random vector X.

Conditional Expectation and Its Properties The notion of conditional expectation is defined with the help of results of set and measure theories. We present the general concept and important properties and illustrate the important special cases.

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a fixed probability space, and let X be a random variable whose expected value exists. Let $\mathcal{C}$ be an arbitrary sub-σ-algebra of $\mathcal{A}$. We wish to define the conditional expectation $Z = \mathbf{E}(X|\mathcal{C})$ of X given $\mathcal{C}$ as a $\mathcal{C}$-measurable random variable for which the random variable satisfies the condition $\mathbf{E}\left(\mathbf{E}(X|\mathcal{C})\mathcal{I}_{\{C\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{C\}}\right)$ for all $C \in \mathcal{C}$. As a consequence of the Radon–Nikodym theorem, a random variable Z exists with probability 1 that satisfies the required conditions.

Definition 1.41. Random variable Z is called the **conditional expectation** of X given σ-algebra $\mathcal{C}$ if the following conditions hold:

(a) Z is a $\mathcal{C}$-measurable random variable.
(b) $\mathbf{E}\left(\mathbf{E}(X|\mathcal{C})\mathcal{I}_{\{C\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{C\}}\right)$ for all $C \in \mathcal{C}$.

Definition 1.42. Let $A \in \mathcal{A}$ be an event. The random variable $\mathbf{P}(A|\mathcal{C}) = \mathbf{E}\left(\mathcal{I}_{\{A\}}|\mathcal{C}\right)$ is called the **conditional expectation** of event A given σ-algebra $\mathcal{C}$.

Important Properties of Conditional Expectation Let $\mathcal{C}$, $\mathcal{C}_1$, and $\mathcal{C}_2$ be sub-σ-algebras of $\mathcal{A}$, and let X, X_1, and X_2 be random variables with finite expected values. Then the following relations hold with probability 1:

1. $\mathbf{E}(\mathbf{E}(X|\mathcal{C})) = \mathbf{E}(X)$.
2. $\mathbf{E}(cX|\mathcal{C}) = c\mathbf{E}(X|\mathcal{C})$ for all constant c.
3. If $\mathcal{C}_0 = \{\varnothing, \Omega\}$ is the trivial σ-algebra, then $\mathbf{E}(X|\mathcal{C}_0) = \mathbf{E}(X)$.
4. If $\mathcal{C}_1 \subset \mathcal{C}_2$, then $\mathbf{E}(\mathbf{E}(X|\mathcal{C}_1)|\mathcal{C}_2) = \mathbf{E}(\mathbf{E}(X|\mathcal{C}_2)|\mathcal{C}_1) = \mathbf{E}(X|\mathcal{C}_1)$.
5. If random variable X does not depend on the σ-algebra $\mathcal{C}$, i.e., if for all Borel sets $D \in \mathcal{B}$ and for all events $A \in \mathcal{C}$ the equality $\mathbf{P}(X \in D, A) = \mathbf{P}(X \in D)\mathbf{P}(A)$ holds, then $\mathbf{E}(X|\mathcal{C}) = \mathbf{E}(X)$.
6. $\mathbf{E}(X_1 + X_2|\mathcal{C}) = \mathbf{E}(X_1|\mathcal{C}) + \mathbf{E}(X_2|\mathcal{C})$.
7. If the random variable X_1 is $\mathcal{C}$-measurable, then $\mathbf{E}(X_1X_2|\mathcal{C}) = X_1\mathbf{E}(X_2|\mathcal{C})$.

Definition 1.43. Let Y be a random variable, and denote by $\mathcal{A}_Y$ the σ-algebra generated by random variable Y, i.e., let $\mathcal{A}_Y$ be the minimal sub-σ-algebra of $\mathcal{A}$ for which Y is $\mathcal{A}_Y$-measurable. The random variable $\mathbf{E}(X|Y) = \mathbf{E}(X|\mathcal{C}_Y)$ is called the **conditional expectation** of X given random variable Y.

Main Properties of Conditional Expectation Firstly, consider the case where random variable Y is discrete and takes values in the set $\mathcal{Y} = \{y_1, \dots, y_n\}$ and

$\mathbf{P}(Y = y_i) > 0$, $1 \le i \le n$. We then define the events $C_i = \{Y = y_i\}$, $1 \le i \le n$. It is clear that the collection of events $\{C_1, \ldots, C_n\}$ forms a complete system of events, i.e., they are mutually exclusive, $\mathbf{P}(C_i) > 0$, $\le i \le n$ and $\mathbf{P}(C_1) + \ldots + \mathbf{P}(C_n) = 1$. The σ-algebra $\mathcal{C}_Y = \sigma(C_1, \ldots, C_n) \subset \mathcal{A}$, which is generated by random variable Y, is the set of events consisting of all subsets of $\{C_1, \ldots, C_n\}$. Note that here we can write "algebra" instead of "σ-algebra" because the set $\{C_1, \ldots, C_n\}$ is finite. Since the events C_i have positive probability, the conditional probabilities

$$\mathbf{E}(X|C_i) = \frac{\mathbf{E}\left(X\mathcal{I}_{\{C_i\}}\right)}{\mathbf{P}(C_i)}$$

are well defined.

Theorem 1.44. *The conditional expectation $\boldsymbol{E}(X|\mathcal{C}_Y)$ satisfies the relation*

$$\boldsymbol{E}(X|\mathcal{C}_Y) = \boldsymbol{E}(X|\mathcal{C}_Y)(\omega) = \sum_{k=1}^{n} \boldsymbol{E}(X|C_k)\mathcal{I}_{\{C_k\}} \text{ with probability } 1. \tag{1.5}$$

Note that Eq. (1.5) can also be rewritten in the form

$$\mathbf{E}(X|Y) = \mathbf{E}(X|Y)(\omega) = \sum_{k=1}^{n} \mathbf{E}(X|Y = y_k)\mathcal{I}_{\{Y=y_k\}}. \tag{1.6}$$

Proof. Since the relation

$$\{\mathbf{E}(X|\mathcal{C}_Y) \le x\} = \cup\{C_i : \ \mathbf{E}(X|C_i) \le x\} \in \mathcal{C}_Y$$

holds for all $x \in \mathbb{R}$, then $\mathbf{E}(X|\mathcal{C}_Y)$ is a $\mathcal{C}_Y$-measurable random variable. On the other hand, if $C \in \mathcal{C}_Y$, $C \ne \{\varnothing\}$, then $C = \cup\{C_i : i \in K\}$ stands with an appropriately chosen set of indices $K \subset \{1, \ldots, n\}$, and we obtain

$$\begin{aligned}\mathbf{E}\left(\mathbf{E}(X|\mathcal{C}_Y)\mathcal{I}_{\{C\}}\right) &= \mathbf{E}\left(\sum_{k\in K} \mathbf{E}(X|C_k)\mathcal{I}_{\{C_k\}}\right) \\ &= \sum_{k\in K} \mathbf{E}(X|C_k)\mathbf{P}(C_k) = \sum_{k\in K} \mathbf{E}\left(X\mathcal{I}_{\{C_k\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{C\}}\right).\end{aligned}$$

If $C = \{\varnothing\}$, then $\mathbf{E}\left(\mathbf{E}(X|\mathcal{C}_Y)\mathcal{I}_{\{C\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{C\}}\right) = 0$. Thus we have proved that random variable (1.5) satisfies all the required properties of conditional expectation. □

Comment 1.45. *From expression (1.6) the following relation can be obtained:*

$$\boldsymbol{E}(X) = \boldsymbol{E}(\boldsymbol{E}(X|Y)) = \int_{-\infty}^{\infty} \boldsymbol{E}(X|Y = y)\mathrm{d}F_Y(y). \tag{1.7}$$

This relation remains valid if, instead of the finite set $\mathcal{Y} = \{y_1, \ldots, y_n\}$*, we choose the countable infinite set* $\mathcal{Y} = \{y_i,\ i \in I\}$ *for which* $\boldsymbol{P}(Y = y_i) > 0$*,* $i \in I$*.*

Comment 1.46. *Denote the function* g *by the relation*

$$g(y) = \begin{cases} \boldsymbol{E}(X|Y = y_k), & \text{if } y = y_k \text{ for an index } k, \\ 0, & \text{otherwise.} \end{cases} \tag{1.8}$$

Then, using formula (1.6), the conditional expectation of X *given* Y *can be obtained with the help of the function* g *as follows:*

$$\boldsymbol{E}(X|Y) = g(Y) \tag{1.9}$$

with probability 1.

Continuous Random Variables (X, Y) Consider a pair of random variables (X, Y) having joint density $f_{X,Y}(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$, respectively. Then the conditional density $f_{X|Y}(x|y)$ exists and, according to (1.1), can be defined as

$$f_{X|Y}(x|y) = \begin{cases} \dfrac{f_{XY}(x, y)}{f_Y(y)}, & \text{if } f_Y(y) > 0, \\ 0, & \text{otherwise}. \end{cases}$$

Define $g(y) = \mathbf{E}(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\mathrm{d}x$. Then the conditional expectation of X given Y can be determined with probability 1 as follows:

$$\mathbf{E}(X|Y) = g(Y),$$

and so we can define

$$\mathbf{E}(X|Y = y) = g(y).$$

Proof. It is clear that $g(Y)$ is a $\mathcal{C}_Y$-measurable random variable; therefore, it is enough to prove that the equality

$$\mathbf{E}\left(\mathbf{E}(X|Y)\mathcal{I}_{\{Y \in D\}}\right) = \mathbf{E}\left(X\mathcal{I}_{\{Y \in D\}}\right)$$

holds for all Borel sets D of a real line. It is not difficult to see that

$$\mathbf{E}\left(\mathbf{E}(X|Y)\mathcal{I}_{\{Y \in D\}}\right) = \mathbf{E}\left(g(Y)\mathcal{I}_{\{Y \in D\}}\right)$$

$$= \int_D \int_{-\infty}^{\infty} x \, \frac{f_{XY}(x,y)}{f_Y(y)} \, f_Y(y) \mathrm{d}x \mathrm{d}y$$

$$= \int_D \int_{-\infty}^{\infty} x f_{XY}(x,y) \mathrm{d}x \mathrm{d}y$$

and, on other hand,

$$\mathbf{E}\left(X \mathcal{I}_{\{Y \in D\}}\right) = \int_D \int_{-\infty}^{\infty} x f_{XY}(x,y) \mathrm{d}x \mathrm{d}y.$$

□

Comment 1.47. *In the case where a pair of random variables has a joint normal distribution, the conditional expectation* $\boldsymbol{E}(X|Y)$ *is a linear function of random variable* Y *with probability* 1*, that is, the regression function* g *is a linear function and the relation*

$$\boldsymbol{E}(X|Y) = \boldsymbol{E}(X) + \frac{\operatorname{cov}(X,Y)}{\boldsymbol{D}(X)}(X - \boldsymbol{E}(X))$$

holds.

General Case By the definition of conditional expectation, $\mathbf{E}(X|Y)$ is $\mathcal{C}_Y$-measurable; therefore, there is a Borel-measurable function g such that $\mathbf{E}(X|Y)$ can be given with probability 1 in the form

$$\mathbf{E}(X|Y) = g(Y). \tag{1.10}$$

This relation makes it possible to give the conditional expectation $\mathbf{E}(X|Y = y)$ as the function

$$\mathbf{E}(X|Y = y) = g(y),$$

which is called a **regression function**. It is clear that the regression function is not necessarily unique and is determined on a Borel set of the real line D satisfying the condition $\mathbf{P}(Y \in D) = 1$.

Comment 1.48. *Let* X *and* Y *be two random variables. Assume that* X *has finite variation. Consider the quadratic distance* $\boldsymbol{E}\left([X - h(Y)]^2\right)$ *for the set* $\mathcal{H}_Y$ *of all Borel-measurable functions* h*, for which* $h(Y)$ *has finite variation. Then the assertion*

$$\min\left\{\boldsymbol{E}\left([X - h(Y)]^2\right) : \; h \in \mathcal{H}_Y\right\} = \boldsymbol{E}\left([X - g(Y)]^2\right)$$

holds. This relation implies that the best approximation of X by Borel-measurable functions of Y in a quadratic mean is the regression $\boldsymbol{E}(X|Y) = g(Y)$.

Formula of Total Expected Value A useful formula can be given to compute the expected value of random variable X if the regression function $\mathbf{E}(X|Y = y)$ can be determined.

Making use of relation 1 given as a general property of conditional expectation and Eq. (1.10), it is clear that

$$\begin{aligned}\mathbf{E}(X) &= \mathbf{E}(\mathbf{E}(X|Y)) = \mathbf{E}(g(Y)) \\ &= \int_{-\infty}^{\infty} g(y)\mathrm{d}F_Y(y) = \int_{-\infty}^{\infty} \mathbf{E}(X|Y = y)\mathrm{d}F_Y(y).\end{aligned}$$

From this relation we have the so-called **formula of total expected value**. If random variable Y has discrete or continuous distributions, then we have the formulas

$$\mathbf{E}(X) = \sum_{i \in I} \mathbf{E}(X|Y = y_i)\mathbf{P}(Y = y_i)$$

and

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} \mathbf{E}(X|Y = y) f_Y(y)\mathrm{d}y.$$

1.2 Frequently Used Discrete and Continuous Distributions

In this part we consider some frequently used distributions and give their definitions and important characteristics. In addition to the formal description of the distributions, we will give appropriate mathematical models that lead to a given distribution. If the distribution function of a random variable is given as a function $F_X(x; a_1, \ldots, a_n)$ depending on a positive integer n and constants $a_1, \ldots, a_n$, then $a_1, \ldots, a_n$ are called the parameters of the density function F_X.

1.2.1 Discrete Distributions

Bernoulli Distribution $Be(p)$, $0 \leq p \leq 1$. The PDF of random variable X with values $\{0, 1\}$ is called a Bernoulli distribution if

$$p_k = \mathbf{P}(X = k) = \begin{cases} p, & \text{if } k = 1, \\ 1 - p, & \text{if } k = 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = p,\ \mathbf{D}^2(X) = p(1-p)$;
Generating function: $1 - p + pz$;
Characteristic function: $1 - p + pe^{it}$.

Example. Let X be the number of heads appearing in one toss of a coin, where

$$p = \mathbf{P}(\text{head appearing in a toss}).$$

Then X has a $Be(p)$ distribution.

Binomial Distribution $B(n, p)$. The distribution of a discrete random variable X with values $\{0, 1, \ldots, n\}$ is called binomial with the parameters n and p, $0 < p < 1$, if its PDF is

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

Expected value and variance: $\mathbf{E}(X) = np,\ \ \mathbf{D}^2(X) = np(1-p)$;
Generating function: $G(z) = (pz + (1-p))^n$;
Characteristic function: $\varphi(t) = (1 + p(e^{it} - 1))^n$.

Example. Consider an experiment in which we observe that an event A with probability $p = \mathbf{P}(A)$, $0 < p < 1$, occurs (success) or not (failure). Repeating the experiment n times independently, define random variable X by the frequency of event A. Then the random variable has a $B(n, p)$ PDF.

Note that if the $Be(n, p)$ random variables $X_1, \ldots, X_n$ are independent, then the random variable $X = X_1 + \ldots + X_n$ has a $B(n, p)$ distribution.

Polynomial Distribution The PDF of a random vector $X = (X_1, \ldots, X_k)^{\mathrm{T}}$ taking values in the set $\{(n_1, \ldots, n_k) : n_i \geq 0,\ n_1 + \ldots + n_k = n\}$ is called polynomial with the parameters n and $p_1, \ldots, p_k$ ($p_i > 0,\ p_1 + \ldots + p_k = 1$) if X has a PDF

$$p_{n_1,\ldots,n_k} = \mathbf{P}(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k}.$$

Note that each coordinate variable X_i of random vector X has a $B(p_i, n)$ binomial distribution whose expected value and variance are np_i and $np_i(1-p_i)$.

Expected value $\mathbf{E}(X) = (np_1, \ldots, np_n)^{\mathrm{T}}$;
Covariance matrix $R = (R_{ij})_{1 \leq i,j \leq k}$, where $R_{ij} = \begin{cases} np_i(1-p_i), & \text{if } i = j, \\ np_i p_j, & \text{if } i \neq j; \end{cases}$
Characteristic function: $\varphi(t_1, \ldots, t_k) = (p_1 e^{it_1} + \ldots + p_k e^{it_k})^n$.

Example. Let $A_1, \ldots, A_k$ be k disjoint events for which $p_i = \mathbf{P}(A_i) > 0,\ p_1 + \ldots + p_k = 1$. Consider an experiment with possible outcomes $A_1, \ldots, A_k$ and repeat it n times independently. Denote by X_i the frequency of event A_i in the series of n

observations. Then the distribution of X is polynomial with the parameters n and $p_1, \ldots, p_k$.

Geometric Distribution The PDF of random variable X taking values in $\{1, 2, \ldots\}$ is called a geometric distribution with the parameter p, $0 < p < 1$, if its PDF is

$$p_k = \mathbf{P}(X = k) = (1-p)^{k-1} p,\ k = 1, 2, \ldots .$$

Expected value and variance:	$\mathbf{E}(X) = \frac{1}{p}$, $\mathbf{D}^2(X) = \frac{1-p}{p^2}$;
Generating function:	$G(z) = \frac{pz}{1-(1-p)z}$;
Characteristic function:	$\varphi(t) = \frac{p}{1-(1-p)e^{it}}$.

Theorem 1.49. *If X has a geometric distribution, then X has a so-called **memoryless property**, that is, for all nonnegative integer numbers i, j the following relation holds:*

$$\boldsymbol{P}(X \geq i + j \,|\, X \geq i) = \boldsymbol{P}(X \geq j).$$

Proof. It is easy to verify that for $k \geq 1$

$$\begin{aligned}\mathbf{P}(X \geq k) &= \sum_{\ell=k}^{\infty} \mathbf{P}(X = \ell) = \sum_{\ell=k}^{\infty} (1-p)^{\ell-1} p \\ &= (1-p)^{k-1} p \sum_{\ell=0}^{\infty} (1-p)^{\ell} = (1-p)^{k-1};\end{aligned}$$

therefore,

$$\begin{aligned}\mathbf{P}(X \geq i + j \,|\, X \geq i) &= \frac{\mathbf{P}(X \geq i + j, X \geq i)}{\mathbf{P}(X \geq i)} \\ &= \frac{\mathbf{P}(X \geq i + j)}{\mathbf{P}(X \geq i)} \\ &= \frac{(1-p)^{i+j-1}}{(1-p)^{i-1}} = (1-p)^j,\ \ j = 0, 1, \ldots .\end{aligned}$$

□

Note that a geometric distribution is sometimes defined on the set $\{0, 1, 2, \ldots\}$ instead of $\{1, 2, \ldots, \}$; in this case, the PDF is determined by

$$p_k = (1-p)^k p,\ k = 0, 1, 2, \ldots .$$

Example. Consider a sequence of experiments and observe whether an event A, $p = \mathbf{P}(A) > 0$, occurs (success) or does not (failure) in each step. If the event

occurs in the kth step first, then define the random variable as $X = k$. In other words, let X be the number of Bernoulli trials of the first success. Then random variable X has a geometric distribution with the parameter p.

Negative Binomial Distribution The distribution of random variable X taking values in $\{0, 1, \ldots\}$ is called a negative binomial distribution with the parameter p, $0 < p < 1$, if

$$p_k = \mathbf{P}(X = k + r) = \binom{r + k - 1}{k} (1 - p)^k p^r, \quad k = 0, 1, \ldots .$$

Expected value and variance: $\mathbf{E}(X) = r\frac{1}{p}, \quad \mathbf{D}^2(X) = r\frac{1-p}{p^2}$;

Generating function: $G(z) = \left(\frac{pz}{1-(1-p)z}\right)^r$;

Characteristic function: $\varphi(t) = p^r \left(1 - (1 - p)\mathrm{e}^{it}\right)^{-r}$.

Example. Let p, $0 < p < 1$, and the positive integer r be two given constants. Suppose that we are given a coin that has a probability p of coming up heads. Toss the coin repeatedly until the rth head appears and define by X the number of tosses. Then random variable X has a negative binomial distribution with parameters (p, r).

Note that from this example it immediately follows that X has a geometric distribution with the parameter p when $r = 1$.

Poisson Distribution The PDF of a random variable X is called a Poisson distribution with the parameter λ ($\lambda > 0$) if X takes values in $\{0, 1, \ldots\}$ and

$$p_k = \mathbf{P}(X = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \quad k = 0, 1, \ldots .$$

Expected value and variance: $\mathbf{E}(X) = \lambda, \quad \mathbf{D}^2(X) = \lambda$;

Generating function: $G(z) = \mathrm{e}^{\lambda(z-1)}$;

Characteristic function: $\varphi(t) = \mathrm{e}^{\lambda(\mathrm{e}^{it}-1)}$.

The following theorem establishes that a binomial distribution can be approximated with a Poisson distribution with the parameter λ when the parameters (p, n) of the binomial distribution satisfy the condition $np \to \lambda$, $n \to \infty$.

Theorem 1.50. *Consider a binomial distribution with the parameter (p, n). Assume that for a fixed constant λ, $\lambda > 0$, the convergence $np \to \lambda$, $n \to \infty$, holds; then the limit of probabilities satisfies the relation*

$$\binom{n}{k} p^k (1 - p)^{n-k} \to \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \quad k = 0, 1, \ldots .$$

Proof. For any fixed $k \geq 0$ integer number we have

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{(np)((n-1)p)\dots((n-k+1)p)}{k!} \mathrm{e}^{(n-k)\log(1-p)}.$$

Since $np \to \lambda$, $n \to \infty$, therefore $p \to 0$, and we obtain

$$\frac{(np)((n-1)p)\dots((n-k+1)p)}{1 \cdot 2 \cdot \dots \cdot k} \to \frac{\lambda^k}{k!}, \quad np \to \lambda.$$

On the other hand, if $p \to 0$, then we get the asymptotic relation $\log(1-p) = -p + o(p)$. Consequently,

$$(n-k)\log(1-p) = -(n-k)(p+o(p)) \to -\lambda, \quad np \to \lambda,\ n \to \infty;$$

therefore, using the last two asymptotic relations, the assertion of the theorem immediately follows. □

1.2.2 Continuous Distributions

Uniform Distribution Let a, b $(a < b)$ be two real numbers. The distribution of random variable X is called uniform on the interval (a, b) if its PDF is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{ha } x \in (a,b), \\ 0, & \text{ha } x \notin (a,b). \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = \frac{a+b}{2}$, $\mathbf{D}^2(X) = \frac{(b-a)^2}{12}$;
Characteristic function: $\varphi(t) = \frac{1}{b-a} \frac{\mathrm{e}^{itb} - \mathrm{e}^{ita}}{it}$.

Note that if X has a uniform distribution on the interval (a, b), then the random variable $Y = \frac{X-a}{b-a}$ is distributed uniformly on the interval $(0, 1)$.

Exponential Distribution $\mathrm{Exp}(\lambda)$, $\lambda > 0$. The distribution of a random variable X is called exponential with the parameter λ, $\lambda > 0$, if its PDF

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = \frac{1}{\lambda}$, $\mathbf{D}^2(X) = \frac{1}{\lambda^2}$;
Characteristic function: $\varphi(t) = \frac{\lambda}{\lambda - it}$.

The Laplace and Laplace–Stieltjes transforms of the density and distribution function of an $\mathrm{Exp}(\lambda)$ distribution are determined as

$$\mathbf{E}\left(\mathrm{e}^{-sX}\right) = f^*(s) = F^{\sim}(s) = \frac{\lambda}{s+\lambda}.$$

The exponential distribution, similarly to the geometric distribution, has the memoryless property.

Theorem 1.51. *For arbitrary constants $t, s > 0$ the relation*

$$\boldsymbol{P}(X > t+s | X > t) = \boldsymbol{P}(X > s)$$

holds.

Proof. It is clear that

$$\mathbf{P}(X > t+s | X > t) = \frac{\mathbf{P}(X > t+s, X > t)}{\mathbf{P}(X > t)} = \\ = \frac{\mathbf{P}(X > t+s)}{\mathbf{P}(X > t)} = \frac{\mathrm{e}^{-\lambda(t+s)}}{\mathrm{e}^{-\lambda t}} = \mathrm{e}^{-\lambda s}.$$

□

Hyperexponential Distribution Let the PDF of random variable X be a mixture of exponential distributions with the parameters $\lambda_1, \ldots, \lambda_n$ and with weights $a_1, \ldots, a_n$ ($a_k > 0, a_1 + \ldots + a_n = 1$). Then the PDF

$$f(x) = \begin{cases} \sum\limits_{k=1}^{n} a_k \lambda_k \mathrm{e}^{-\lambda_k x} & \text{if } x > 0, \\ 0, & \text{if } x \le 0, \end{cases}$$

of random variable X is called hyperexponential.

Expected value and variance: $\mathbf{E}(X) = \sum\limits_{k=1}^{n} \frac{a_k}{\lambda_k}$, $\mathbf{D}^2(X) = 2\sum\limits_{k=1}^{n} \frac{a_k}{\lambda_k^2} - \left(\sum\limits_{k=1}^{n} \frac{a_k}{\lambda_k}\right)^2$;

Characteristic function: $\varphi(t) = \sum\limits_{k=1}^{n} a_k \frac{\lambda_k}{\lambda_k - it}$.

Denote by $\Gamma(x) = \int\limits_0^{\infty} y^{x-1}\mathrm{e}^{-y}\mathrm{d}y$, $x > -1$ the well-known **gamma function** Γ in analysis, which is necessary for the definition of the gamma distribution.

Gamma Distribution Gamma(α, λ), $\alpha, \lambda > 0$.

The distribution of a random variable X is called a **gamma distribution** with the parameters $\alpha, \lambda > 0$, if its PDF is

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = \frac{\alpha}{\lambda}, \quad \mathbf{D}^2(X) = \frac{\alpha}{\lambda^2}$;
Characteristic function: $\varphi(t) = \left(\frac{\lambda}{\lambda - it}\right)^\alpha$.

Comment 1.52. *A gamma distribution with the parameters $\alpha = n$, $\lambda = n\mu$ is called an **Erlang distribution**.*

Comment 1.53. *If the independent identically distributed random variables $X_1, X_2, \ldots$ have an exponential distribution with the parameter λ, then the distribution of the sum $Z = X_1 + \ldots + X_n$ is a gamma distribution with the parameter (n, λ). This relation is easy to see because the characteristic function of an exponential distribution with the parameter λ is $(1 - it/\lambda)^{-1}$; then the characteristic function of its nth convolution power is $(1 - it/\lambda)^{-n}$, which equals the characteristic function of a Gamma(n, λ) distribution.*

Beta Distribution Beta(a, b), $a, b > 0$. The distribution of random variable X is called a beta distribution if its PDF is

$$f(x) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, & \text{if } x \in (0,1), \\ 0, & \text{if } x \notin (0,1). \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = \frac{a}{a+b}, \quad \mathbf{D}^2(X) = \frac{ab}{(a+b)^2(a+b+1)}$;
Characteristic function in the form of power series: $\varphi(t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha+\beta+k)}$.

Gaussian (Also Called Normal) Distribution $N(\mu, \lambda)$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$. The distribution of random variable X is called Gaussian with the parameters (μ,σ) if it has a PDF

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty.$$

Expected value and variance: $\mu = \mathbf{E}(X)$ and $\sigma^2 = \mathbf{D}^2(X)$;
Characteristic function: $\varphi(t) = \exp\left\{i\mu t - \frac{\sigma^2}{2} t^2\right\}$.

The $N(0, 1)$ distribution is usually called a standard Gaussian or standard normal distribution, and its PDF is equal to

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}, \quad -\infty < x < \infty.$$

It is easy to verify that if a random variable has an $N(\mu, \sigma)$ distribution, then the centered and linearly normed random variable $Y = (X - \mu)/\sigma$ has a standard Gaussian distribution.

Multidimensional Gaussian (Normal) Distribution $N(\mu, \mathbf{R})$ Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ be an n-dimensional random vector whose coordinates $Z_1, \ldots, Z_n$ are independent and have a standard $N(0, 1)$ Gaussian distribution. Let $\mathbf{V} \in \mathbb{R}^{m \times n}$ be an $(m \times n)$ matrix and $\mu = (\mu_1, \ldots, \mu_m)^{\mathrm{T}} \in \mathbb{R}^m$ an m-dimensional vector. Then the distribution of the m-dimensional random vector X defined by the equation $\mathbf{X} = \mathbf{VZ} + \mu$ is called an m-dimensional Gaussian distribution.

Expected value and variance matrix:

$$\mathbf{E}(\mathbf{X}) = \mu_{\mathbf{X}} = \mu \qquad \text{and} \qquad \mathbf{D}^2(\mathbf{X}) = \mathbf{R}_{\mathbf{X}} = \mathbf{E}\left((\mathbf{X} - \mu)(\mathbf{X} - \lambda)^{\mathrm{T}}\right) = \mathbf{V}\mathbf{V}^{\mathrm{T}};$$

Characteristic function:

$$\varphi(\mathbf{t}) = \exp\left\{ i\mathbf{t}^{\mathrm{T}}\mu - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{R}_{\mathbf{X}}\mathbf{t} \right\}, \text{ where } \mathbf{t} = (t_1, \ldots, t_m)^{\mathrm{T}} \in \mathbb{R}^m.$$

If $\mathbf{V}$ is a nonsingular quadratic matrix ($m = n$ and $\det \mathbf{V} \neq 0$), then the random vector $\mathbf{X}$ has a density in the form

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi \det \mathbf{R}_X)^{n/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\mathbf{R}_X^{-1}(\mathbf{x}-\mu) \right\}, \quad \mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n.$$

Example. If the random vector $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$ has a two-dimensional Gaussian distribution with expected value $\mu = (\mu_1, \mu_2)^{\mathrm{T}}$ and covariance matrix

$$R_X = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

then its PDF has the form

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\sqrt{ac - b^2}}{2\pi} \exp\left\{ -\frac{1}{2}[a(x_1 - \mu_1)^2 + 2b(x_1 - \mu_1)(x_2 - \mu_2) + c(x_2 - \mu_2)^2] \right\},$$

where a, b, c, μ_1, μ_2 are constants satisfying the conditions $a > 0$, $c > 0$, and $b^2 < ac$.

Note that the marginal distributions of random variables X_1 and X_2 are $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ Gaussian, respectively, where

$$\sigma_1 = \sqrt{\frac{a}{ac - b^2}}, \quad \sigma_2 = \sqrt{\frac{c}{ac - b^2}} \quad \text{and} \quad b = \operatorname{cov}(X_1, X_2).$$

Distribution Functions Associated with Gaussian Distributions Let $Z, Z_1, Z_2, \ldots$ be independent random variables whose distributions are standard Gaussian, i.e., with the parameters $(0, 1)$. There are many distributions, for example the χ^2 and the logarithmically normal distributions defined subsequently (further examples are the frequently used t, F, and Wishart distributions in statistics [46]), that can be given as distributions of appropriately chosen functions of random variables $Z, Z_1, Z_2, \ldots$.

χ^2 Distribution The distribution of the random variable $X = Z_1^2 + \ldots + Z_n^2$ is called a χ^2 distribution with parameter n. The PDF is

$$f_n(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2a)} x^{n/2-1} \mathrm{e}^{-x/2}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = n, \quad \mathbf{D}^2(X) = 2n;$
Characteristic function: $\varphi(t) = (1 - 2it)^{-n/2}.$

Logarithmic Gaussian (Normal) Distribution If random variable Z has an $N(\mu, \sigma)$ Gaussian distribution, then the distribution of the random variable $X = \mathrm{e}^Z$ is called a logarithmic Gaussian (normal) distribution. The PDF is

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{\frac{(\log x - \mu)^2}{2\sigma^2}\right\}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance: $\mathbf{E}(X) = \mathrm{e}^{\sigma^2/2+\mu}, \quad \mathbf{D}^2(X) = \mathrm{e}^{\sigma^2/2+\mu}\left(\mathrm{e}^{\mathrm{e}^2} - 1\right).$

Weibull Distribution The Weibull distribution is a generalization of the exponential distribution for which the behavior of the tail distribution is modified by a positive constant k as follows:

$$F(x) = \begin{cases} 1 - \mathrm{e}^{-(x/\lambda)^k}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0; \end{cases}$$

$$f(x) = \begin{cases} \left(\frac{k}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{k-1} \mathrm{e}^{-(x/\lambda)^k}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Expected value and variance:

$$\mathbf{E}(X) = \lambda\Gamma(1 + 1/k), \quad \mathbf{D}^2(X) = \lambda^2\left(\Gamma(1 + 2/k) - \Gamma^2(1 + 1/k)\right).$$

Pareto Distribution Let c and λ be positive numbers. The density function and the PDF of a Pareto distribution are defined as follows:

$$F(x) = \begin{cases} 1 - (\frac{x}{c})^{-\lambda}, & \text{if } x > c, \\ 0, & \text{if } x \le 0; \end{cases}$$

$$f(x) = \begin{cases} \left(\frac{\lambda}{c}\right)\left(\frac{x}{c}\right)^{-\lambda-1} & \text{if } x > c, \\ 0, & \text{if } x \le c. \end{cases}$$

Since the PDF of the Pareto distribution is a simple power function in consequence of this property, it tends to zero with polynomial order as x goes to infinity and the nth moment exists if and only if $n < \lambda$.
Expected value (if $k > 1$) and variance (if $k > 2$):

$$\mathbf{E}(X) = \frac{ck}{k-1}, \quad \mathbf{D}(X) = \frac{c^2 k}{(k-1)^2(k-2)}.$$

1.3 Limit Theorems

1.3.1 Convergence Notions

There are many convergence notions in the theory of analysis, for example, pointwise convergence, uniform convergence, and convergences defined by various metrics. In the theory of probability, several kinds of convergences are also used that are related to the sequences of random variables or to their sequence of distribution functions. The following notion is the so-called weak convergence of distribution functions.

Definition 1.54. The sequence of distribution functions $F_n,\ n = 1, 2, \ldots$ **weakly converges** to the distribution function F (abbreviated $F_n \xrightarrow{w} F,\ n \to \infty$) if the convergence $F_n(x) \to F(x),\ n \to \infty$, holds in all continuity points of F.

If the distribution function F is continuous, then the convergence $F_n \xrightarrow{w} F,\ n \to \infty$ holds if and only if $F_n(x) \to F(x),\ n \to \infty$ for all $x \in \mathbb{R}$. The weak convergence of the sequence $F_n,\ n = 1, 2, \ldots$ is equivalent to the condition that the convergence

$$\int_{-\infty}^{\infty} g(x)\mathrm{d}F_n(x) \to \int_{-\infty}^{\infty} g(x)\mathrm{d}F(x)$$

is true for all bounded and continuous functions g.

In addition, the weak convergence of a distribution function can be given with the help of an appropriate metric in the space $\mathbb{F} = \{F\}$ of all distribution functions.

Let G and H be two distribution functions (i.e., $G, H \in \mathbb{F}$), and define the **Levy metric** [96] as follows:

$$L(G,H) = \inf\{\varepsilon :\ G(x) \le H(x+\varepsilon) + \varepsilon,\ H(x) \le G(x+\varepsilon) + \varepsilon, \text{ for all } x \in \mathbb{R}\}.$$

Then it can be proved that the weak convergence $F_n \overset{w}{\to} F,\ n \to \infty$, of the distribution functions $F,\ F_n,\ n = 1, 2, \ldots$, holds if and only if $\lim\limits_{n\to\infty} L(F_n, F) = 0$.

The most frequently used convergence notions in probability theory for a sequence of random variables are the convergence in distribution, convergence in probability, convergence with probability 1, or almost surely (a.s.), and convergence in mean square (convergence in L_2), which will be introduced subsequently. In cases of the last three convergences, it is assumed that the random variables are defined on a common probability space $(\Omega, \mathcal{A}, \mathbf{P}())$.

Definition 1.55. The sequence of random variables $X_1, X_2, \ldots$ **converges in distribution** to a random variable X (abbreviated $X_n \overset{d}{\to} X,\ n \to \infty$) if their distribution functions satisfy the weak convergence

$$F_{X_n} \overset{w}{\to} F_X,\ n = 1, 2, \ldots .$$

Definition 1.56. The sequence of random variables $X_1, X_2, \ldots$ **converges in probability** to a random variable X ($X_n \overset{P}{\to} X,\ n \to \infty$) if the convergence

$$\lim_{n\to\infty} \mathbf{P}(|X_n - X| > \varepsilon) = 0$$

holds for all positive constants ε.

Definition 1.57. The random variables $X_1, X_2, \ldots$ **converge with probability** 1 (or **almost surely**) to a random variable X (abbreviated $X_n \overset{\text{a.s.}}{\to} X,\ n \to \infty$) if the condition

$$\mathbf{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$$

holds.

The limit $\lim\limits_{n\to\infty} X_n = X$ exists if there are defined random variables with probability 1 $X' = \limsup\limits_{n\to\infty} X_n$ and $X''(\omega) = \liminf\limits_{n\to\infty} X_n$ for which the relation

$$\mathbf{P}\left(X'(\omega) = X''(\omega) = X(\omega)\right) = 1$$

is true. This means that there is an event $A \in \mathcal{A}$, $\mathbf{P}(A) = 0$, such that the equality

$$X'(\omega) = X''(\omega) = X(\omega),\ \ \omega \in \Omega \setminus A$$

holds.

Theorem 1.58 (*[84]*). *The convergence* $\lim_{n\to\infty} X_n = X$ *with probability 1 is true if and only if for all* $\varepsilon > 0$

$$\boldsymbol{P}\left(\sup_{k\geq n}|X_k - X| > \varepsilon\right) = 0.$$

Definition 1.59. Let $X_n,\ n \geq 1$ and X be random variables with finite variance. The sequence $X_1, X_2, \ldots$ **converges in mean square** to random variable X (abbreviated $X_n \xrightarrow{L_2} X,\ n \to \infty$) if

$$\mathbf{E}\left(|X_n - X|^2\right) \to 0,\ n \to \infty.$$

This type of convergence is often called an L_2 convergence of random variables.

The enumerated convergence notions are not equivalent to each other, but we can mention several connections between them. The convergence in distribution follows from all the others. The convergence in probability follows from the convergence with probability 1 and from the convergence in mean square. It can be proved that if the sequence $X_1, X_2, \ldots$ is convergent in probability to the random variable X, then there exists a subsequence $X_{n_1}, X_{n_2}, \ldots$ such that it converges with probability 1 to random variable X.

1.3.2 Laws of Large Numbers

The intuitive introduction of probability implicitly uses the limit behavior of the average

$$\overline{S}_n = \frac{X_1 + \ldots + X_n}{n},\ n = 1, 2, \ldots,$$

of independent identically distributed random variables $X_1, X_2, \ldots$. The main question is: under what condition does the sequence $\overline{S}_n$ converge to a constant μ in probability (weak law of large numbers) or with probability 1 (strong law of large numbers) as n goes to infinity?

Consider an experiment in which we observe that an event A occurs or not. Repeating the experiment n times independently, define the frequency of event A by $S_n(A)$ and the relative frequency by $\overline{S}_n(A)$.

Theorem 1.60 (*Bernoulli*). *The relative frequency of an event A tends in probability to the probability of the event* $p = \mathbf{P}(A)$, *that is, for all* $\varepsilon > 0$ *the relation*

$$\lim_{n\to\infty} \boldsymbol{P}\left(\left|\overline{S}_n(A) - p\right| > \varepsilon\right) = 0$$

holds.

If we introduce the notation

$$X_i = \begin{cases} 1, \text{ if the } i\text{-th outcome in } A, \\ 0, \text{ otherwise,} \end{cases}$$

then the assertion of the last theorem can be formulated as follows:

$$\overline{S}_n = \frac{X_1 + \ldots + X_n}{n} \xrightarrow{p} p,\ n \to \infty,$$

which is a simple consequence of the Chebyshev inequality because the X_i are independent and identically distributed and $\mathbf{E}(X_i) = p = \mathbf{P}(A)$, $\mathbf{D}^2(X_i) = p(1-p)$, $i = 1, 2, \ldots$. This result can be generalized without any difficulties as follows.

Theorem 1.61. *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with common expected value μ and finite variance σ^2. Then the convergence in probability*

$$\overline{S}_n = \frac{X_1 + \ldots + X_n}{n} \xrightarrow{p} \mu,\ n \to \infty,$$

is true.

Proof. Example 1.38, which is given after the proof of the Chebyshev inequality, shows that for all $\varepsilon > 0$ the inequality

$$\mathbf{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

is valid. From this the convergence in probability $\overline{S}_n \xrightarrow{p} \mu$, $n \to \infty$ follows. It is not difficult to see that the convergence in L_2 is also true, i.e., $\overline{S}_n \xrightarrow{L_2} \mu$, $n \to \infty$. □

It should be noted that the inequality $\mathbf{P}\left(\left|\frac{X_1+\ldots+X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$, which guarantees the convergence in probability, gives an upper bound for the probability $\mathbf{P}\left(\left|\frac{X_1+\ldots+X_n}{n} - \mu\right| \geq \varepsilon\right)$ also.

The Kolmogorov strong law of large numbers gives a necessary and sufficient condition for convergence with probability 1.

Theorem 1.62 (*Kolmogorov*). *If the sequence of random variables $X_1, X_2, \ldots$ is independent and identically distributed, then the convergence*

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{a.s.} \mu,\ n \to \infty$$

holds for a constant μ if and only if the random variables X_i have finite expected value and $\boldsymbol{E}(X_i) = \mu$.

Corollary 1.63. *If $\overline{S}_n(A)$ defines the relative frequency of an event A occurring in n independent experiments, then the Bernoulli law of large numbers*

$$\overline{S}_n(A) \xrightarrow{p} p = \boldsymbol{P}(A),\ n \to \infty,$$

is valid. By the Kolmogorov law of large numbers, this convergence is true with probability 1 *also, that is,*

$$\overline{S}_n(A) \xrightarrow{a.s.} p = \boldsymbol{P}(A),\ n \to \infty.$$

1.3.3 Central Limit Theorem, Lindeberg–Feller Theorem

The basic problem of central limit theorems is as follows. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with a common distribution function $F_X(x)$. The question is, under what conditions does a sequence of constants μ_n and σ_n, $\sigma_n \neq 0$, $n = 1, 2, \ldots$ exist such that the sequence of centered and linearly normed sums

$$\overline{S}_n = \frac{X_1 + \ldots + X_n - \mu_n}{\sigma_n},\ n = 1, 2, \ldots \tag{1.11}$$

converges in the distributions

$$F_{\overline{S}_n} \xrightarrow{w} F,\ n \to \infty$$

and have a nondegenerate limit distribution function F? A distribution function $F(x)$ is nondegenerate if there is no point $x_0 \in \mathbb{R}$ satisfying the condition $F(x_0) - F(x_0-) = 1$, that is, the distribution does not concentrate at one point.

Theorem 1.64. *If the random variables $X_1, X_2, \ldots$ are independent and identically distributed with finite expected value $\mu = \boldsymbol{E}(X_1)$ and variance $\sigma^2 = \mathbf{D}^2(X_1)$, then*

$$\boldsymbol{P}\left(\frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}\sigma} \leq x\right) \to \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u^2/2} \mathrm{d}u$$

holds for all $x \in \mathbb{R}$, where the function $\Phi(x)$ denotes the distribution function of standard normal random variables.

If the random variables $X_1, X_2, \ldots$ are independent but not necessarily identically distributed, then a general, so-called Lindeberg–Feller theorem is valid.

Theorem 1.65. *Let $X_1, X_2, \ldots$ be independent random variables whose variances are finite. Denote*

$$\mu_n = \boldsymbol{E}(X_1) + \ldots + \boldsymbol{E}(X_n), \quad \sigma_n = \sqrt{\boldsymbol{D}^2(X_1) + \ldots + \boldsymbol{D}^2(X_n)}, \quad n = 1, 2, \ldots .$$

The limit

$$\boldsymbol{P}\left(\frac{X_1 + \ldots + X_n - \mu_n}{\sigma_n} \le x\right) \to \Phi(x), \quad n \to \infty,$$

is true for all $x \in \mathbb{R}$ if and only if the Lindeberg–Feller condition holds:

$$\lim_{n\to\infty} \max_{1\le j\le n} \frac{1}{\sigma_j^2} \boldsymbol{E}\left(X_j^2 \mathcal{I}_{\{|X_j| > \varepsilon\sigma_n\}}\right) = 0, \ x \in \mathbb{R}, \ \varepsilon > 0,$$

where $\mathcal{I}_{\{\}}$ denotes the indicator variable.

1.3.4 Infinitely Divisible Distributions and Convergence to the Poisson Distribution

There are many practical problems for which model (1.11) and results related to it are not satisfactory. The reason is that the class of possible limit distributions is insufficiently large; for instance, it does not consist of discrete distributions. An example of this is a Poisson distribution, which is an often-used distribution in queueing theory.

As a generalization of model (1.11), consider the sequence of series of random variables (sometimes called a sequence of random variables of triangular arrays)

$$\{X_{n,1}, \ldots, X_{n,k_n}\}, \ n = 1, 2, \ldots, \ k_n \to \infty,$$

satisfying the following conditions for all fixed positive integers n:

1. The random variables $X_{n,1}, \ldots, X_{n,k_n}$ are independent.
2. The random variables $X_{n,1}, \ldots, X_{n,k_n}$ are **infinitesimal** (in other words, **asymptotically negligible**) if the limit for all $\varepsilon > 0$

$$\lim_{n\to\infty} \max_{1\le j\le k_n} \mathbf{P}\left(\left|X_{n,j}\right| > \varepsilon\right) = 0$$

holds.

Considering the sums of series of random variables

$$S_n = X_{n,1} + \ldots + X_{n,k_n}, \ n = 1, 2, \ldots,$$

the class of possible limit distributions (so-called infinitely divisible distributions) is already a sufficiently large class containing, for example, a Poisson distribution.

Definition 1.66. A random variable X is called **infinitely divisible** if it can be given in the form

$$X \stackrel{d}{=} X_{n,1} + \ldots + X_{n,n}$$

for every $n = 1, 2, \ldots$, where the random variables $X_{n,1}, \ldots, X_{n,n}$ are independent and identically distributed.

Infinitely divisible distributions (to which, for example, the normal and Poisson distributions belong) can be given with the help of their characteristic functions.

Theorem 1.67. *If random variable X is infinitely divisible, then its characteristic function has the form (**Lévy–Khinchin canonical form**)*

$$\log f(t) = i\mu t - \frac{\sigma^2}{2}t^2 + \int_{-\infty}^{0} \left(\mathrm{e}^{itx} - 1 - \frac{itx}{1+x^2}\right) \mathrm{d}L(x)$$

$$+ \int_{0}^{\infty} \left(\mathrm{e}^{itx} - 1 - \frac{itx}{1+x^2}\right) \mathrm{d}R(x),$$

where the functions L and R satisfy the following conditions:

(a) μ and σ $(\sigma \geq 0)$ are real constants.
(b) $L(x)$, $x \in (-\infty, 0)$ and $R(x)$, $x \in (0, \infty)$ are monotonically increasing functions on the intervals $(-\infty, 0)$ and $(0, \infty)$, respectively.
(c) $L(-\infty) = R(\infty) = 0$ and the inequality condition

$$\int_{-\infty}^{0} x^2 \mathrm{d}L(x) + \int_{0}^{\infty} x^2 \mathrm{d}R(x) < \infty$$

holds.

If an infinitely divisible distribution has finite variation, then its characteristic function can be given in a more simple form (**Kolmogorov formula**):

$$\log f(t) = i\mu t + \int_{-\infty}^{\infty} \left(\mathrm{e}^{itx} - 1 - itx\right) \frac{1}{x^2} \mathrm{d}K(x),$$

where μ is a constant and $K(x)$ $(K(-\infty) = 0)$ is a monotonically nondecreasing function.

As special cases of the Kolmogorov formula, we get the normal and Poisson distributions.

(a) An infinitely divisible distribution is normal with the parameters (μ,σ) if the function $K(x)$ is defined as

$$K(x) = \begin{cases} 0, & \text{if } x \le 0, \\ \sigma^2, & \text{if } x > 0. \end{cases}$$

Then the characteristic function is

$$f(t) = i\mu t - \frac{\sigma^2}{2}t^2.$$

(b) An infinitely divisible distribution is Poisson with the parameter λ ($\lambda > 0$) if $\mu = \lambda$ and the function $K(x)$ is defined as

$$K(x) = \begin{cases} 0, & \text{if } x \le 1, \\ \lambda, & \text{if } x > 1. \end{cases}$$

In this case the characteristic function can be given as follows:

$$f(t) = i\mu t + \int_{-\infty}^{\infty} \left(e^{itx} - 1 - itx\right) \frac{1}{x^2} dK(x) = \lambda(e^{it} - 1).$$

The following theorem gives an answer to the question of the conditions under which the limit distribution of sums of independent infinitesimal random variables is Poisson. This result will be used later when considering sums of independent arrival processes of queues.

Theorem 1.68 (*Gnedenko, Marczinkiewicz*). *Let* $\{X_{1,n}, \ldots, X_{k_n,n}\}$, $n = 1, 2, \ldots$, *be a sequence of series of independent infinitesimal random variables. The sequence of distributions of sums*

$$X_n = X_{n1} + \ldots + X_{n,k_n}, \quad n \ge 1,$$

converges weakly to a Poisson distribution with the parameter λ ($\lambda > 0$) *as* $n \to \infty$ *if and only if the following conditions hold for all* ε ($0 < \varepsilon < 1$)*:*

(A) $$\sum_{j=1}^{k_n} \int_{\mathbb{R}_\varepsilon} dF_{nj}(x) \to 0.$$

(B) $$\sum_{j=1}^{k_n} \int_{|x-1|<\varepsilon} dF_{nj}(x) \to \lambda.$$

(C) $\sum_{j=1}^{k_n} \int\limits_{|x|<\varepsilon} \mathrm{d}F_{nj}(x) \to 0.$

(D) $\sum_{j=1}^{k_n} \left[\int\limits_{|x|<\varepsilon} x^2 \mathrm{d}F_{nj}(x) - \left(\int\limits_{|x|<\varepsilon} x \mathrm{d}F_{nj}(x) \right)^2 \right] \to 0,$

where $F_{nj}(x) = \boldsymbol{P}\left(X_{nj} \le x\right)$ *and* $\mathbb{R}_\varepsilon = \mathbb{R} \setminus (\{|x| < \varepsilon\} \cup \{|x-1| < \varepsilon\})$.

Note that conditions (A) and (B) guarantee the convergence of the Poisson part to the appropriate Poisson distribution of the limit, (C) means that there is no centralization, and from (D) it follows that the limit distribution does not contain a Gaussian part.

1.4 Exercises

Exercise 1.1. Let X be a nonnegative random variable with CDF F_X. Given $0 \le t \le X$ $[\mathbf{P}(X > t) \ne 0]$, find the CDF of residual lifetime X.

Exercise 1.2. Let X and Y be independent random variables with a Poisson distribution of parameters λ and μ, respectively. Verify that

(a) The sum $X + Y$ has a Poisson distribution with the parameter $\lambda + \mu$;
(b) For any nonnegative integers $m \le n$ the conditional distribution $\mathbf{P}(X = m \mid X + Y = n)$ is binomial with the parameter $(n, \frac{\lambda}{\lambda+\mu})$, i.e.,

$$\mathbf{P}(X = m \mid X + Y = n) = \binom{m}{n} \left(\frac{\lambda}{\lambda+\mu}\right)^m \left(1 - \frac{\lambda}{\lambda+\mu}\right)^{n-m}.$$

Exercise 1.3. Let X and Y be independent random variables having a uniform distribution on the interval $(0, 1)$ and an exponential distribution with the parameter 1, respectively. Find the probability (concrete number) that $X < Y$.

Exercise 1.4. Divide the interval $(0, 1)$ into three parts with two independently and randomly chosen points U_1 and U_2 of the interval $(0, 1)$. Find the probability of event A that the three parts can determine a triangle.

Exercise 1.5. Show that for a nonnegative random variable X with a finite nth ($n \ge 1$) moment it is true that $\mathbf{E}(X^n) = \int\limits_0^\infty \mathbf{P}(x < X) n x^{n-1} \mathrm{d}x$.

Exercise 1.6. Let X and Y be independent random variables with a uniform distribution on the interval $(0, 1)$. Find the quantities

(a) $\mathbf{E}(|X - Y|)$, $\mathbf{D}^2(|X - Y|)$,
(b) $\mathbf{P}(|X - Y|) > \frac{1}{2}$.

Exercise 1.7. Let X and Y be independent random variables having an exponential distribution with the parameters λ and μ, respectively.

(a) Determine the density function of the random variable $Z = X + Y$.
(b) Find the density function of the random variable $W = \min(X, Y)$.

Exercise 1.8. Let $X_1, \ldots, X_n$ be independent random variables having an exponential distribution with the parameter λ.
Find the expected values of the random variables $V_n = \max(X_1, \ldots, X_n)$, and $W_n = \min(X_1, \ldots, X_n)$.

Exercise 1.9. Let X and Y be independent random variables with density functions $f_X(x)$ and $f_Y(x)$, respectively. Determine the conditional expected value $E(X \mid X < Y)$.

Exercise 1.10. Determine the conditional expectations $\mathbf{E}(X \,|Y = y)$ and $\mathbf{E}(X \,|Y)$ if the joint PDF of the random variables X and Y has the form

(a) $f_{X,Y}(x, y) = \begin{cases} 2, \text{ if } 0 < x, y \text{ and } x + y < 1, \\ 0, \text{ otherwise;} \end{cases}$

(b) $f_{X,Y}(x, y) = \begin{cases} 3(x + y), \text{ if } 0 < x, y \text{ and } x + y < 1, \\ 0, \qquad\quad \text{otherwise.} \end{cases}$

Exercise 1.11. Let $X_1, X_2, \ldots$ be independent random variables with an exponential distribution of the parameter λ. Let N be a geometrically distributed random variable with the parameter p [$p_k = \mathbf{P}(N = k) = p(1 - p)^k$, $k = 1, 2, \ldots$], which does not depend on random variables $(X_1, X_2, \ldots)$. Prove that the sum $Y = X_1 + \ldots + X_N$ has an exponential distribution with the parameter $p\lambda$.

Exercise 1.12. Consider the distribution function of the sum Y_{40} of independent random variables $X_1, \ldots, X_{40}$ having an exponential distribution with the parameter 1. Give an estimate for the probability $p = \mathbf{P}\left(\frac{|Y_{40} - \mathbf{E}(Y_{40})|}{\mathbf{D}(Y_{40})} > 0.05\right)$ calculated with the help of the central limit theorem. We can numerically calculate this probability because the random variable Y_{40} has a gamma distribution with the parameter $(40, 1)$. Using this fact, what result can we obtain for the considered probability? (On the numerical calculation of the gamma distribution see, for example, [72] or [63].)

Chapter 2
Introduction to Stochastic Processes

2.1 Stochastic Processes

When considering technical, economic, ecological, or other problems, in several cases the quantities $\{X_t,\ t \in \mathcal{T}\}$ being examined can be regarded as a collection of random variables. This collection describes the changes (usually in time and in space) of considered quantities. If the set $\mathcal{T}$ is a subset of the set of real numbers, then the set $\{t \in \mathcal{T}\}$ can be interpreted as time and we can say that the random quantities X_t vary in time. In this case the collection of random variables $\{X_t,\ t \in \mathcal{T}\}$ is called a **stochastic process**. In mathematical modeling of randomly varying quantities in time, one might rely on the highly developed theory of stochastic processes.

Definition 2.1. Let $\mathcal{T} \subset \mathbb{R}$. A **stochastic process** X is defined as a collection $X = \{X_t,\ t \in \mathcal{T}\}$ of indexed random variables X_t, which are given on the same probability space $(\Omega, \mathcal{A}, \mathbf{P}())$.

Depending on the notational complexity of the parameter, we occasionally interchange the notation X_t with $X(t)$.

It is clear that $X_t = X_t(\omega)$ is a function of two variables. For fixed $t \in \mathcal{T}$, X_t is a random variable, and for fixed $\omega \in \Omega$, X_t is a function of the variable $t \in \mathcal{T}$, which is called a **sample path** of the stochastic process.

Depending on the set $\mathcal{T}$, X is called a **discrete-time** stochastic process if the index set $\mathcal{T}$ consists of consecutive integers, for example, $\mathcal{T} = \{0, 1, \ldots\}$ or $\mathcal{T} = \{\ldots, -1, 0, 1, \ldots\}$. Further, X is called a **continuous-time** stochastic process if $\mathcal{T}$ equals an interval of the real line, for example, $\mathcal{T} = [a, b]$, $\mathcal{T} = [0, \infty)$ or $\mathcal{T} = (-\infty, \infty)$.

Note that in the case of discrete time, X is a sequence $\{X_n,\ n \in \mathcal{T}\}$ of random variables, while it determines a random function in the continuous-time case. It should be noted that similarly to the notion of real-valued stochastic processes, we may define complex or vector valued stochastic processes also if X_t take values in a complex plane or in higher-dimensional Euclidean space.

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, 55
DOI 10.1007/978-1-4614-5317-8_2, © Springer Science+Business Media, LLC 2013

2.2 Finite-Dimensional Distributions of Stochastic Processes

A stochastic process $\{X_t,\ t \in \mathcal{T}\}$ can be characterized in a statistical sense by its finite-dimensional distributions.

Definition 2.2. The **finite-dimensional distributions** of a stochastic process $\{X_t,\ t \in \mathcal{T}\}$ are defined by the family of all joint distribution functions

$$F_{t_1,\ldots,t_n}(x_1,\ldots,x_n) = \mathbf{P}\,(X_{t_1} < x_1,\ldots,X_{t_n} < x_n),$$

where $n = 1,2,\ldots$ and $t_1,\ldots,t_n \in \mathcal{T}$.

The family of introduced distribution functions

$$\mathcal{F} = \{F_{t_1,\ldots,t_n},\ \ t_1,\ldots,t_n \in \mathcal{T},\ n = 1,2,\ldots\}$$

satisfies the following, specified consistency conditions:

(a) For all positive integers n, m and indices $t_1,\ldots,t_{n+m} \in \mathcal{T}$

$$\lim_{x_{n+1}\to\infty} \ldots \lim_{x_{n+m}\to\infty} F_{t_1,\ldots,t_n,t_{n+1},\ldots,t_{n+m}}(x_1,\ldots,x_n,x_{n+1},\ldots,x_{n+m})$$
$$= F_{t_1,\ldots,t_n}(x_1,\ldots,x_n),\ x_1,\ldots,x_n \in R\ .$$

(b) For all permutations $(i_1,\ldots,i_n)$ of the numbers $\{1,2,\ldots,n\}$

$$F_{s_1,\ldots,s_n}(x_{i_1},\ldots,x_{i_n}) = F_{t_1,\ldots,t_n}(x_1,\ldots,x_n),\ x_1,\ldots,x_n \in R\ ,$$

where $s_j = t_{i_j},\ j = 1,\ldots,n$.

Definition 2.3. If the family $\mathcal{F}$ of joint distribution functions defined previously satisfies conditions (a) and (b), then we say that $\mathcal{F}$ satisfies the **consistency conditions**.

The following theorem is a basic one in probability theory and ensures the existence of a stochastic process (in general of a collection of random variables) with given finite-dimensional distribution functions satisfying the consistency conditions.

Theorem 2.4 (Kolmogorov consistency theorem). *Suppose a family of distribution functions* $\mathcal{F} = \{F_{t_1,\ldots,t_n},\ t_1,\ldots,t_n \in \mathcal{T},\ n = 1,2,\ldots\}$ *satisfies the consistency conditions (a) and (b). Then there exists a probability space* $(\Omega, \mathcal{A}, \mathbf{P})$*, and on that a stochastic process* $\{X_t,\ t \in \mathcal{T}\}$*, whose finite-dimensional distributions are identical to* $\mathcal{F}$.

For our considerations, it usually suffices to provide the finite-dimensional distribution functions of the stochastic processes, in which case the process is

defined in a **weak sense** and it is irrelevant on which probability space it is given. In some instances the behavior of the random path is significant (e.g., continuity in time), which is related to a given probability space $(\Omega, \mathcal{A}, P)$ where the process $\{X_t,\ t \in \mathcal{T}\}$ is defined. In this case the process is given in a **strict sense**.

2.3 Stationary Processes

The class of stochastic processes that show a stationary statistical property in time plays a significant role in practice. Among these processes the most important ones are the stationary processes in strict and weak senses. The main notions are given here for one-dimensional processes, but the notion for high-dimensional processes can be introduced similarly.

Definition 2.5. A process $\{X_t,\ t \in \mathcal{T}\}$ is called **stationary in a strict sense** if the joint distribution functions of random variables

$$(X_{t_1}, \ldots, X_{t_n}) \quad and \quad (X_{t_1+t}, \ldots, X_{t_n+t})$$

are identical for all t, positive integer n, and $t_1, \ldots, t_n \in \mathcal{T}$ satisfying the conditions $t_i + t \in \mathcal{T},\ i = 1, \ldots, n$.

Note that this definition remains valid in the case of vector-valued stochastic processes. Consider a stochastic process X with finite second moment, that is, $\mathbf{E}(X_t^2) < \infty$, for all $t \in \mathcal{T}$. Denote the expected value and covariance functions by

$$\begin{aligned} \mu_X(t) &= \mathbf{E}\,(X_t)\,,\ t \in \mathcal{T}, \\ R_X(s,t) &= \operatorname{cov}(X_s, X_t) \\ &= \mathbf{E}\,((X_t - \mu_X(t))(X_s - \mu_X(s)))\,,\ s,t \in \mathcal{T}. \end{aligned}$$

Definition 2.6. A process $\{X_t,\ t \in \mathcal{T}\}$ is called **stationary in a weak sense** if X_t has finite second moment for all $t \in \mathcal{T}$ and the expected value and covariance function satisfy the following relation:

$$\begin{aligned} \mu_X(t) &= \mu_X,\ t \in \mathcal{T}, \\ R_X(s,t) &= R_X(t-s),\ s,t \in \mathcal{T}. \end{aligned}$$

The function R_X is called the **covariance function**.

It is clear that if a stochastic process with finite second moment is stationary in a strict sense, then it is stationary in a weak sense also, because the expected value and covariance function depend also on the two-dimensional joint distribution,

which is time-invariant if the time shifts. Besides the covariance function $R_X(t)$, the **correlation function** $r_X(t)$ is also used, which is defined as follows:

$$r_X(t) = \frac{1}{R_X(0)} R_X(t) = \frac{1}{\sigma_X^2} R_X(t).$$

2.4 Gaussian Process

In practice, we often encounter stochastic processes whose finite-dimensional distributions are Gaussian. These stochastic processes are called **Gaussian**. In queueing theory Gaussian processes often appear when asymptotic methods are applied.

Note that the expected values and covariances determine the finite-dimensional distributions of the Gaussian process; therefore, it is easy to verify that a Gaussian process is stationary in a strict sense if and only if it is stationary in a weak sense. We also mention here that the discrete-time Gaussian process consists of independent Gaussian random variables if these random variables are uncorrelated.

2.5 Stochastic Process with Independent and Stationary Increments

In several practical modeling problems, stochastic processes have independent and stationary increments. These processes play a significant role both in theory and practice. Among such processes the Wiener and the Poisson processes are defined below.

Definition 2.7. If for any integer $n \geq 1$ and parameters $t_0, \ldots, t_n \in \mathcal{T},\ t_0 < \ldots < t_n$, the increments

$$X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}}$$

of a stochastic process $X = \{X_t,\ t \in \mathcal{T}\}$ are independent random variables, then X is called a stochastic process with **independent increments**. The process X has **stationary increments** if the distribution of $X_{t+h} - X_t,\ t, t+h \in \mathcal{T}$ does not depend on t.

2.6 Wiener Process

As a special but important case of stochastic processes with independent and stationary increments, we mention here the **Wiener process** (also called **process of Brownian motion**), which gives the mathematical model of diffusion. A process

$X = \{X_t,\ t \in [0,\infty)\}$ is called a Wiener process if the increments of the process are independent and for any positive integer n and $0 \le t_0 < \ldots < t_n$ the joint density function of random variables $X_{t_0}, \ldots, X_{t_n}$ can be given in the form

$$f(x_0,\ldots,x_n;t_0,\ldots,t_n) = (2\pi)^{-n/2}\left[t_0(t_1-t_0)\ldots(t_n-t_{n-1})\right]^{-1/2} \times \exp\left\{-\frac{1}{2}\left(\frac{x_0^2}{t_0}+\frac{(x_1-x_0)^2}{t_1-t_0}+\ldots+\frac{(x_n-x_{n-1})^2}{t_n-t_{n-1}}\right)\right\}.$$

It can be seen from this formula that the Wiener process is Gaussian and the increments

$$X_{t_j} - X_{t_{j-1}},\ j = 1,\ldots,n,$$

are independent Gaussian random variables with expected values 0 and variances $t_j - t_{j-1}$. The expected value function and the covariance function are determined as

$$\mu_X(t) = 0, \quad R_X(s,t) = \min(t,s), \quad t,s \ge 0.$$

2.7 Poisson Process

2.7.1 Definition of Poisson Process

Besides the Wiener process defined above, we discuss in this chapter another important process with independent increments in probability theory, the Poisson process. This process plays a fundamental role not only in the field of queueing theory but in many areas of theoretical and applied sciences, and we will deal with this process later as a Markov arrival process, birth-and-death process, and renewal process. Its significance in probability theory and practice is that it can be used to model different event occurrences in time and space in, for example, queueing systems, physics, insurance, population biology. There are several introductions and equivalent definitions of the Poisson process in the literature according to its different characterizations. First we present the notion in the simple (classical) form and after that in a more general context.

In queueing theory, a frequently used model for the description of the arrival process of costumers is as follows. Assume that costumers arrive at the system one after another at $t_1 < t_2 < \ldots$; $t_n \to \infty$ as $n \to \infty$. The differences in occurrence times, called **interarrival times**, are denoted by

$$X_1 = t_1,\ X_2 = t_2 - t_1, \ldots,\ X_n = t_n - t_{n-1}, \ldots .$$

Define the process $\{N(t),\ t \ge 0\}$ with $N(0) = 0$ and

$$N(t) = \max\{n :\ t_n \le t\} = \max\{n :\ X_1 + \ldots + X_n \le t\},\ t > 0.$$

This process counts the number of customers arriving at the system in the time interval $(0, t]$ and is called the **counting process** for the sequence $t_1 < t_2 < \ldots$. Obviously, the process takes nonnegative integer values only, is nondecreasing, and $N(t) - N(s)$ equals the number of occurrences in the time interval $(s, t]$ for all $0 < s < t$.

In the special case, when $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with exponential distribution $\mathrm{Exp}(\lambda)$, the increments $N(t) - N(s)$ have a Poisson distribution with the parameter $\lambda(t - s)$. In addition, the counting process $N(t)$ possesses an essential property, that is, it evolves in time **without aftereffects**. This means that the past and current occurrences have no effect on subsequent occurrences. This feature leads to the property of independent increments.

Definition 2.8. We say that the process $N(t)$ is a **Poisson process** with **rate** λ if

1. $N(0) = 0$,
2. $N(t),\ t \geq 0$ is a process with independent increments,
3. The distribution of increments is Poisson with the parameter $\lambda(t - s)$ for all $0 < s < t$.

By definition, the distributions of the increments $N(t + h) - N(t)$, $t \geq 0$, $h > 0$, do not depend on the moment t; therefore, it is a process with stationary increments and is called a **homogeneous** Poisson process at rate λ. Next, we introduce the Poisson process in a more general setting, and as a special case we have the homogeneous case. After that we will deal with the different characterizations of Poisson processes, which in some cases can serve as a definition of the process. At the end of this chapter, we will introduce the notion of the high-dimensional Poisson process (sometimes called a spatial Poisson process) and give its basic properties.

Let $\{\Lambda(t),\ t \geq 0\}$ be a nonnegative, monotonically nondecreasing, continuous-from-right real-valued function for which $\Lambda(0) = 0$.

Definition 2.9. We say that a stochastic process $\{N(t),\ t \geq 0\}$ taking nonnegative integers is a **Poisson process** if

1. $N(0) = 0$,
2. $N(t)$ is a process with independent increments,
3. The CDFs of the increments $N(t) - N(s)$ are Poisson with the parameter $\Lambda(t) - \Lambda(s)$ for all $0 \leq s \leq t$, that is,

$$\mathbf{P}\left(N(t) - N(s) = k\right) = \frac{(\Lambda(t) - \Lambda(s))^k}{k!} \mathrm{e}^{-(\Lambda(t) - \Lambda(s))},\ k = 0, 1, \ldots .$$

Since for any fixed $t > 0$ the distribution of $N(t) = N(t) - N(0)$ is Poisson with mean $\Lambda(t)$, that is the reason that $N(t)$ is called a Poisson process. We can state that the process $N(t)$ is a monotonically nondecreasing jumping process whose increments $N(t) - N(s)$, $0 \leq s < t$, take nonnegative integers only and the increments have Poisson distributions with the parameter $(\Lambda(t) - \Lambda(s))$. Thus the

random variables $N(t),\ t \geq 0$ have Poisson distributions with the parameter $\Lambda(t)$; therefore, the expected value of $N(t)$ is $\mathbf{E}\,(N(t)) = \Lambda(t),\ t \geq 0$, which is called a **mean value function**.

We also note that using the property of independent increments, the joint distribution of the random variables $N(t_1), \ldots, N(t_n)$ can be derived for all positive integers n and all $0 < t_1 < \ldots < t_n$ without difficulty because for any integers $0 \leq k_1 \leq \ldots \leq k_n$ we get

$$\begin{aligned}
&\mathbf{P}\,(N(t_1) = k_1, \ldots, N(t_n) = k_n) \\
&= \mathbf{P}\,(N(t_1) = k_1, N(t_2) - N(t_1) = k_2 - k_1, \ldots, N(t_n) - N(t_{n-1}) = k_n - k_{n-1}) \\
&= \frac{(\Lambda(t_1))^{k_1}}{k_1!} \mathrm{e}^{-\Lambda(t_1)} \prod_{i=2}^{n} \frac{(\Lambda(t_i) - \Lambda(t_i - 1))^{k_i - k_{i-1}}}{(k_i - k_{i-1})!} \mathrm{e}^{-(\Lambda(t_i) - \Lambda(t_{i-1}))}.
\end{aligned}$$

Since the mean value function $\Lambda(t) = \mathbf{E}\,(N(t))$ is monotonically nondecreasing, the set of discontinuity points $\{\tau_n\}$ of $\Lambda(t)$ is finite or countably infinite. It can happen that the set of discontinuity points $\{\tau_n\}$ has more than one convergence point, and in this case we cannot give the points of $\{\tau_n\}$ as an ordered sequence $\tau_1 < \tau_2 < \ldots$. Define the jumps of the function $\Lambda(t)$ at discontinuity points τ_n as follows:

$$\lambda_n = \Lambda(\tau_n + 0) - \Lambda(\tau_n - 0) = \Lambda(\tau_n) - \Lambda(\tau_n - 0).$$

By definition, the increments of a Poisson process are independent; thus it is easy to check that the following decomposition exists:

$$N(t) = N_r(t) + N_s(t),$$

where $N_r(t)$ and $N_s(t)$ are independent Poisson processes with mean value functions

$$\Lambda_r(t) = \Lambda(t) - \sum_{\tau_n < t} \lambda_n \ \text{ and } \ \Lambda_s(t) = \sum_{\tau_n < t} \lambda_n.$$

The **regular** part $N_r(t)$ of $N(t)$ has jumps equal to 1 only, whose mean value function $\Lambda_r(t)$ is continuous. Thus we can state that the process $N_r(t)$ is continuous in probability, that is, for any point t, $0 \leq t < \infty$, the relation

$$\lim_{s \to 0} \mathbf{P}\,(N_r(t+s) - N_r(t) > 0) = \lim_{s \to 0} \mathbf{P}\,(N_r(t+s) - N_r(t) \geq 1) = 0$$

is true. The second part $N_s(t)$ of $N(t)$ is called a **singular** Poisson process because it can have jumps only in discrete points $\{\tau_n\}$. Then

$$\mathbf{P}\,(N_s(\tau_n) - N_s(\tau_n - 0)) = k) = \frac{\lambda_n^k}{k!}\, \mathrm{e}^{-\lambda_n}, k = 0, 1, 2, \ldots .$$

Definition 2.10. If the mean value function $\Lambda(t)$ of a Poisson process $\{N(t),\ t \geq 0\}$ is differentiable with the derivative $\lambda(s),\ s \geq 0$ satisfying $\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$, then the function $\lambda(s)$ is called a **rate** (or **intensity**) **function** of the process.

In accordance with our first definition (2.8), we say that the Poisson process $N(t)$ is **homogeneous** with the rate λ if the rate function is a constant $\lambda(t) = \lambda,\ t \geq 0$. In this case, $\Lambda(t) = \lambda t,\ t \geq 0$ is satisfied; consequently, the distributions of all increments $N(t) - N(s),\ 0 \leq s < t$ are Poisson with the parameter $\lambda(t - s)$ and $\mathbf{E}\,(N(t) - N(s)) = \lambda(t - s)$. This shows that the average number of occurrences is proportional to the length of the corresponding interval and the constant of proportionality is λ. These circumstances justify the name of the rate λ.

If the rate can vary with time, that is, the rate function does not equal a constant, the Poisson process is called **inhomogeneous**.

2.7.2 Construction of Poisson Process

The construction of Poisson processes plays an essential role both from a theoretical and a practical point of view. In particular, it is essential in simulation methods. The Poisson process $N(t)$ and the sequence of the random jumping points $t_1 < t_2 < \ldots$ of the process uniquely determine each other. This fact provides an opportunity to give another definition of the Poisson process on the real number line. We prove that the following two constructions of Poisson processes are valid (see, for example, pp. 117–118 in [85]).

Theorem 2.11 (*Construction I). Let $X_1, X_2, \ldots$ be independent and identically distributed random variables whose common CDF is exponential with parameter* 1. *Define*

$$M(t) = \sum_{m=1}^{\infty} \mathcal{I}_{\{X_1+\ldots+X_m \leq t\}},\ t \geq 0. \tag{2.1}$$

Then the process $M(t)$ is a homogeneous Poisson process with an intensity rate equal to 1.

Theorem 2.12 (*Construction II). Let $U_1, U_2, \ldots$ be a sequence of independent and identically distributed random variables having common uniform distribution on the interval $(0, T)$, and let N be a random variable independent of U_i with a Poisson distribution with the parameter λT. Define*

$$N(t) = \sum_{m=1}^{N} \mathcal{I}_{\{U_m \leq t\}},\ 0 \leq t \leq T. \tag{2.2}$$

Then $N(t)$ is a homogeneous Poisson process on the interval $[0, T]$ at rate λ.

We begin with the proof of Construction II. Then, using this result, we verify Construction I.

Proof (Construction II). Let K be a positive integer and $t_1, \dots, t_K$ positive constants such that $t_0 = 0 < t_1 < t_2 < \dots < t_K = T$. Since, by Eq. (2.2), $N(T) = N$ and $N(t) = \sum_{m=1}^{N} \mathcal{I}_{\{U_m \le t\}}$, the increments of $N(t)$ on the intervals $(t_{k-1}, t_k]$, $k = 1, \dots, K$, can be given in the form

$$N(t_k) - N(t_{k-1}) = \sum_{n=1}^{N} \mathcal{I}_{\{t_{k-1} < U_n \le t_k\}}, \ k = 1, \dots, K.$$

Determine the joint characteristic function of the increments $N(t_k) - N(t_{k-1})$. Let $s_k \in \mathbb{R}$, $k = 1, \dots, K$, be arbitrary; then

$$\begin{aligned}
\varphi(s_1, \dots, s_K) &= \mathbf{E}\left(\exp\left\{\sum_{k=1}^{K} i s_k (N(t_k) - N(t_{k-1}))\right\}\right) \\
&= \mathbf{P}(N = 0) + \sum_{n=1}^{\infty} \mathbf{E}\left(\exp\left\{\sum_{k=1}^{K} i s_k (N(t_k) - N(t_{k-1}))\right\} \,\middle|\, N = n\right) \mathbf{P}(N = n) \\
&= \mathrm{e}^{-\lambda T} + \sum_{n=1}^{\infty} \mathbf{E}\left(\exp\left\{\sum_{k=1}^{K} i s_k \sum_{\ell=1}^{n} \mathcal{I}_{\{t_{k-1} < U_\ell \le t_k\}}\right\}\right) \mathbf{P}(N = n) \\
&= \mathrm{e}^{-\lambda T} + \sum_{n=1}^{\infty} \prod_{\ell=1}^{n} \mathbf{E}\left(\exp\left\{\sum_{k=1}^{K} i s_k \mathcal{I}_{\{t_{k-1} < U_\ell \le t_k\}}\right\}\right) \frac{(\lambda T)^n}{n!} \mathrm{e}^{-\lambda T} \\
&= \mathrm{e}^{-\lambda T} \sum_{n=0}^{\infty} \left(\sum_{k=1}^{K} \frac{t_k - t_{k-1}}{T} \mathrm{e}^{i s_k}\right)^n \frac{(\lambda T)^n}{n!} = \mathrm{e}^{-\lambda T} \exp\left\{\sum_{k=1}^{K} \mathrm{e}^{i s_k} \lambda (t_k - t_{k-1})\right\},
\end{aligned}$$

and using the relation $T = t_K - t_0 = \sum_{k=1}^{K} (t_k - t_{k-1})$ we get

$$\varphi(s_1, \dots, s_K) = \prod_{k=1}^{K} \exp\left\{\lambda (t_k - t_{k-1})(\mathrm{e}^{i s_k} - 1)\right\}.$$

Since the characteristic function $\varphi(s_1, \dots, s_K)$ derived here is equal to the joint characteristic function of independent random variables having a Poisson distribution with the parameters $\lambda(t_k - t_{k-1})$, $k = 1, \dots, K$, the proof is complete. □

For the proof of Construction I we need the following well-known lemma of probability theory.

Lemma 2.13. *Let T be a positive constant and k be a positive integer. Let $U_1, \ldots, U_k$ be independent and identically distributed random variables having a common uniform distribution on the interval $(0, T)$. Define by $U_{1k} \le \ldots \le U_{kk}$ the ordered random variables $U_1, \ldots, U_k$. Then the joint PDF of random variables $U_{1k}, \ldots, U_{kk}$ is*

$$f_{U_{1k},\ldots,U_{kk}}(t_1, \ldots, t_k) = \begin{cases} \frac{k!}{T^k}, \text{ if } 0 < t_1 \le t_2 \le \ldots \le t_k < T, \\ 0, \qquad\qquad \text{otherwise.} \end{cases}$$

Proof. Since $U_{1k} \le \ldots \le U_{kk}$, it is enough to determine the joint PDF of random variables $U_{1k}, \ldots, U_{kk}$ on the set

$$\mathcal{K} = \{(t_1, \ldots, t_k) : \ 0 \le t_1 \le \ldots \le t_k < T\}.$$

Under the assumptions of the lemma, the random variables $U_1, \ldots, U_k$ are independent and uniformly distributed on the interval $(0, T)$; thus for every permutation $i_1, \ldots, i_k$ of the numbers $1, 2, \ldots, k$ (the number of all permutations is equal to $k!$)

$$\begin{aligned} &\mathbf{P}\left(U_{i_1} \le \ldots \le U_{i_k}, U_{i_1} \le t_1, \ldots, U_{i_k} \le t_k\right) \\ &\quad = \mathbf{P}\left(U_1 \le \ldots \le U_k, U_1 \le t_1, \ldots, U_k \le t_k\right), \end{aligned}$$

then

$$\begin{aligned} F_{U_{1k},\ldots,U_{kk}}(t_1, \ldots, t_k) &= \mathbf{P}\left(U_{1k} \le t_1, \ldots, U_{kk} \le t_k\right) \\ &= k!\mathbf{P}\left(U_1 \le \ldots \le U_k, U_1 \le t_1, \ldots, U_k \le t_k\right) \\ &= k! \int_0^{t_1} \ldots \int_0^{t_k} \frac{1}{T^k} \mathcal{I}_{\{u_1 \le \ldots \le u_k\}} \mathrm{d}u_k \ldots \mathrm{d}u_1 \\ &= \frac{k!}{T^k} \int_0^{t_1} \int_{u_1}^{t_2} \ldots \int_{u_{k-1}}^{t_k} \mathrm{d}u_k \ldots \mathrm{d}u_1. \end{aligned}$$

From this we immediately have

$$f_{U_{1k},\ldots,U_{kk}}(t_1, \ldots, t_k) = \frac{k!}{T^k}, \ \ (t_1, \ldots, t_k) \in \mathcal{K},$$

which completes the proof. □

Proof (Construction I). We verify that for any $T > 0$ the process $M(t)$, $0 \le t \le T$ is a homogeneous Poisson process with rate λ. By Construction II, $N(T) = N$,

where the distribution of random variable N is Poisson with the parameter λT. From Eq. (2.2) it follows that the process $N(t),\ 0 \leq t < T$, can be rewritten in the form

$$N(t) = \sum_{m=1}^{N} \mathcal{I}_{\{U_m \leq t\}} = \sum_{n=1}^{N} \mathcal{I}_{\{U_{nN} \leq t\}},$$

where for every $k \geq 1$ and under the condition $N(T) = k$ the random variables $U_1, \ldots U_k$ are independent and uniformly distributed on the interval $(0, T)$ and $U_{1k} \leq U_{2k} \leq \ldots \leq U_{kk}$ are the ordered random variables $U_1, \ldots U_k$. Note that we used these properties only to determine the joint characteristic function of the increments. Define

$$T_n = X_1 + \ldots + X_n,\ \ n = 1, 2, \ldots,$$

where, by assumption, $X_1, X_2, \ldots$ are independent and identically distributed random variables with a common exponential CDF of parameter 1. Then, using the relation (2.1), for any $0 \leq t \leq T$,

$$M(t) = \sum_{n=1}^{\infty} \mathcal{I}_{\{T_n \leq t\}} = \begin{cases} \sum\limits_{n=1}^{M(T)} \mathcal{I}_{\{T_n \leq t\}}, & \text{if } T \geq T_1, \\ 0, & \text{if } T < T_1. \end{cases}$$

By the previous note it is enough to prove that

(a) The random variable $M(T)$ has a Poisson CDF with the parameter λT;
(b) For every positive integer k and under the condition $M(T) = k$, the joint CDF of the random variables $T_1, \ldots, T_n$ are identical with the CDF of the random variables $U_{1k}, \ldots, U_{kk}$.

(a) First we prove that for any positive t the CDF of the random variable $M(t)$ is Poisson with the parameter (λT). Since the common CDF of independent and identically distributed random variables X_i is exponential with the parameter λ, the random variable T_n has a gamma(n, λ) distribution whose PDF (see the description of gamma distribution in Sect. 1.2.2.) is

$$f_{T_n}(x) = \begin{cases} \frac{\lambda^n}{\Gamma(n)} x^{n-1} \mathrm{e}^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases}$$

From the exponential distribution of the first arrival we have

$$\mathbf{P}\left(M(t) = 0\right) = \mathbf{P}\left(X_1 > t\right) = \mathrm{e}^{-\lambda t}.$$

Using the theorem of the total expected value, for every positive integer k we obtain

$$\begin{aligned}
\mathbf{P}\,(M(t)=k) &= \mathbf{P}\,(X_1+\ldots+X_k \le t < X_1+\ldots+X_{k+1}) \\
&= \mathbf{P}\,(T_k \le t < T_k + X_{k+1}) \\
&= \int_0^t \mathbf{P}\,(T_k \le t < T_k + X_{k+1} \,|T_k = z)\frac{\lambda^k}{\Gamma(k)} z^{k-1}\mathrm{e}^{-\lambda z}\mathrm{d}z \\
&= \int_0^t \mathbf{P}\,(t-z < X_{k+1})\frac{\lambda^k}{\Gamma(k)} z^{k-1}\mathrm{e}^{-\lambda z}\mathrm{d}z \\
&= \int_0^t \mathrm{e}^{-\lambda(t-z)}\frac{\lambda^k}{\Gamma(k)} z^{k-1}\mathrm{e}^{-\lambda z}\mathrm{d}z \\
&= \frac{\lambda^k}{\Gamma(k)}\mathrm{e}^{-\lambda t}\int_0^t z^{k-1}\mathrm{d}z = \frac{(\lambda t)^k}{\Gamma(k)k}\mathrm{e}^{-\lambda t} = \frac{(\lambda t)^k}{k!}\mathrm{e}^{-\lambda t};
\end{aligned}$$

thus the random variable $M(t)$, $t \ge 0$ has a Poisson distribution with the parameter λt.

(b) Let T be a fixed positive number and let $U_1, \ldots U_k$ be independent random variables uniformly distributed on the interval $(0,1)$. Denote by $U_{1k} \le \ldots \le U_{kk}$ the ordered random variables $U_1, \ldots U_k$. Now we verify that for every positive integer k the joint CDF of random variables $T_1, \ldots, T_k$ under the condition $M(T) = k$ is identical with the joint CDF of the ordered random variables $U_{1k}, \ldots, U_{kk}$ (see Theorem 2.3 of Ch. 4. in [48]).

For any positive numbers $t_1, \ldots, t_k$, the joint conditional CDF of random variables $T_1, \ldots, T_k$ given $M(t) = k$ can be written in the form

$$\mathbf{P}\,(T_1 \le t_1, \ldots, T_k \le t_k \,|M(T)=k) = \frac{\mathbf{P}\,(T_1 \le t_1, \ldots, T_k \le t_k, M(T)=k)}{\mathbf{P}\,(M(T)=k)}.$$

By the result proved in part (a), the denominator has the form

$$\mathbf{P}\,(M(T)=k) = \frac{(\lambda T)^k}{k!}\mathrm{e}^{-\lambda T},\ k=0,1,\ldots,$$

while the numerator can be written as follows:

$$
\begin{aligned}
&\mathbf{P}\,(T_1 \le t_1, \dots, T_k \le t_k, M(T) = k) \\
&\quad = \mathbf{P}\,(X_1 \le t_1, X_1 + X_2 \le t_2, \dots, X_1 + \dots + X_k \le t_k, X_1 + \dots + X_{k+1} > T) \\
&\quad = \int_0^{t_1} \int_0^{t_2-u_1} \int_0^{t_3-(u_1+u_2)} \dots \int_0^{t_k-(u_1+\dots+u_{k-1})} \int_{T-(u_1+\dots+u_k)}^{\infty} \prod_{i=1}^{k+1} \left(\lambda \mathrm{e}^{-\lambda u_i}\right) \mathrm{d}u_{k+1} \dots \mathrm{d}u_1 \\
&\quad = \lambda^k \int_0^{t_1} \int_0^{t_2-u_1} \int_0^{t_3-(u_1+u_2)} \dots \int_0^{t_k-(u_1+\dots+u_{k-1})} \mathrm{e}^{-\lambda(u_1+\dots+u_k)} \mathrm{e}^{-\lambda(T-u_1+\dots+u_k)} \mathrm{d}u_{k+1} \dots \mathrm{d}u_1 \\
&\quad = \lambda^k \mathrm{e}^{-\lambda T} \int_0^{t_1} \int_0^{t_2-u_1} \int_0^{t_3-(u_1+u_2)} \dots \int_0^{t_k-(u_1+\dots+u_{k-1})} \mathrm{d}u_k \dots \mathrm{d}u_1.
\end{aligned}
$$

Setting $v_1 = u_1, v_2 = u_1 + u_2, \dots, v_k = u_1 + \dots + u_k$, the last integral takes the form

$$
\frac{(\lambda T)^k}{k!} \mathrm{e}^{-\lambda T} \frac{k!}{T^k} \int_0^{t_1} \int_{v_1}^{t_2} \int_{v_2}^{t_3} \dots \int_{v_{k-1}}^{t_k} \mathrm{d}v_k \dots \mathrm{d}v_1,
$$

thus

$$
\mathbf{P}\,(T_1 \le t_1, \dots, T_k \le t_k \,|M(T) = k) = \frac{k!}{T^k} \int_0^{t_1} \int_{v_1}^{t_2} \int_{v_2}^{t_3} \dots \int_{v_{k-1}}^{t_k} \mathrm{d}v_k \dots \mathrm{d}v_1.
$$

From this we get that the joint conditional PDF of random variables $T_1, \dots, T_k$ given $M(T) = k$ equals the constant value $\frac{k!}{T^k}$, which, by the preceding lemma, is identical with the joint PDF of random variables $U_{1k}, \dots, U_{kk}$. Using the proof of Construction II, we obtain that Construction I has the result of a homogeneous Poisson process at rate λ on the interval $(0, T]$, and at the same time on the whole interval $(0, \infty)$, because T was chosen arbitrarily. □

2.7.3 *Basic Properties of a Homogeneous Poisson Process*

Let $N(t),\ t \ge 0$ be a homogeneous Poisson process with a rate λ. We enumerate below the main properties of $N(t)$.

(a) For any $t \ge 0$ the CDF of $N(t)$ is Poisson with the parameter λt, that is,

$$
\mathbf{P}\,(N(t) = k) = \frac{(\lambda t)^k}{k!} \mathrm{e}^{-\lambda t},\ k = 0, 1, \dots .
$$

(b) The increments of $N(t)-N(s)$, $0 \le s < t$, are independent and have a Poisson distribution with the parameter $\lambda(t-s)$.
(c) The sum of two independent homogeneous Poisson processes $N_1(t;\lambda_1)$ and $N_2(t;\lambda_2)$ at rates λ_1 and λ_2, respectively, is a homogeneous Poisson process with a rate $(\lambda_1+\lambda_2)$.
(d) Given $0 < t < T < \infty$, a positive integer N_0 and an integer k satisfy the inequality $0 \le k \le N_0$. The conditional CDF of the random variable $N(t)$ given $N(T) = N_0$ is binomial with the parameters $(N_0, 1/T)$.

Proof.

$$\begin{aligned}\mathbf{P}\left(N(t)=k \mid N(T)=N_0\right) &= \frac{\mathbf{P}\left(N(t)=k, N(T)=N_0\right)}{\mathbf{P}\left(N(T)=N_0\right)} \\ &= \frac{\mathbf{P}\left(N(t)=k, N(T)-N(t)=N_0-k\right)}{\mathbf{P}\left(N(T)=N_0\right)} \\ &= \frac{(\lambda t)^k}{k!}\mathrm{e}^{-\lambda t}\frac{(\lambda(T-t))^{N_0-k}}{(N_0-k)!}\mathrm{e}^{-\lambda(T-t)}\left(\frac{(\lambda T)^{N_0}}{N_0!}\mathrm{e}^{-\lambda T}\right)^{-1} \\ &= \binom{N_0}{k}\left(\frac{t}{T}\right)^k\left(1-\frac{t}{T}\right)^{N_0-k}.\end{aligned}$$

□

(e) The following asymptotic relations are valid as $h \to +0$:

$$\begin{aligned}&\mathbf{P}\left(N(h)=0\right) = 1-\lambda h + o(h),\\ &\mathbf{P}\left(N(h)=1\right) = \lambda h + o(h),\\ &\mathbf{P}\left(N(h)\ge 2\right) = o(h). \qquad \text{(orderliness)}\end{aligned}$$

Lemma 2.14. *For every nonnegative integer m the inequality*

$$\left|\mathrm{e}^x - \sum_{k=0}^{m}\frac{x^k}{k!}\right| < \frac{|x|^{m+1}}{(m+1)!}\mathrm{e}^{|x|} = o(|x|^m),\ x \to 0,$$

holds.

Proof. The assertion of the lemma follows from the nth-order Taylor approximation to e^x with the Lagrange form of the remainder term (see Sect. 7.7 of [4]), but one can obtain it by simple computations. Using the Taylor expansion

$$\mathrm{e}^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}$$

of the function e^x, which implies that

$$\left|\mathrm{e}^x - \sum_{k=0}^{m} \frac{x^k}{k!}\right| = \left|\sum_{k=m+1}^{\infty} \frac{x^k}{k!}\right| \le \frac{|x|^{m+1}}{(m+1)!} \sum_{k=0}^{\infty} \frac{(m+1)!}{(m+1+k)!} |x|^k <$$

$$< \frac{|x|^{m+1}}{(m+1)!} \sum_{k=0}^{\infty} \frac{|x|^k}{k!} = \frac{|x|^{m+1}}{(m+1)!} \mathrm{e}^{|x|} = o(|x|^m), \ x \to 0.$$

□

Proof of Property (e). From the preceding lemma we have as $h \to +0$

$$\mathbf{P}(N(h) = 0) = \mathrm{e}^{-\lambda h} = 1 - \lambda h + o(h),$$

$$\mathbf{P}(N(h) = 1) = \frac{\lambda h}{1!} \mathrm{e}^{-\lambda h} = \lambda h(1 - \lambda h + o(h)) = \lambda h + o(h),$$

$$\mathbf{P}(N(h) \ge 2) = 1 - \left(\mathrm{e}^{-\lambda h} + \frac{(\lambda h)^1}{1!} \mathrm{e}^{-\lambda h}\right) = -\left(\mathrm{e}^{-\lambda h} - 1 + \frac{(\lambda h)^1}{1!} \mathrm{e}^{-\lambda h}\right) = o(h).$$

□

(f) Given that exactly one event of a homogeneous Poisson process $[N(t),\ t \ge 0]$ has occurred during the interval $(0, t]$, the time of occurrence of this event is uniformly distributed over $(0, t]$.

Proof of Property (f). Denote by λ the rate of the process $N(t)$. Immediate application of the conditional probability gives for all $0 < x < t$

$$\mathbf{P}(X_1 \le x | N(t) = 1) = \frac{\mathbf{P}(X_1 \le x,\ N(t) = 1)}{\mathbf{P}(N(t) = 1)}$$

$$= \frac{\mathbf{P}(N(x) = 1, N(t) - N(x) = 0)}{\mathbf{P}(N(t) = 1)}$$

$$= \frac{\mathbf{P}(N(x) = 1)\mathbf{P}(N(t - x) = 0)}{\mathbf{P}(N(t) = 1)}$$

$$= \left(\frac{(\lambda x)^1}{1!} \mathrm{e}^{-\lambda x} \frac{[\lambda(t - x)]^0}{0!} \mathrm{e}^{-\lambda(t-x)}\right) \left(\frac{(\lambda t)^1}{1!} \mathrm{e}^{-\lambda t}\right)^{-1} = \frac{x}{t}.$$

□

(g) **Strong Markov property**. Let $\{N(t),\ t \ge 0\}$ be a homogeneous Poisson process with the rate λ, and assume that $N(t)$ is $\mathcal{A}_t$ measurable for all $t \ge 0$, where $\mathcal{A}_t \subset \mathcal{A},\ t \ge 0$, is a monotonically increasing family of σ-algebras. Let τ be a random variable such that the condition $\{\tau \le t\} \in \mathcal{A}_t$ holds for all $t \ge 0$. This type of random variable is called a **Markov point** with respect to

the family of σ-algebra $\mathcal{A}_t, t \geq 0$. For example, the constant $\tau = t$ and the so-called **first hitting time**, $\tau_k = \sup\{s : N(s) < k\}$, where k is a positive integer, are Markov points. Denote

$$N_\tau(t) = N(t+\tau) - N(\tau),\ t \geq 0.$$

Then the process $N_\tau(t),\ t \geq 0$, is a homogeneous Poisson process with the rate λ, which does not depend on the Markov point τ or on the process $\{N(t),\ 0 \leq t \leq \tau\}$.

(h) **Random deletion (filtering) of a Poisson process**. Let $N(t),\ t \geq 0$ be a homogeneous Poisson process with intensity $\lambda > 0$. Let us suppose that we delete points in the process $N(t)$ independently with probability $(1-p)$, where $0 < p < 1$ is a fixed number. Then the new process $M(t),\ t \geq 0$, determined by the undeleted points of $N(t)$ constitutes a homogeneous Poisson process with intensity $p\lambda$.

Proof of the Property (h). Let us represent the Poisson process $N(t)$ in the form

$$N(t) = \sum_{k=1}^{\infty} \mathcal{I}_{\{t_k \leq t\}},\ \ t \geq 0,$$

where $t_k = X_1 + \ldots + X_k,\ k = 1, 2, \ldots$ and $X_1, X_2, \ldots$ are independent exponentially distributed random variables with the parameter λ. The random deletion in the process $N(t)$ can be realized with the help of a sequence of independent and identically distributed random variables $I_1, I_2, \ldots$, which do not depend on the process $N(t),\ t \geq 0$ and have a distribution $\mathbf{P}(I_k = 1) = p$, $\mathbf{P}(I_k = 0) = 1 - p$. The deletion of a point t_k in the process $N(t)$ happens only in the case $I_k = 0$. Let $T_0 = 0$, and denote by $0 < T_1 < T_2 < \ldots$ the sequence of remaining points. Thus the new process can be given in the form

$$M(t) = \sum_{k=1}^{\infty} \mathcal{I}_{\{T_k \leq t\}} = \sum_{k=1}^{\infty} \mathcal{I}_{\{t_k \leq t,\ I_k = 1\}},\ \ t \geq 0.$$

Using the property of the process $N(t)$ and the random sequence $I_k,\ k \geq 1$, it is clear that the sequence of random variables $Y_k = T_k - T_{k-1},\ k = 1, 2, \ldots$, are independent and identically distributed; therefore, it is enough to prove that they have an exponential distribution with the parameter $p\lambda$, i.e., $\mathbf{P}(Y_k < y) = 1 - \mathrm{e}^{p\lambda y}$.

The sequence of the remaining points T_k can be given in the form $T_k = t_{n_k},\ k = 1, 2, \ldots$, where the random variables n_k are defined as follows:

$$n_1 = \min\{j :\ j \geq 1,\ I_j = 1\},$$
$$n_k = \min\{j :\ j > n_{k-1},\ I_j = 1\},\ k \geq 2.$$

Let us compute the distribution of the random variable

$$Y_1 = T_1 = X_1 + \ldots + X_{n_1}.$$

By the use of the formula of total probability, we obtain

$$\begin{aligned}\mathbf{P}(Y_1 < y) &= \mathbf{P}(X_1 + \ldots + X_{n_1} < y) \\ &= \sum_{k=1}^{\infty} \mathbf{P}(X_1 + \ldots + X_{n_1} < y | n_1 = k)\mathbf{P}(n_1 = k) \\ &= \sum_{k=1}^{\infty} \mathbf{P}(X_1 + \ldots + X_k < y)\mathbf{P}(n_1 = k).\end{aligned}$$

The sum $X_1 + \ldots + X_k$ of independent exponentially distributed random variables X_i has a gamma distribution with the density function

$$f(y; k, \lambda) = \frac{\lambda^k}{(k-1)!} y^{k-1} \mathrm{e}^{-\lambda y},\ y > 0,$$

whereas, on the other hand, the random variable n_1 has a geometric distribution with the parameter p, i.e.,

$$\mathbf{P}(n_1 = k) = (1-p)^{k-1} p;$$

therefore, we get

$$\sum_{k=1}^{\infty} \mathbf{P}(X_1 + \ldots + X_k < y)\mathbf{P}(n_1 = k) = \sum_{k=1}^{\infty} \int_0^y \frac{\lambda^k}{(k-1)!} x^{k-1} \mathrm{e}^{-\lambda x} (1-p)^{k-1} p \mathrm{d}x$$

$$= \lambda p \int_0^y \left(\sum_{k=0}^{\infty} \frac{[(1-p)\lambda x]^k}{k!} \right) \mathrm{e}^{-\lambda x} \mathrm{d}x = \lambda p \int_0^y \mathrm{e}^{(1-p)\lambda x} \mathrm{e}^{-\lambda x} \mathrm{d}x$$

$$= \lambda p \int_0^y \mathrm{e}^{-p\lambda x} \mathrm{d}x = 1 - \mathrm{e}^{-p\lambda y}.$$

□

(1) **Modeling an inhomogeneous Poisson process**. Let $\{\Lambda(t),\ t \geq 0\}$ be a non-negative, monotonically nondecreasing, continuous-from-left function such that $\Lambda(0) = 0$. Let $N(t),\ t \geq 0$, be a homogeneous Poisson process with rate 1. Then the process defined by the equation

$$N_\Lambda(t) = N(\Lambda(t)),\ t \geq 0,$$

is a Poisson process with mean value function $\Lambda(t),\ t \geq 0$.

Proof of Property (i). Obviously, $N(\Lambda(0)) = N(0) = 0$, and the increments of the process $N_\Lambda(t)$ are independent and the CDF of the increments are Poissonian, because for any $0 \leq s \leq t$ the CDF of the increment $N_\Lambda(t) - N_\Lambda(s)$ is Poisson with the parameter $\Lambda(t) - \Lambda(s)$,

$$\begin{aligned}\mathbf{P}\left(N_\Lambda(t) - N_\Lambda(s) = k\right) &= \mathbf{P}\left(N(\Lambda(t)) - N(\Lambda(s))\right) \\ &= \frac{(\Lambda(t) - \Lambda(s))^k}{k!} \mathrm{e}^{-(\Lambda(t)-\Lambda(s))},\ k = 0, 1, \ldots .\end{aligned}$$

□

2.7.4 Higher-Dimensional Poisson Process

The Poisson process can be defined, in higher dimensions, as a model of random points in space. To do this, we first concentrate on the process on the real number line, from the aspect of a possible generalization.

Let $\{N(t),\ t \geq 0\}$ be a Poisson process on a probability space $(\Omega, \mathcal{A}, P)$. Assume that it has a rate function $\lambda(t),\ t \geq 0$; thus, the mean value function has the form

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s,\ t \geq 0,$$

where the function $\lambda(t)$ is nonnegative and locally integrable function. Denote by $t_1, t_2, \ldots$ the sequence of the random jumping points of $N(t)$. Since the mean value function is continuous, the jumps of $N(t)$ are exactly 1; moreover, the process $N(t)$ and the random points $\Pi = \{t_1, t_2, \ldots\}$ determine uniquely each other. If we can characterize the countable set Π of random points $\{t_1, t_2, \ldots\}$, then at the same time we can give a new definition of the Poisson process $N(t)$.

Denote by $\mathcal{B}_+ = \mathcal{B}(\mathbb{R}_+)$ the Borel σ-algebra of the half line $\mathbb{R}_+ = [0, \infty)$, i.e., the minimal σ-algebra that consists of all open intervals of $\mathbb{R}_+$. Let $B_i = (a_i, b_i],\ i = 1, \ldots, n$, be nonoverlapping intervals of $\mathbb{R}_+$; then obviously $B_i \in \mathcal{B}_+$. Introduce the random variables

$$\Pi(B_i) = \#\{\Pi \cap B_i\} = \#\left\{t_j : t_j \in B_i\right\},\ i = 1, \ldots, n,$$

where $\#\{\cdot\}$ means the number of elements of a set; then

$$\Pi(B_i) = N(b_i) - N(a_i).$$

By the use of the properties of Poisson processes, the following statements hold:

(1) The random variables $\Pi(B_i)$ are independent because the increments of the process $N(t)$ are independent.
(2) The CDF of $\Pi(B_i)$ is Poisson with the parameter $\Lambda(B_i)$, i.e.,

$$\mathbf{P}\left(\Pi(B_i) = k\right) = \frac{(\Lambda(B_i))^k}{k!}\mathrm{e}^{-\Lambda(B_i)},$$

where $\Lambda(B_i) = \int_{B_i} \lambda(s)\mathrm{d}s,\ 1 \le i \le n$.

Observe that by the definition of random variables $\Pi(B_i)$, it is unimportant whether or not the set of random points $\Pi = \{t_i\}$ is ordered and $\Pi(B_i)$ is determined by the number of points t_i only, which is included in the interval $(a_i, b_i]$. This circumstance is important because we want to define the Poisson processes on higher-dimensional spaces, which do not constitute an ordered set, contrary to the one-dimensional case.

More generally, let $B_i \in \mathcal{B}(\mathbb{R}_+),\ 1 \le i \le n$, be disjoint Borel sets and denote $\Pi(B_i) = \#\{\Pi \cap B_i\}$. It can be checked that $\Pi(B_i)$ are random variables defined by the random points $\Pi = \{t_1, t_2, \ldots\}$ and they satisfy properties (1) and (2). On this basis, the Poisson process can be defined in higher-dimensional Euclidean spaces and, in general, in metric spaces also (see Chap. 2. of [54]).

Consider the d-dimensional Euclidean space $S = \mathbb{R}^d$ and denote by $\mathcal{B}(S)$ the Borel σ-algebra of the subset of S. We will define the Poisson process Π as a random set function satisfying properties (1) and (2). Let $\Pi : \Omega \to \mathcal{S}$ be a random point set in $\mathcal{S}$, where $\mathcal{S}$ denotes the set of all subsets of S consisting of countable points. Then the quantities $\Pi(A) = \#\{\Pi \cap A\}$ define random variables for all $A \in \mathcal{B}(S)$.

Definition 2.15. We say that Π is a Poisson process on the space S if $\Pi \in \mathcal{S}$ is a random countable set of points in S and the following conditions are satisfied:

(1) The random variables $\Pi(A_i) = \#\{\Pi \cap A_i\}$ are independent for all disjoint sets $A_1, \ldots, A_n \in \mathcal{B}(S)$.
(2) For any $A \in \mathcal{B}(S)$ the CDF of random variables $\Pi(A)$ are Poisson with the parameter $\Lambda(A)$, where $0 \le \Lambda(A) \le \infty$.
The function $\Lambda(A),\ A \in \mathcal{B}(S)$ is called a **mean measure** of a Poisson process (see [54], p. 14).

Properties:

1. Since the random variable $\Pi(A)$ has a Poisson distribution with the parameter $\Lambda(A)$, then $\mathbf{E}\left(\Pi(A)\right) = \Lambda(A)$ and $\mathbf{D}^2\left(\Pi(A)\right) = \Lambda(A)$.
2. If $\Lambda(A)$ is finite, then the random variable $\Pi(A)$ is finite with probability 1, and if $\Lambda(A) = \infty$, then the number of elements of the random point set $\Pi \cap A$ is countably infinite with probability 1.
3. For any disjoint sets $A_1, A_2, \ldots \in \mathcal{B}(S)$,

$$\Pi(A) = \sum_{i=1}^{\infty} \Pi(A_i) \quad \text{and} \quad \Lambda(A) = \sum_{i=1}^{\infty} \Lambda(A_i),$$

where $A = \cup_{i=1}^{\infty} A_i$. The last relation means that the mean measure $\Lambda(B),\ B \in \mathcal{B}(S)$ satisfies the conditions of a measure, i.e., it is a nonnegative, σ-additive set function on the measurable space $(S, \mathcal{B}(S))$, which justifies the name of Λ.

Like the one-dimensional case, when the Poisson process has a rate function, it is an important class of Poisson processes for which there exists a nonnegative locally integrable function λ with the property

$$\Lambda(B) = \int_B \lambda(s)\mathrm{d}s,\ B \in \mathcal{B}(S)$$

(here the integral is defined with respect to the Lebesgue measure ds). Then the mean measure Λ is **nonatomic**, that is, there is no point $s_0 \in \mathcal{B}(S)$ such that $\Lambda(\{s_0\}) > 0$.

4. By the use of properties 1 and 3, it is easy to obtain the relation

$$\mathbf{D}^2\left(\Pi(A)\right) = \sum_{i=1}^{\infty} \mathbf{D}^2\left(\Pi(A_i)\right) = \sum_{i=1}^{\infty} \Lambda(A_i) = \Lambda(A).$$

5. For any $B, C \in \mathcal{B}(S)$,

$$\mathrm{cov}(\Pi(B), \Pi(C)) = \Lambda(B \cap C).$$

Proof. Since $\Pi(B) = \Pi(B \cap C) + \Pi(B \backslash C)$ and $\Pi(C) = \Pi(B \cap C) + \Pi(C \backslash B)$, where the sets $A \cap C$, $A \backslash C$ and $C \backslash A$ are disjoint, the $\Pi(A \cap C)$, $\Pi(A \backslash C)$, and $\Pi(C \backslash A)$ are independent random variables, and thus

$$\begin{aligned} \mathrm{cov}(\Pi(A), \Pi(C)) &= \mathrm{cov}(\Pi(A \cap C), \Pi(A \cap C)) \\ &= \mathbf{D}^2\left(\Pi(A \cap C)\right) = \Lambda(A \cap C). \end{aligned}$$

□

6. For any (not necessarily disjoint) sets $A_1, \ldots, A_n \in \mathcal{B}(S)$ the joint distribution of random variables $\Pi(A_1), \ldots, \Pi(A_n)$ is uniquely determined by the mean measure Λ.

Proof. Denote the set of the 2^n pairwise disjoint sets by

$$\mathcal{C} = \{C = B_1 \cap \ldots \cap B_n,\ \text{where } B_i \text{ means the set either } A_i, \text{ or } \bar{A}_i\};$$

then the random variables $\Pi(C)$ are independent and have a Poisson distribution with the parameter $\Lambda(C)$. Consequently, the random variables $\Pi(A_1), \ldots, \Pi(A_n)$

can be given as a sum from a 2^n number of independent random variables $\Pi(C),\ C \in \mathcal{C}$, having a Poisson distribution with the parameter $\Lambda(C)$; therefore, the joint distribution of random variables $\Pi(A_i)$ is uniquely determined by $\Pi(C),\ C \in \mathcal{C}$, and the mean measure Λ. □

Comment 2.16. *Let $S = \mathbb{R}^d$, and assume*

$$\Lambda(A) = \int_A \lambda(x)\mathrm{d}x,\ A \in \mathcal{B}(S),$$

where $\lambda(x)$ is a nonnegative and locally integrable function and $dx = dx_{1\ldots}dx_n$. If $|A|$ denotes the n-dimensional (Lebesgue) measure of a set A and the function $\lambda(x)$ is continuous at a point $x_0 \in S$, then

$$\Lambda(A) \sim \lambda(x_0)\,|A|$$

if the set A is included in a small neighborhood of the point x_0.

The Poisson process Π is called ***homogeneous*** *if $\lambda(x) = \lambda$ for a positive constant λ. In this case for any $A \in \mathcal{B}(S)$ the inequality $\Lambda(A) = \lambda\,|A|$ holds.*

The following three theorems state general assertions on the Poisson processes defined in higher-dimensional spaces (see Chap. 2 of [54]).

Theorem 2.17 (*Existence theorem). If the mean measure Λ is nonatomic on the space S and it is σ-finite, i.e., it can be expressed in the form*

$$\Lambda = \sum_{i=1}^{\infty} \Lambda_i,\ \ \textit{where}\ \ \Lambda_i(S) < \infty,$$

then there exists a Poisson process Π on the space S and has mean measure Λ.

Theorem 2.18 (*Superposition theorem). If $\Pi_i, i = 1, 2, \ldots$, is a sequence of independent Poisson processes with mean measure $\Lambda_1, \Lambda_2, \ldots$ on the space S, then the superposition $\Pi = \cup_{i=1}^{\infty}\Pi_i$ is a Poisson process with mean measure $\Lambda = \sum_{i=1}^{\infty}\Lambda_i$.*

Theorem 2.19 (*Restriction theorem). Let Π be a Poisson process on the space S with mean measure Λ. Then for any $S_0 \in \mathcal{B}(S)$ the process*

$$\Pi_0 = \Pi \cap S_0$$

can be defined as a Poisson process on S with mean measure

$$\Lambda_0(A) = \Lambda(A \cap S_0).$$

The process Π_0 can be interpreted as a Poisson process on the space S_0 with mean measure Λ_0, where Λ_0 is called the restriction of mean measure Λ to S_0.

2.8 Exercises

Exercise 2.1. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with finite absolute moment $\mathbf{E}(|X_1|) < \infty$. Let N be a random variable taking positive integer numbers and independent of the random variable $(X_i, i = 1, 2, \ldots)$. Prove that

(a) $\mathbf{E}(X_1 + \ldots + X_N) = \mathbf{E}(X_1)\,\mathbf{E}(N)$,
(b) $\mathbf{D}^2(X_1 + \ldots + X_N) = \mathbf{D}^2(X_1) + (\mathbf{E}(X_1))^2\,(\mathbf{E}(N))^2$
(Wald identities or Wald lemma).

Exercise 2.2. Let $X_0, X_1, \ldots$ be independent random variables with joint distribution $\mathbf{P}(X_i = 1) = \mathbf{P}(X_i = -1) = \frac{1}{2}$.

Define $Z_0 = 0$, $Z_k = Z_{k-1} + X_k$, $k = 0, 1, \ldots$. Determine the expectation and covariance function of the process $(Z_k, k = 1, 2, \ldots)$ (random walk on the integer numbers).

Let a and b be real numbers, $|b| < 1$. Denote $W_0 = aX_0$, $W_k = bW_{k-1} + X_k$, $k = 1, 2, \ldots$ [here the process $(W_k, k = 0, 1, \ldots)$ constitutes a first-degree autoregressive process with the initial value aX_0 and with the innovation process $(X_k, k = 1, 2, \ldots)$]. If we fix the value b, for which value of a will the process W_k be stationary in a weak sense?

Exercise 2.3. Let a and b be real numbers, and let U be a random variable uniformly distributed on the interval $(0, 2\pi)$. Denote $X_t = a\cos(bt + U)$, $-\infty < t < \infty$. Prove that the random cosine process $(X_t, -\infty < t < \infty)$ is stationary.

Exercise 2.4. Let $N(t), T \geq 0$ be a homogeneous Poisson process with intensity λ.

(a) Determine the covariance and correlation functions of $N(t)$.
(b) Determine the conditional expectation $\mathbf{E}(N(t + s) \mid N(t))$.

Chapter 3
Markov Chains

In the early twentieth century, Markov (1856–1922) introduced in [67] a new class of models called Markov chains, applying sequences of dependent random variables that enable one to capture dependencies over time. Since that time, Markov chains have developed significantly, which is reflected in the achievements of Kolmogorov, Feller, Doob, Dynkin, and many others. The significance of the extensive theory of Markov chains and the continuous-time variant called Markov processes is that it can be successfully applied to the modeling behavior of many problems in, for example, physics, biology, and economics, where the outcome of one experiment can affect the outcome of subsequent experiments. The terminology is not consistent in the literature, and many authors use the same name (Markov chain) for both discrete and continuous cases. We also apply this terminology.

Heuristically, the property that characterizes Markov chains can be expressed by the so-called memoryless notion (Markov property) as follows: a Markov chain is a stochastic process for which future behavior, given the past and the present, depends only on the present and not on the past.

This chapter presents a brief introduction to the theory of discrete-time Markov chains (DTMCs) and to the continuous-time variant, continuous-time Markov chains (CTMCs), that will be applied to the modeling and analysis of queueing systems. Note that DTMCs and CTMCs taking values in a set of countable elements have many similar properties; however, in contrast to discrete-time processes, the characteristics of a sample path essentially differ in continuous cases.

We limit ourselves here to the definition of Markov processes and to their basic properties with countable state space in discrete time $\mathcal{T} = \{0, 1, \ldots\}$ and continuous time $\mathcal{T} = [0, \infty)$. In connection with the classic results discussed in this chapter, we refer mainly to the classic works [35, 36].

Consider a discrete-time or continuous-time stochastic process $X = (X_t,\ t \in \mathcal{T})$ given on a probability space $(\Omega, \mathcal{A}, P)$ and taking values in a countable set, called the **state space**, $\mathcal{X} = \{x_0, x_1, \ldots\}$. The state space $\mathcal{X}$ is called **finite** if it consists of a finite number of elements. The sample path of a discrete-time process with discrete

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, 77
DOI 10.1007/978-1-4614-5317-8_3, © Springer Science+Business Media, LLC 2013

sample space is defined in the space of sequences $\mathcal{S} = \{x_{k_0}, x_{k_1}, \ldots\}$, $x_{k_i} \in \mathcal{X}$, while it is an element of the space of all functions $\mathcal{S} = \{x_t : x_t \in \mathcal{X},\ t \geq 0\}$ in continuous-time cases.

We say that the process is **in the state** $x \in \mathcal{X}$ at the time $t \in \mathcal{T}$ if $X_t = x$. The process starts from a state $x_0 \in \mathcal{X}$ determined by the distribution of the random variable X_0, which is the **initial distribution** of the process. If there exists a state $x_0 \in \mathcal{X}$ for which $\mathbf{P}(X_0 = x_0) = 1$, then the state x_0 is called the **initial state**. The state of the process can change from time to time, and these changes in state are known as **transitions**. The probabilities of these state changes are called **transition probabilities**, which with the initial distribution determine the statistical behavior of the process.

If we denote by $\mathcal{B}_X$ the σ-algebra of all subsets of the state space $\mathcal{X}$, then the pair $(\mathcal{X}, \mathcal{B}_X)$ is a measurable space and the connection $\{X_t \in A\} \in \mathcal{A}$ holds for all $t \in \mathcal{T}$ and $\{A \in \mathcal{B}_X\} \in \mathcal{A}$.

Definition 3.1. A stochastic process $(X_t, t \in \mathcal{T})$ with the discrete state space $\mathcal{X}$ is called a **Markov chain** if for every nonnegative integer n and for all $t_0 < \ldots < t_n < t_{n+1},\ t_i \in \mathcal{T},\ x_0, \ldots, x_{n+1} \in \mathcal{X}$

$$\mathbf{P}\left(X_{t_{n+1}} = x_{n+1} \mid X_{t_0} = x_0, \ldots, X_{t_n} = x_n\right) = \mathbf{P}\left(X_{t_{n+1}} = x_{n+1} \mid X_{t_n} = x_n\right), \tag{3.1}$$

provided that this conditional probability exists. Let $x, y \in \mathcal{X},\ s \leq t,\ s, t \in \mathcal{T}$; then the function

$$p_{x,y}(s,t) = \mathbf{P}\left(X_t = y \mid X_s = x\right)$$

is called a **transition probability function** of a Markov chain. If the equation $p_{x,y}(s,t) = p_{x,y}(t-s)$ holds for all $x, y \in \mathcal{X},\ s \leq t,\ s, t \in \mathcal{T}$, then the Markov chain is called (**time**) **homogeneous**; otherwise it is known as **inhomogeneous**.

In both discrete- and continuous-time cases, this definition expresses the aforementioned memoryless property of a Markov chain, and it ensures that the transition probabilities depend only on the present state X_s, not on how the present state was reached. We start with a discussion of DTMCs.

3.1 Discrete-Time Markov Chains with Discrete State Space

Given a Markov chain $X = (X_t,\ t \in \mathcal{T})$, $\mathcal{T} = \{0, 1, \ldots\}$ on a probability space $(\Omega, \mathcal{A}, P)$ taking values in a finite or countably infinite set of elements $\mathcal{X}$. It is conventional to denote the finite state space by the set $\mathcal{X} = \{0, 1, \ldots, K\}$ $(0 < K < \infty)$ and the countably infinite one by $\mathcal{X} = \{0, 1, \ldots\}$. This notation is quite reasonable for queueing systems, and in general, it does not lead to a separate problem if the elements of $\mathcal{X}$ serve to distinguish the states only; otherwise, the state space is chosen based on practical requirements. Assume that the events $\{X_t = i\},\ i \in \mathcal{X}$, are disjoint for all $t \in \mathcal{T}$.

In the discrete-time case we can give an alternative definition of a Markov chain instead of Eq. (3.1).

Definition 3.2. A discrete-time stochastic process X with state space $\mathcal{X}$ is called a Markov chain if for every $n = 0, 1, \ldots$ and for all states $i_0, \ldots, i_{n+1} \in \mathcal{X}$

$$\begin{aligned} p_{i_n, i_{n+1}}(n, n+1) &= \mathbf{P}\left(X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots, X_n = i_n\right) \\ &= \mathbf{P}\left(X_{n+1} = i_{n+1} \mid X_n = i_n\right), \end{aligned} \tag{3.2}$$

provided that a conditional probability exists. The probability

$$p_{i,j}(n, n+1) = \mathbf{P}\left(X_{n+1} = j \mid X_n = i\right),\ i, j \in \mathcal{X}, n = 0, 1, \ldots,$$

is called a **one-step transition probability**, which is the probability of a transition from a state i to a state j in a single step from time n to time $n + 1$.

Relation (3.2) is simpler in our case than that of Eq. (3.1), but it can be easily checked that they are equivalent to each other. Here we can define, from a practical point of view, the transition probability $p_{i,j}(s, t) = 0$, when the probability of the event $\{X_s = i\}$ equals 0 at the time point s because if $\mathbf{P}\{X_s = i\} = 0$ holds, then the sample path arrives at the state i with probability 0 at time s; therefore, the quantity $p_{i,j}(s, t)$ can be defined freely in this case.

Definition 3.3. We say that a stochastic process X with state space $\mathcal{X}$ is a **Markov chain of m-order** (or a **Markov chain with memory** m) if for every $n = 1, 2, \ldots$ and for arbitrary states $i_k \in \mathcal{X}$, $k = 0, \ldots, n + m$,

$$\begin{aligned} &\mathbf{P}\left(X_{n+m} = i_{n+m} \mid X_0 = i_0, \ldots, X_{n+m-1} = i_{n+m-1}\right) \\ &= \mathbf{P}\left(X_{n+m} = i_{n+m} \mid X_n = i_n, \ldots, X_{n+m-1} = i_{n+m-1}\right), \end{aligned}$$

provided that a conditional probability exists.

It is not difficult to verify that an m-order Markov chain can be represented as a first-order one if we introduce a new m-dimensional process as follows. Define the vector-valued process $Y = (Y_0, Y_1, \ldots)$,

$$Y_n = (X_n, \ldots, X_{n+m-1}),\ n = 0, 1, \ldots,$$

with state space

$$\mathcal{X}' = \{(k_1, \ldots, k_m) :\ k_1, \ldots, k_m \in \mathcal{X}\}.$$

Then the process Y is a first-order Markov chain because

$$\begin{aligned} &\mathbf{P}\left(Y_{n+1} = (i_{n+1}, \ldots, i_{n+m}) \mid Y_0 = (i_0, \ldots, i_{m-1}), \ldots, Y_n = (i_n, \ldots, i_{n+m-1})\right) \\ &\quad = \mathbf{P}\left(X_{n+m} = i_{n+m}, \ldots, X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots, X_{n+m-1} = i_{n+m-1}\right) \\ &\quad = \mathbf{P}\left(X_{n+m} = i_{n+m}, \ldots, X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots, X_{n+m-1} = i_{n+m-1}\right) \\ &\quad = \mathbf{P}\left(Y_{n+1} = (i_{n+1}, \ldots, i_{n+m}) \mid Y_n = (i_n, \ldots, i_{n+m-1})\right). \end{aligned}$$

This is why we consider only first-order Markov chains and why, later on, we will write only *Markov chain* instead of *Markov chain of first order*.

In the theory of DTMCs, the initial distribution

$$P = (p_i,\ i \in \mathcal{X}),\ \text{where } p_i = \mathbf{P}(X_0 = i),$$

and the transition probabilities [see Eq. (3.2)]

$$p_{ij}(n, n+1),\ i, j \in \mathcal{X}, n = 0, 1, \ldots,$$

play a fundamental role because the statistical behavior of a Markov chain is completely determined by them (Theorem 3.4).

The states i and j, which play a role in Definition 3.2, can be identical, which means that the process can remain in the same state at the next time point. We say that a Markov chain X is (**time**) **homogeneous** if the transition probabilities do not depend on time shifting, that is,

$$p_{ij} = \mathbf{P}(X_{n+1} = j \mid X_n = i) = \mathbf{P}(X_1 = j \mid X_0 = i),\ \ i, j \in \mathcal{X}, n = 0, 1, \ldots.$$

If a Markov chain is not homogeneous, then it is called **inhomogeneous**.

3.1.1 Homogeneous Markov Chains

From a practical point of view, the class of homogeneous Markov chains plays a significant role; therefore, in this chapter we will investigate the properties of this class of processes. However, many results for homogeneous cases remain valid in the inhomogeneous case, too.

By definition, for a homogeneous Markov chain the one-step transition probability (or simply transition probability) $p_{i,j},\ i, j \in \mathcal{X}$, equals the probability that, starting from the initial state $X_0 = i$ at time 0, the process will be in the state j at the next time point 1, and this probability does not change if we take the transition probability in arbitrary time $n = 1, 2, \ldots,$

$$p_{ij} = \mathbf{P}\{X_1 = j \mid X_0 = i\} = \mathbf{P}\{X_{n+1} = j \mid X_n = i\}.$$

The transition probabilities satisfy the equation

$$\sum_{j \in \mathcal{X}} p_{ij} = 1.$$

This equation expresses the obvious fact that starting in a state i at the next time point the process takes certainly some state $j \in \mathcal{X}$. The following theorem states that the initial distribution and the transition probabilities determine the finite-dimensional distribution of a homogeneous Markov chain, and as a consequence we obtain that a Markov chain can be given in a statistical sense with the state space, the initial distribution, and the transition probabilities.

Theorem 3.4. *The finite-dimensional distributions of a Markov chain X are uniquely determined by the initial distribution and the transition probabilities and*

$$\mathbf{P}(X_0 = i_0, \dots, X_n = i_n) = p_{i_{n-1}i_n} p_{i_{n-2}i_{n-1}} \cdot \dots \cdot p_{i_0 i_1} p_{i_0}. \tag{3.3}$$

Proof. Let n be a positive integer, and let $i_0, \dots, i_n \in \mathcal{X}$. First, assume that $\mathbf{P}(X_0 = i_0, \dots, X_n = i_n) > 0$. By the definition of conditional probability,

$$\begin{aligned}
&\mathbf{P}(X_0 = i_0, \dots, X_n = i_n) \\
&\quad = \mathbf{P}(X_n = i_n \mid X_0 = i_0, \dots, X_{n-1} = i_{n-1}) \mathbf{P}(X_0 = i_0, \dots, X_{n-1} = i_{n-1}) = \dots \\
&\quad = \mathbf{P}(X_n = i_n \mid X_0 = i_0, \dots, X_{n-1} = i_{n-1}) \\
&\qquad \times \mathbf{P}(X_{n-1} = i_{n-1} \mid X_0 = i_0, \dots, X_{n-2} = i_{n-2}) \cdot \dots \cdot \mathbf{P}(X_1 = i_1 \mid X_0 = i_0) \\
&\qquad\quad \mathbf{P}(X_0 = i_0).
\end{aligned}$$

Using the Markov property we can rewrite this formula in the form

$$\begin{aligned}
&\mathbf{P}(X_0 = i_0, \dots, X_n = i_n) \\
&\quad = \mathbf{P}(X_n = i_n \mid X_{n-1} = i_{n-1}) \cdot \dots \cdot \mathbf{P}(X_1 = i_1 \mid X_0 = i_0) \mathbf{P}(X_0 = i_0) \\
&\quad = p_{i_{n-1}i_n} p_{i_{n-2}i_{n-1}} \cdot \dots \cdot p_{i_0 i_1} p_{i_0}.
\end{aligned}$$

If $\mathbf{P}(X_0 = i_0, \dots, X_n = i_n) = 0$, then either $\mathbf{P}(X_0 = i_0) = p_{i_0} = 0$ or there exists an index m, $0 \le m \le n-1$, for which

$$\mathbf{P}(X_0 = i_0, \dots, X_m = i_m) > 0 \text{ and } \mathbf{P}(X_0 = i_0, \dots, X_{m+1} = i_{m+1}) = 0.$$

Consequently,

$$\begin{aligned}
&\mathbf{P}(X_0 = i_0, \dots, X_{m+1} = i_{m+1}) \\
&\quad = \mathbf{P}(X_{m+1} = i_{m+1} \mid X_0 = i_0, \dots, X_m = i_m) \mathbf{P}(X_0 = i_0, \dots, X_m = i_m) \\
&\quad = p_{i_m i_{m+1}} \mathbf{P}(X_0 = i_0, \dots, X_m = i_m),
\end{aligned}$$

and therefore $p_{i_m i_{m+1}} = 0$. This means that the product $p_{i_{n-1}i_n} p_{i_{n-2}i_{n-1}} \cdot \dots \cdot p_{i_0 i_1} p_{i_0}$ equals 0 in both cases, and so assertion (3.3) of the theorem is true. □

Comment 3.5. *From relation (3.4) it immediately follows that for any $A_i \subset \mathcal{X}$, $0 \le i \le n$ the probability $\mathbf{P}(X_0 \in A_0, \dots, X_n \in A_n)$ can be given in the form*

$$\mathbf{P}(X_0 \in A_0, \dots, X_n \in A_n) = \sum_{i_0 \in A_0} \dots \sum_{i_n \in A_n} \mathbf{P}(X_0 = i_0, \dots, X_n = i_n),$$

where the probabilities are determined by relation (3.3), that is, with the help of the initial distribution and the transition probabilities.

The following remark clarifies an essential property of the homogeneous Markov chain, and on that basis limit theorems can be proved. This property relates the behavior of Markov chains to renewal and regenerative processes, which we will discuss later on in Sects. 4.1 and 4.2.

Comment 3.6. *From the memoryless property of a Markov chain X it follows that we can divide the time access into disjoint parts where the process behavior is mutually independent and follows the same probabilistic rules. We define the limits of these independent parts by the time instants when the process stays in the state $i_0 \in \mathcal{X}$.*

Formally, we define the sequence of random time points $0 \le \tau_1 < \tau_2 < \dots$ by the condition $X_{\tau_n} = i_0$, $n = 1, 2, \dots$, and $X_s \ne i_0$ if $s \notin \{\tau_1, \tau_2, \dots\}$. In this way $0 \le \tau_1 < \tau_2 < \dots$ are the times of the first, second, etc. visits to the state i_0, and i_0 is not visited between τ_n and τ_{n+1}, $n = 1, 2, \dots$. We define Y_n and $Z_{n,k}$ by $Y_n = \tau_{n+1} - \tau_n$ and $Z_{n,k} = X_{\tau_n + k}$, $0 \le k < Y_n$. Y_n is the time between the nth and the $n+1$th visits to i_0, and $Z_{n,k}$ is the state of the process at k steps after the nth visit to i_0, having that the next visit to i_0 is after $\tau_n + k$. Using the memoryless property of the Markov chain X we obtain that the random variables $(Y_n,\ Z_{n,k},\ 0 \le k < Y_n)$, $n = 1, 2, \dots$, are independent and their stochastic behaviors are identical. This fact ensures that the process is regenerative (Sect. 4.2).

In many cases, the study of Markov chains will be made simpler by the use of transition probability matrices.

Definition 3.7. The matrices associated with the transition probabilities of a Markov chain $\mathcal{X}$ with finite or countably infinite elements are

$$\Pi = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0N} \\ p_{10} & p_{11} & \cdots & p_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N0} & p_{N1} & \cdots & p_{NN} \end{bmatrix} \quad \text{and} \quad \Pi = \begin{bmatrix} p_{00} & p_{01} & p_{12} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

These matrices are called (one-step) **transition probability matrices**.

A matrix with nonnegative entries $\mathbf{A} = \left[a_{ij}\right]_{i,j \in \mathcal{X}}$ is called a **stochastic matrix** if for every row the sum of row elements equals 1. Then all transition probability matrices are stochastic ones:

(a) The elements of Π are obviously nonnegative,

$$p_{ij} \ge 0,\ i, j \in \mathcal{X}.$$

(b) For every i the sum of the ith row elements of Π equals 1,

$$\sum_{j \in \mathcal{X}} p_{ij} = 1,\ i \in \mathcal{X}.$$

The first of the following three examples shows that a sequence of independent and identically distributed discrete random variables is a homogeneous Markov chain. The second one shows that the sequence of sums of these random variables also constitutes a homogeneous Markov chain. If in the second case the random variables are independent, but not identically distributed, then the defined sequences will be an inhomogeneous Markov chain. The third example describes the stochastic behavior of a random walk on the real number line; in this case it is reasonable to choose the state space to be the set of all integer numbers, that is, $\mathcal{X} = \{0, \pm 1, \pm 2, \ldots\}$.

Let $Z_0, Z_1, \ldots$ be a sequence of independent and identically distributed random variables with a common CDF

$$\mathbf{P}(Z_m = k) = p_k,\ p_k \geq 0,\ k = 0, 1, \ldots,\ m = 0, 1, \ldots .$$

Example 3.8. Define the discrete-time stochastic process X with the relation $X_n = Z_n,\ n = 0, 1, \ldots$. Then X is a homogeneous Markov chain with initial distribution $\mathbf{P}(X_0 = k) = p_k,\ k = 0, 1, \ldots$ and transition probability matrix

$$\Pi = \begin{bmatrix} p_0 & p_1 & p_2 & \cdots \\ p_0 & p_1 & p_2 & \cdots \\ p_0 & p_1 & p_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Example 3.9. Consider the process $X_n = Z_1 + \ldots + Z_n,\ n = 0, 1, \ldots$, with the initial distribution $\mathbf{P}(X_0 = 0) = 1$, i.e., the initial state is 0. The one-step transition probabilities are

$$\begin{aligned} p_{ij}(n, n+1) &= \mathbf{P}(X_{n+1} = j \mid X_n = i) \\ &= \mathbf{P}(Z_1 + \ldots + Z_{n+1} = j \mid Z_1 + \ldots + Z_n = i) \\ &= \mathbf{P}(Z_{n+1} = j - i) = \begin{cases} p_{j-i}, & \text{ha } j \geq i, \\ 0, & \text{ha } j < i. \end{cases} \end{aligned}$$

This means that the process X is a homogeneous Markov chain with the transition probability matrix

$$\Pi = \begin{bmatrix} p_0 & p_1 & p_2 & p_3 & p_4 & \cdots \\ 0 & p_0 & p_1 & p_2 & p_3 & \cdots \\ 0 & 0 & p_0 & p_1 & p_2 & \cdots \\ 0 & 0 & 0 & p_0 & p_1 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Example 3.10. Now let

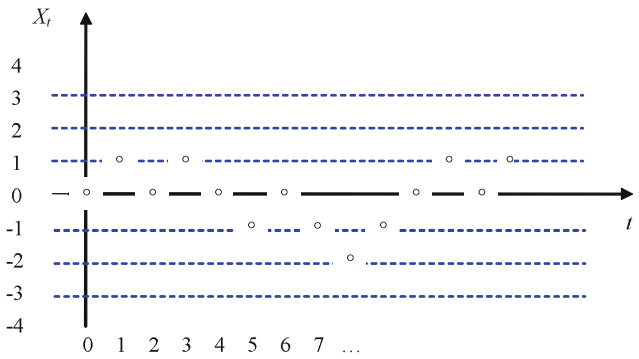


Fig. 3.1 Random walk

$$\mathbf{P}(Z_i = +1) = p,\ \mathbf{P}(Z_i = -1) = 1 - p \quad (0 < p < 1),\ i = 1, 2, \ldots,$$

be the common distribution function of a sequence of independent random variables $Z_0, Z_1, \ldots$, and define the process $X_n = Z_1 + \ldots + Z_n$, $n = 1, 2, \ldots$. Let $\mathbf{P}(X_0 = 0) = 1$ be the initial distribution of the process X. Then the process X is a homogeneous Markov chain with initial state $X_0 = 0$ and transition probability matrix

$$\begin{aligned} p_{ij}(n, n+1) &= \mathbf{P}(X_{n+1} = j \mid X_n = i) \\ &= \mathbf{P}(Z_1 + \ldots + Z_{n+1} = j \mid Z_1 + \ldots + Z_n = i) \\ &= \mathbf{P}(Z_{n+1} = j - i) = \begin{cases} p, & \text{if } j = i + 1, \\ 1 - p, & \text{if } j = i - 1, \\ 0, & \text{if } |i - j| \neq 1. \end{cases} \end{aligned}$$

The process X describes the random walk on the number line starting from the origin and moves at every step one unit to the right with probability p and to the left with probability $(1 - p)$, with these moves being independent of each other. The case $p = 1/2$ corresponds to the symmetric random walk.

Figure 3.1 demonstrates the transitions of the random walk, while Fig. 3.2 shows the transitions of a Markov chain with a finite state space.

3.1.2 The m-Step Transition Probabilities

Let X be a DTMC with discrete state space $\mathcal{X}$. Denote by

$$p_{ij}(s, t) = \mathbf{P}(X_t = j \mid X_s = i)$$

the transition probabilities of X and by

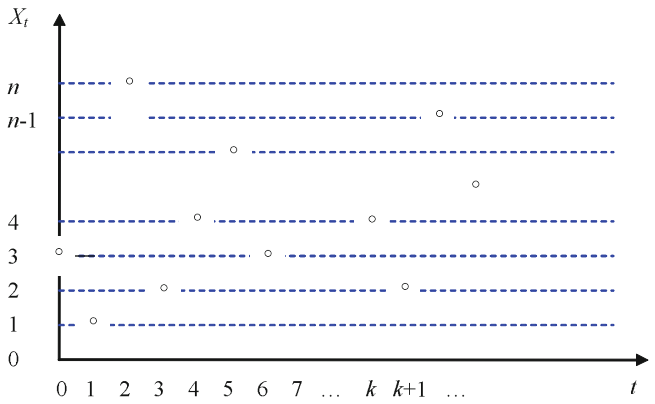


Fig. 3.2 Markov chain with finite state space

$$\Pi(s,t) = \left[p_{ij}(s,t)\right]_{i,j\in\mathcal{X}},\ i,j \in \mathcal{X} \text{ and } 0 \le s \le t < \infty$$

the transition probability matrices. We set for $s = t$

$$p_{ij}(s,s) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \ne j. \end{cases}$$

If the Markov chain X is homogeneous, then the transition probability $p_{ij}(s,t)$ depends only on the difference $t - s$. Thus, using the notation $t = s + m$, we have

$$p_{ij}(s, s+m) = p_{ij}(m),\ s, m = 0, 1, \ldots,\ i, j \in \mathcal{X}.$$

Definition 3.11. The quantities $p_{ij}(m)$, $m = 0, 1, \ldots$, $i, j \in \mathcal{X}$ are called the **m-step transition probabilities** of the Markov chain X, and the matrix $\Pi(m) = \left[p_{ij}(m)\right]_{i,j\in\mathcal{X}}$ associated with them is called an **m-step transition probability matrix**.

Theorem 3.12 (*Chapman–Kolmogorov equation). For every nonnegative integer number r, s, the $(r + s)$-step transition probabilities of the homogeneous Markov chain satisfy the equation*

$$p_{ij}(r+s) = \sum_{k\in\mathcal{X}} p_{ik}(r)\, p_{kj}(s). \tag{3.4}$$

Proof. Assume the initial state of the process is i, that is, the process starts from the state i at the time point 0. First we note that the relation

$$p_{ik}(r) = \mathbf{P}\left(X_r = k \mid X_0 = i\right) = \frac{\mathbf{P}\left(X_0 = i, X_r = k\right)}{\mathbf{P}\left(X_0 = i\right)} = 0$$

holds for some state k if and only if $\mathbf{P}(X_0 = i, X_r = k) = 0$. On the other hand, since $\{X_r = k\},\ k \in \mathcal{X}$ form a complete system of events, $\sum_{k\in\mathcal{X}} \mathbf{P}(X_r = k) = 1$, and, in accordance with the definitions of the $(r+s)$-step transition probability and the conditional probability, we obtain

$$\begin{aligned} p_{ij}(r+s) &= \mathbf{P}(X_{r+s} = j \mid X_0 = i) \\ &= \frac{\mathbf{P}(X_{r+s} = j,\ X_0 = i)}{\mathbf{P}(X_0 = i)} = \sum_{k\in\mathcal{X}} \frac{\mathbf{P}(X_{r+s} = j, X_0 = i, X_r = k)}{\mathbf{P}(X_0 = i)} \\ &= \sum_{k\in\mathcal{X}} \mathcal{I}_{\{p_{ik}\neq 0\}} \frac{\mathbf{P}(X_0 = i, X_r = k)}{\mathbf{P}(X_0 = i)} \frac{\mathbf{P}(X_{r+s} = j, X_0 = i, X_r = k)}{\mathbf{P}(X_0 = i, X_r = k)} \\ &= \sum_{k\in\mathcal{X}} \mathcal{I}_{\{p_{ik}\neq 0\}} \mathbf{P}(X_r = k \mid X_0 = i) \mathbf{P}(X_{r+s} = j \mid X_r = k, X_0 = i) \\ &= \sum_{k\in\mathcal{X}} \mathcal{I}_{\{p_{ik}\neq 0\}} p_{ik}(0,r)\, p_{kj}(r, r+s) = \sum_{k\in\mathcal{X}} p_{ik}(r)\, p_{kj}(s). \end{aligned}$$

□

If we use the matrix notation $\Pi(s,t) = \left[p_{ij}(s,t)\right]_{i,j\in\mathcal{X}}$, then the Chapman–Kolmogorov equation can be rewritten in the matrix form

$$\Pi(s,t) = \Pi(s,r)\Pi(r,t),$$

where s, r, t, and n are integer numbers satisfying the inequality $0 \le s \le r \le t, n \ge 1$. Successively repeating this relation we have

$$\Pi(0,n) = \Pi(0,1)\Pi(1,n) = \ldots = \Pi(0,1)\Pi(1,2)\cdot\ldots\cdot\Pi(n-1,n).$$

Consequently, the m-step transition probability matrix of a homogeneous Markov chain can be given in the form

$$\Pi(m) = \Pi^m,$$

where $\Pi = \Pi(0,1)$ is the (one-step) transition probability matrix of the Markov chain.

3.1.3 Classification of States of Homogeneous Markov Chains

The behavior of a Markov chain and its asymptotic properties essentially depend on the transition probabilities, which reflect the connections among the different states.

Denote by $P_i(t) = \mathbf{P}(X_t = i), i \in \mathcal{X}$ the distribution of the Markov chain X at the time $t \ge 0$. One of the most important questions in the theory of Markov chains

concerns the conditions under which a limit distribution exists for all initial states $X_0 = k \in \mathcal{X}$,

$$\lim_{t\to\infty} P(t) = \pi = (\pi_i, i \in \mathcal{X}),$$

of the time-dependent distribution $P(t) = (P_i(t), i \in \mathcal{X})$, where $\pi_i \geq 0$, $\sum_{i\in\mathcal{X}} \pi_i = 1$. In the answer to this question, the arithmetic properties of the transition probabilities play an important role.

To demonstrate this fact, consider the case where the sample space $\mathcal{X}$ can be divided into two disjoint (nonempty) sets $\mathcal{X}_1$ and $\mathcal{X}_2$ such that

$$p_{ij} = p_{ji} = 0, \text{ for all } i \in \mathcal{X}_1 \text{ and } j \in \mathcal{X}_2.$$

Obviously, if $X_0 = i_0 \in \mathcal{X}_1$ is the initial state, then the relation $X_t \in \mathcal{X}_1$ is valid for all $t \geq 0$, and in the opposite case, $X_t \in \mathcal{X}_2$ for all $t \geq 0$ holds if the initial state i_0 satisfies the condition $i_0 \in \mathcal{X}_2$. This means that in this case we can in fact consider two Markov chains $(\mathcal{X}_k, (P_i(0),\ i \in \mathcal{X}_k), \Pi_k)$, $k = 1, 2$, that can be investigated independently of each other.

Definition 3.13. The state $j \in \mathcal{X}$ is **accessible** from the state $i \in \mathcal{X}$ (denoted by $i \to j$) if there exists a positive integer m such that $p_{ij}(m) > 0$. If the states $i, j \in \mathcal{X}$ are mutually accessible from each other, then we say that they **communicate** (denoted by $i \longleftrightarrow j$).

$p_{ii}(0) = 1,\ i \in \mathcal{X}$ represents the assumption that "every state is accessible in 0 steps from itself." If the state $j \in \mathcal{X}$ is not accessible from the state $i \in \mathcal{X}$ (denoted by $i \nrightarrow j$), then $p_{ij}(m) = 0,\ m \geq 1$. It is easy to check that $i \longleftrightarrow j$ is an **equivalence relation**: it is reflexive, transitive, and symmetric. Furthermore, if the states i and j do not communicate, then either $p_{ij}(m) = 0,\ m \geq 1$, or $p_{ji}(m) = 0$, $m \geq 1$. If a state i satisfies the condition $p_{ii} = p_{ii}(1) = 1$, then the state i is called **absorbing**. This means that if the process visits an absorbing state at time t, then it remains there forever and no more state transitions occur.

If the state space $\mathcal{X}$ does not consist of the states i and j such that $i \to j$, but $j \nrightarrow i$, then $\mathcal{X}$ can be given as a union of finite or countable disjoint sets

$$\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \ldots,$$

where for every k the states of $\mathcal{X}_k$ communicate, while for every k, n, $k \neq n$, the states of $\mathcal{X}_k$ cannot be accessible from the states of $\mathcal{X}_n$.

Definition 3.14. A set of states is called **irreducible** if all pairs of its elements communicate.

In the theory of Markov chains, irreducible classes play an important role because they can be independently analyzed.

Definition 3.15. A Markov chain is called **irreducible** if all pairs of its states communicate.

Clearly, if a Markov chain is irreducible, then it consists of only one irreducible class of states, that is, for every $i, j \in \mathcal{X}$ there exists an integer $m \geq 1$ (depending on i and j) such that $p_{ij}(m) > 0$.

Definition 3.16. For every i denote by $d(i)$ the greatest common divisor of integer numbers $m \geq 1$ for which $p_{ii}(m) > 0$. If $p_{ii}(m) = 0$ for every m, then we set $d(i) = 0$. Then the number $d(i)$ is called the **period** of the Markov chain. If $d(i)=1$ for every state, then the Markov chain is called **aperiodic**.

Example 3.17 (Periodic Markov chain). Consider the random walk on the number line demonstrated earlier in Example 3.10. Starting from an arbitrary state i we can return to state i with positive probabilities in steps $2, 4, \ldots$ only. It is clear that in this case, $p_{ii}(2k) > 0$ and $p_{ii}(2(k-1)+1) = 0$ for every $i \in \mathcal{X}$ and $k = 1, 2, \ldots$; therefore, $d(i) = 2$. At the same time, the Markov chain is obviously irreducible.

Theorem 3.18. *Let X be a homogeneous Markov chain with state space $\mathcal{X}$, and let $\mathcal{X}' \subset \mathcal{X}$ be a nonempty irreducible class. Then for every $i, j \in \mathcal{X}'$, the periods of i and j are the same, i.e., $d(i) = d(j)$.*

Proof. Let $i, j \in \mathcal{X}', i \neq j$, be two arbitrary states. Since $\mathcal{X}'$ is an irreducible class, there exist $t, s \geq 1$ integers such that the inequalities $p_{ij}(t) > 0$ and $p_{ji}(s) > 0$ hold. From this, by the Chapman–Kolmogorov equation, we obtain

$$p_{ii}(t+s) \geq p_{ij}(t)p_{ji}(s) > 0 \quad \text{and} \quad p_{jj}(t+s) \geq p_{ji}(s)p_{ij}(t) > 0;$$

therefore, the numbers $d(i)$ and $d(j)$ differ from 0. Choose arbitrarily an integer $m \geq 1$ such that $p_{ii}(m) > 0$. Repeatedly applying the Chapman–Kolmogorov equation, we have for any $k \geq 1$

$$p_{jj}(t+s+km) \geq p_{ji}(s)p_{ii}(km)p_{ij}(t) \geq p_{ji}(t)\,(p_{ii}(m))^k\, p_{ij}(s) > 0.$$

Thus by the definition of the period of the state j, $d(j)$ is a divisor of both $(t+s+m)$ and $(t+s+2m)$, and hence it is also a divisor of their difference $(t+s+2m)-(t+s+m) = m$. From this it immediately follows that $d(j)$ is a divisor of every m for which $p_{ii}(m) > 0$, and thus it is a divisor of $d(i)$; therefore, $d(j) \leq d(i)$. Changing the role of i and j we get the reverse inequality $d(j) \geq d(i)$, and consequently $d(j) = d(i)$. □

Notice that from this theorem it follows that the states of an irreducible class have a common period $d(\mathcal{X}')$ called the **period of the class**. As a consequence, we have the following assertion.

Corollary 3.19. *If the Markov chain X is homogeneous and irreducible with state space $\mathcal{X}$, then every state has the same period $d = d(\mathcal{X}) > 0$ and is periodic or aperiodic depending on $d > 1$ or $d = 1$, respectively.*

The main property of the numbers for which the probabilities of returning to a state i in k steps are positive, i.e., $p_{ii}(k) > 0$, is given by the following assertion.

Theorem 3.20. *Let X be a homogeneous irreducible Markov chain with state space $\mathcal{X}$. Then for every state $i \in \mathcal{X}$ there exists an integer M_i such that $p_{ii}(d(i)m) > 0$ if $m \geq M_i$.*

Proof. By the previous theorem, $d(i) \geq 1$. Let $m_1, \ldots, m_L$ be different positive integer numbers such that, on the one hand, $p_{ii}(m_k) > 0$, $1 \leq k \leq L$, and on the other hand, $d(i)$ can be given as the greatest common divisor of integers $m_1, \ldots, m_L$. Then, using the well-known assertion from the number theory that there exists an integer M_i such that for every integer $m \geq M_i$, the equation $md(i) = r_1 m_1 + \ldots + r_L m_L$ has a solution with nonnegative integers $i_1, \ldots, i_L$. Applying this fact and the Chapman–Kolmogorov equation we obtain

$$p_{ii}(md(i)) \geq (p_{ii}(m_1))^{r_1} \cdot \ldots \cdot (p_{ii}(m_L))^{r_L} > 0,$$

and consequently the assertion of the theorem is true. □

Consider now the homogeneous irreducible Markov chain with period $d(\mathcal{X}) > 1$. We show that for the transitions among the states there exists a cyclic property, demonstrated in Example 3.28, of the random walk on a number line: if the walk starts from state 0, then the process can take only even integers in even steps and only odd integers in odd steps. The cyclic property in this case means that after even numbers follow odd numbers and after odd numbers follow even numbers as states. This division of states is generalized subsequently for Markov chains with arbitrary period d.

Let $i_0 \in \mathcal{X}$ be an arbitrarily fixed state, and define the sets

$$\mathcal{X}_k = \{j \in \mathcal{X} : \ p_{i_0 j}(k + md) > 0, \text{ for some } m \geq 0\}, k = 0, 1, \ldots, d-1.$$

That is, $\mathcal{X}_k$ is the set of states that are available from i_0 in $k + md$ $(m = 0, 1, \ldots,)$ steps.

Theorem 3.21. *The sets $\mathcal{X}_1, \ldots, \mathcal{X}_{d-1}$ are disjoint, $\mathcal{X} = \mathcal{X}_0 \cup \ldots \cup \mathcal{X}_{d-1}$, and the Markov chain allows for only the following cyclic transitions among the sets $\mathcal{X}_k$:*

$$\mathcal{X}_0 \to \mathcal{X}_1 \to \ldots \to \mathcal{X}_{d-1} \to \mathcal{X}_0. \tag{3.5}$$

Proof. First we prove that the sets $\mathcal{X}_0, \ldots, \mathcal{X}_{d-1}$ are disjoint and their union is $\mathcal{X}$. In contrast, assume that there exist integers k_1, k_2, m_1, m_2 such that $0 \leq k_1 < k_2 \leq d-1, m_1, m_2 \geq 1$, $p_{i_0 j}(k_1 + m_1 d) > 0$, and $p_{i_0 j}(k_2 + m_2 d) > 0$. Since the Markov chain is irreducible, there exists an integer $K \geq 1$ such that $p_{i i_0}(K) > 0$. Using the Chapman–Kolmogorov equation we have

$$p_{i_0 i_0}(k_1 + m_1 d + K) \geq p_{i_0 j}(k_1 + m_1 d) p_{j i_0}(K) > 0,$$
$$p_{i_0 i_0}(k_2 + m_2 d + K) \geq p_{i_0 j}(k_2 + m_2 d) p_{j i_0}(K) > 0.$$

By the definition of the period d, d is a divisor of both $(k_1 + m_1 d + K)$ and $(k_2 + m_2 d + K)$, thus it is also a divisor of their difference, that is, $(k_2 - k_1) + (m_2 - m_1)d$. Consequently, d is a divisor of the difference $(k_2 - k_1)$, which is a contradiction, because $0 < k_2 - k_1 \leq d - 1$. The irreducibility condition ensures that if all states $i \in \mathcal{X}$ are accessible from the state i_0, then $\mathcal{X} = \mathcal{X}_0 \cup \ldots \cup \mathcal{X}_{d-1}$.

We now verify that for every k, $0 \leq k \leq d - 1$, $i \in \mathcal{X}_k$ and $j \in \mathcal{X}$ such that $p_{ij} > 0$, the relation $j \in \mathcal{X}_K$, $0 \leq K < d$, is true, where

$$K = \begin{cases} k + 1, & \text{if } \ 0 \leq k < d - 1, \\ 0, & \text{if } \ k = d - 1. \end{cases}$$

This property guarantees the transitions between the states in (3.5).

Since $j \in \mathcal{X}_k$, then, by the definition of the sets $\mathcal{X}_k$, there exists an integer $m \geq 0$ such that $p_{i_0 j}(k + md) > 0$. From this, by the use of the Chapman–Kolmogorov equality, we have

$$p_{i_0 \ell}(k + 1 + md) \geq p_{i_0 j}(k + md)\, p_{j\ell} > 0.$$

In view of the fact that

$$k + 1 + md = \begin{cases} K + m\, d, & if \ \ 0 \leq k < d - 1, \\ 0 + (m + 1)\, d, & if \ \ k = d - 1, \end{cases}$$

from the definition of $\mathcal{X}_K$ follows the relation $j \in \mathcal{X}_K$. □

As a consequence of Theorem 3.21, we have the next important corollary, which allows us to consider an aperiodic Markov chain instead of a periodic one.

Corollary 3.22. *Theorem 3.21 states that starting from a state of $\mathcal{X}_k$, $k = 0, 1, \ldots, d - 1$, after exactly d steps the process returns to a state of $\mathcal{X}_k$. If we define the quantities*

$$p_{ij}^{(k)} = \mathbf{P}\,(X_d = i \mid X_0 = j),\ i, j \in \mathcal{X}_k,$$

then $\sum_{j \in \mathcal{X}_k} p_{ij}^{(k)} = 1$, $i \in \mathcal{X}_k$ follows. This means that the matrices $\mathbf{P}^{(k)} = \left[p_{ij}^{(k)}\right]_{i,j \in \mathcal{X}_k}$ are stochastic; they can be interpreted as one-step transition probability matrices, and consequently the processes

$$Y^{(k)} = (Y_0, Y_1, \ldots),\ k = 0, 1, \ldots, d - 1,$$

with the state space $\mathcal{X}_k$ and transition probability matrix $\mathbf{P}^{(k)}$, are homogeneous and irreducible Markov chains, and so, instead of the original chain, d homogeneous irreducible Markov chains can be considered independently.

If the states of the Markov chain are numbered according to the $\mathcal{X}_k$, $k = 0, 1, \ldots, d-1$, sets, then the transition probability matrix has the following structure:

$$
\begin{array}{r|ccccc}
\mathcal{X}_0\{ & 0 & P_{0\to 1} & & & \\
\mathcal{X}_1\{ & & 0 & P_{1\to 2} & & \\
 & & & \ddots & \ddots & \\
\mathcal{X}_{d-2}\{ & & & & 0 & P_{d-2\to d-1} \\
\mathcal{X}_{d-1}\{ & P_{d-1\to 0} & & & & 0
\end{array}
.
$$

3.1.4 Recurrent Markov Chains

We consider the question of what conditions ensure the existence of limit theorems for homogeneous aperiodic Markov chains, that is, under what conditions does there exist the limit distribution $\pi = (\pi_i, i \in \mathcal{X})$, $(\pi_i \geq 0, \sum\limits_{i\in\mathcal{X}} \pi_i = 1)$, such that, independently of the initial distribution $(p_i, i \in \mathcal{X})$, the limit is

$$\lim_{n\to\infty} P_i(n) = \lim_{n\to\infty} \mathbf{P}(X_n = i) = \pi_i,\ i \in \mathcal{X}?$$

To provide an answer to this question, it is necessary to consider some quantities such as the probability and the expected value of returning to a given state of a Markov chain or arriving at a state j from another state i. Let $i, j \in \mathcal{X}$ be two arbitrary states, and introduce the following notations:

$$
\begin{aligned}
T_{ij} &= \inf\{n :\ n > 1,\ X_n = j \mid X_0 = i\},\\
f_{ij}(0) &= 0,\\
f_{ij}(1) &= \mathbf{P}(X_1 = j \mid X_0 = i),\\
f_{ij}(n) &= \mathbf{P}(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j \mid X_0 = i),\ n = 2, 3, \ldots.
\end{aligned}
$$

If $i \neq j$, then the quantities $f_{ij}(n) = \mathbf{P}\{T_{ij} = n\}$ mean the **first hit** (or **first passage**) **probabilities** for the state j from i, which is the probability that starting from the state i at time point 0, the process will be first in the state j during n steps (or in time n). If $i = j$, then the quantity $f_{ii}(n)$ means the **first return probability** in the state i in n steps.

Denote $f_{ij} = \sum\limits_{k=1}^{\infty} f_{ij}(k)$, $i, j \in \mathcal{X}$. Obviously, the quantity f_{ij} means the probability that the Markov chain starts from a state i at time 0 and at some time arrives at the state j, that is, $f_{ij} = \mathbf{P}\{T_{ij} < \infty\}$.

Definition 3.23. A state i is called **recurrent** if the process returns to the state i with probability 1, that is, $f_{ii} = \mathbf{P}\{T_{ii} < \infty\} = 1$. If $f_{ii} < 1$, then the state i is called **transient**.

From the definition it follows that when i is a transient state, then a process with positive probability will never return to the state i. The following theorem describes the connection between the return probabilities and the m-step transition probabilities of a Markov chain in the form of a so-called discrete renewal equation.

Theorem 3.24. *For every* $i, j \in \mathcal{X},\ n = 1, 2, \ldots,$

$$p_{ij}(n) = \sum_{k=1}^{n} f_{ij}(k)\, p_{jj}(n-k). \tag{3.6}$$

Proof. By the definition $p_{jj}(0) = 1$, in the case $n = 1$ we have $p_{ij}(1) = f_{ij}(1)\, p_{jj}(0) = f_{ij}(1)$. Now let $n \geq 2$. Using conditional probability and the Markov property we get

$$\begin{aligned}\mathbf{P}(X_n = j, X_1 = j \mid X_0 = i) &= \mathbf{P}(X_n = j \mid X_1 = j, X_0 = i)\mathbf{P}(X_1 = j \mid X_0 = i) \\ &= p_{jj}(n-1)\, p_{ij}(1) = f_{ij}(1)\, p_{jj}(n-1).\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}&\mathbf{P}(X_n = j, X_k = j, X_m \neq j, 1 \leq m \leq k-1 \mid X_0 = i) \\ &\quad = f_{ij}(k)\, p_{jj}(n-k),\ \ n = 1, 2, \ldots.\end{aligned}$$

On the basis of the last two equations, it follows that

$$\begin{aligned}p_{ij}(n) &= \mathbf{P}(X_n = j, X_1 = j \mid X_0 = i) \\ &\quad + \sum_{k=2}^{n} \mathbf{P}(X_n = j, X_k = j, X_m \neq j, 1 \leq m \leq k-1 \mid X_0 = i) \\ &= f_{ij}(1)\, p_{jj}(n-1) + \sum_{k=2}^{n} f_{ij}(k)\, p_{jj}(n-k),\ \ n = 1, 2, \ldots.\end{aligned}$$

□

The notion of the recurrence of a state is defined by the return probabilities, but the following theorem makes it possible to provide a condition for it with the use of n-step transition probabilities $p_{ii}(n)$ and to classify the Markov chains.

Theorem 3.25. *(a) The state* $i \in \mathcal{X}$ *is recurrent if and only if*

$$\sum_{n=1}^{\infty} p_{ii}(n) = \infty.$$

(b) If i and j are communicating states and i is recurrent, then j is also recurrent.

(c) If a state $j \in \mathcal{X}$ is transient, then for arbitrary $i \in \mathcal{X}$

$$\sum_{n=1}^{\infty} p_{ij}(n) < \infty \quad \textit{and consequently} \quad \lim_{n\to\infty} p_{ij}(n) = 0.$$

Proof. (a) By the definition $p_{ii}(0) = 1$ and using relation (3.6) of the preceding theorem we obtain

$$\sum_{n=1}^{\infty} p_{ii}(n) = \sum_{n=1}^{\infty}\sum_{k=1}^{n} f_{ii}(k)\,p_{ii}(n-k) = \sum_{k=1}^{\infty}\sum_{n=k}^{\infty} f_{ii}(k)\,p_{ii}(n-k)$$
$$= \sum_{k=1}^{\infty} f_{ii}(k)\left(p_{ii}(0) + \sum_{n=1}^{\infty} p_{ii}(n)\right).$$

From this equation, if the sum $\sum_{n=1}^{\infty} p_{ii}(n)$ is finite, then we get

$$f_{ii} = \left(1 + \sum_{n=1}^{\infty} p_{ii}(n)\right)^{-1} \sum_{n=1}^{\infty} p_{ii}(n) < 1;$$

consequently, i is not a recurrent state.
If $\sum_{n=1}^{\infty} p_{ii}(n) = \infty$, then obviously $\lim_{N\to\infty} \sum_{n=1}^{N} p_{ii}(n) = \infty$. Since for all positive integers N the relation

$$\sum_{n=1}^{N} p_{ii}(n) = \sum_{n=1}^{N}\sum_{k=1}^{n} f_{ii}(k)\,p_{ii}(n-k)$$
$$= \sum_{k=1}^{N}\sum_{n=k}^{N} f_{ii}(k)\,p_{ii}(n-k) \le \sum_{k=1}^{N} f_{ii}(k) \sum_{n=0}^{N} p_{ii}(n)$$
$$\le \left(1 + \sum_{n=1}^{N} p_{ii}(n)\right) \sum_{k=1}^{N} f_{ii}(k)$$

holds, from the limit $\sum_{n=1}^{N} p_{ii}(n) \to \infty$

$$1 \ge f_{ii} = \sum_{k=1}^{\infty} f_{ii}(k) \ge \sum_{k=1}^{N} f_{ii}(k) \ge \left(1 + \sum_{k=1}^{N} p_{ii}(k)\right)^{-1} \sum_{k=1}^{N} p_{ii}(k) \to 1,\ N \to \infty$$

follows. Consequently, $f_{ii} = 1$, and thus the state i is recurrent.

(b) Since the states i and j communicate, there exist integers $n, m \geq 1$ such that $p_{ij}(m) > 0$ and $p_{ji}(n) > 0$. By the Chapman–Kolmogorov equation for every integer $k \geq 1$,

$$p_{ii}(m+k+n) \geq p_{ij}(m)p_{jj}(k)p_{ji}(n),$$
$$p_{jj}(m+k+n) \geq p_{ji}(n)p_{ii}(k)p_{ij}(m).$$

From this

$$\sum_{k=1}^{\infty} p_{ii}(k) \geq \sum_{k=1}^{\infty} p_{ii}(m+n+k) \geq p_{ij}(m)p_{ji}(n) \sum_{k=1}^{\infty} p_{jj}(k),$$
$$\sum_{k=1}^{\infty} p_{jj}(k) \geq \sum_{k=1}^{\infty} p_{jj}(m+n+k) \geq p_{ij}(m)p_{ji}(n) \sum_{k=1}^{\infty} p_{ii}(k).$$

Both series $\sum_{k=1}^{\infty} p_{ii}(k)$ and $\sum_{k=1}^{\infty} p_{jj}(k)$ are simultaneously convergent or divergent because $p_{ij}(m) > 0$ and $p_{ji}(n) > 0$; thus, by assertion (a) of the theorem, the states i and j are recurrent or transient at the same time.

(c) Applying the discrete renewal Eq. (3.6) and result (a), assertion (c) immediately follows.

□

Definition 3.26. A Markov chain is called **recurrent** or **transient** if every state is recurrent or transient.

Comment 3.27. *Using the n-step transition probabilities $p_{ii}(n)$, a simple formula can be given for the expected value of the number of returns to a state $i \in \mathcal{X}$. Let $X_0 = i$ be the initial state of the Markov chain. The expected value of the return number is expressed as*

$$\mathbf{E}\left(\sum_{k=1}^{\infty} \mathcal{I}_{\{X_k=i\}} | X_0 = i\right) = \sum_{k=1}^{\infty} \mathbf{E}\left(\mathcal{I}_{\{X_k=i\}} | X_0 = i\right)$$
$$= \sum_{k=1}^{\infty} \mathbf{P}\left(X_k = i \mid X_0 = i\right) = \sum_{k=1}^{\infty} p_{ii}(k).$$

The assertion of Theorem 3.25 can be interpreted in another way: a state $i \in \mathcal{X}$ is recurrent if and only if the expected value of the number of returns equals infinity.

Example 3.28 (Recurrent Markov chain). Consider the random walk process $X = (X_n,\ n = 0, 1, \ldots)$ described in Example 3.10. The process, starting from the origin, at all steps moves one unit to the right with probability p and to the left

with probability $(1-p)$, independently of each other. We have proved earlier that the process X is a homogeneous, irreducible, and periodic Markov chain with period 2. Here we discuss the conditions under which the Markov chain will be recurrent.

By the condition $X_0 = 0$, it is clear that $p_{00}(2k+1) = 0$, $k = 0, 1, \ldots$. The process can return in $2k$ steps to the state 0 only if it moves, in some way, k times to the left and k times to the right, the probability of which is

$$p_{00}(2k) = \binom{2k}{k} p^k (1-p)^k = \frac{(2k)!}{k!k!} [p(1-p)]^k.$$

Using the well-known Stirling's formula, which gives an asymptotic relation for $k!$ as $k \to \infty$ as follows (see p. 616 of [5]):

$$\sqrt{2\pi} k^{k+1/2} \mathrm{e}^{-k} < k! < \sqrt{2\pi} k^{k+1/2} \mathrm{e}^{-k} \left(1 + \frac{1}{4k}\right);$$

then

$$k! \approx \left(\frac{k}{\mathrm{e}}\right)^k \sqrt{2\pi k};$$

and thus we have

$$p_{00}(2k) \approx \left(\frac{2k}{\mathrm{e}}\right)^{2k} \sqrt{2\pi(2k)} \left(\left(\frac{k}{\mathrm{e}}\right)^k \sqrt{2\pi k}\right)^{-2} [p(1-p)]^k = \frac{[4p(1-p)]^k}{\sqrt{\pi k}}.$$

By the inequality between arithmetic and geometrical means, the numerator has an upper bound

$$4\,[p(1-p)] \le 4 \left[\frac{p + (1-p)}{2}\right]^2 = 1,$$

where the equality holds if and only if $p = 1-p$, that is, $p = 1/2$. In all other cases the product is less than 1; consequently, the sum of return probabilities $p_{00}(2k)$ is divergent if and only if $p = 1/2$ (symmetric random walk); otherwise, it is convergent. As a consequence of Theorem 3.25, we obtain that the state 0, and together with it all states of the Markov chain, is recurrent if and only if $p = 1/2$.

Note that a similar result is valid if we consider the random walk with integer coordinates in a plane. It can be verified that only in the case of a symmetric random walk will the state $(0, 0)$ be recurrent, when the probabilities of the movements left-right-up-down are $1/4 - 1/4$. In addition, if a random walk is defined in a similar way in higher-dimensional (≥ 3) spaces, then the Markov chain will no longer be recurrent.

3.2 Fundamental Limit Theorem of Homogeneous Markov Chains

3.2.1 *Positive Recurrent and Null Recurrent Markov Chains*

Let X be a homogeneous Markov chain with the finite ($N < \infty$) or countably infinite ($N = \infty$) state space $\mathcal{X} = \{0, 1, \ldots, N\}$ and (one-step) transition probability matrix $\Pi = \left[p_{ij}\right]_{i \in \mathcal{X}}$. Let $P = (p_i = \mathbf{P}(X_0 = i),\ i \in \mathcal{X})$ be the initial distribution. Denote by $P(n) = (P_i(n) = \mathbf{P}(X_n = i),\ i \in \mathcal{X}), n = 0, 1, \ldots,$ the time-dependent distribution of the Markov chain; then $P(0) = P$.

The main question to be investigated here concerns the conditions under which there exists a limit distribution of m-step transition probabilities

$$\lim_{m \to \infty} p_{ij}(m) = \pi_j, \quad \text{where } \pi_j \geq 0 \text{ and } \sum_{i \in \mathcal{X}} \pi_i = 1$$

and how it can be determined. The answer is closely related to the behavior of the recurrent states i of a Markov chain. Note that the condition of recurrence $f_{ii} = \sum_{k=1}^{\infty} f_{ii}(k) = 1$ does not ensure the existence of a limit distribution. The main characteristics are the expected values of the return times $\mu_i = T_{ii} = \sum_{k=1}^{\infty} k f_{ii}(k)$, and the recurrent states will be classified according to whether or not the μ_i are finite because the condition $\mu_i < \infty,\ i \in \mathcal{X}$, guarantees the existence of a limit distribution.

Definition 3.29. A recurrent state $i \in \mathcal{X}$ is called **positive recurrent** (or **nonnull recurrent**) if the return time has a finite expected value μ_i; otherwise, if $\mu_i = \infty$, then it is called **null recurrent**.

Theorem 3.30. *Let X be a homogeneous, irreducible, aperiodic, and recurrent Markov chain. Then for all states $i, j \in \mathcal{X}$,*

$$\lim_{m \to \infty} p_{ij}(m) = \frac{1}{\mu_j}.$$

Note that this theorem not only gives the limit of the m-step transition probabilities with the help of the expected value of the return times, but it interprets the notion of positive and null recurrence. By definition, a recurrent state j is positive recurrent if $1/\mu_j > 0$ and null recurrent if $1/\mu_j = 0$ (here and subsequently, we write $1/\infty = 0$). The assertion given in the theorem is closely related to the discrete renewal Eq. (3.6), and using it we can prove a limit theorem, as the following lemma shows (see [29] and Chap. XIII of [31]).

Lemma 3.31 (*Erdős, Feller, Pollard*). *Let $(q_i,\ i \geq 0)$ be an arbitrary distribution on the natural numbers, i.e., $q_i \geq 0,\ \sum_{i=0}^{\infty} q_i = 1$. Assume that the distribution $(q_i,\ i \geq 0)$ is not latticed, that is, the greatest common divisor of the indices with*

the probabilities $q_i > 0$ equals 1. *If the sequence $\{v_n,\ n \geq 0\}$, satisfies the discrete renewal equation*

$$v_0 = 1,\ \ v_n = \sum_{k=1}^{n} q_k v_{n-k},\ n \geq 1,$$

then

$$\lim_{n\to\infty} v_n = \frac{1}{\mu},$$

where $\mu = \sum\limits_{k=1}^{\infty} k q_k$ and $\frac{1}{\mu} = 0$ if $\mu = \infty$.

The proof of Theorem 3.30 uses the following result from analysis.

Lemma 3.32. *Assume that the sequence $(q_1, q_2, \ldots)$ of nonnegative real numbers satisfies the condition $\sum_{i=0}^{\infty} q_i = 1$. If the sequence of real numbers $(w_n,\ n \geq 0)$ is convergent, $\lim\limits_{n\to\infty} w_n = w$, then*

$$\lim_{n\to\infty} \sum_{k=0}^{n} q_{n-k} w_n = w.$$

Proof. It is clear that the elements of $\{w_n\}$ are bounded; then there exists a number W such that $|w_n| \leq W,\ n \geq 0$. From the conditions $\lim\limits_{n\to\infty} w_n = w$ and $\sum_{i=0}^{\infty} q_i = 1$ it follows for any $\varepsilon > 0$ that there exist integers $N(\varepsilon)$ and $K(\varepsilon)$ such that

$$|w_n - w| < \varepsilon \quad \text{and} \quad \sum_{k=K(\varepsilon)}^{\infty} q_k < \varepsilon.$$

It is easy to check that for every $n > n(\varepsilon) = \max(N(\varepsilon), K(\varepsilon))$,

$$\begin{aligned}
|w_n - w| &\leq \left| \sum_{k=0}^{n} q_k w_{n-k} - \sum_{k=0}^{n} q_k w \right| \\
&\leq \sum_{k=0}^{n(\varepsilon)} q_k |w_{n-k} - w| + \sum_{k=n(\varepsilon)+1}^{n} q_k |w_{n-k} - w| + \sum_{k=n+1}^{\infty} q_k |w| \\
&\leq \sum_{k=0}^{n(\varepsilon)} q_k \varepsilon + \sum_{k=n(\varepsilon)+1}^{n} q_k (W + |w|) + \sum_{k=n+1}^{\infty} q_k |w| \\
&\leq \varepsilon + \varepsilon(W + |w|) + \varepsilon |w| = \varepsilon(1 + W + 2|w|).
\end{aligned}$$

Since $\varepsilon > 0$ can be chosen freely, we get the convergence $w_n \to w,\ n \to \infty$. □

Proof (Theorem 3.30).

(a) We prove firstly the assertion for the case $i = j$. By the discrete renewal equation

$$p_{ii}(0) = 1, \quad p_{ii}(n) = \sum_{k=1}^{n} f_{ii}(k)\, p_{ii}(n-k), \quad n = 1, 2, \ldots,$$

where the state i is recurrent, $f_{ii} = \sum_{k=1}^{\infty} f_{ii}(k) = 1$ ($f_{ii} \geq 0$). Using the assertion of Lemma 3.31 we have

$$\lim_{n\to\infty} p_{ii}(n) = \frac{1}{\mu_i}.$$

(b) Now let $i \neq j$, and apply Lemma 3.32. Since the Markov chain is irreducible and recurrent, $f_{ij} = \sum_{k=1}^{\infty} f_{ij}(k) = 1$ ($f_{ij} \geq 0$). Then, as $n \to \infty$,

$$\lim_{n\to\infty} p_{ij}(n) = \lim_{n\to\infty} \sum_{k=1}^{n} f_{ij}(k)\, p_{jj}(n-k) = \sum_{k=1}^{\infty} f_{ij}(k) \frac{1}{\mu_j} = \frac{1}{\mu_j}. \qquad \square$$

Similar results can be easily proven for periodic cases. Let X be a homogeneous, irreducible, and recurrent Markov chain with period $d > 1$. Then the state space $\mathcal{X}$ can be decomposed into disjoint subsets $\mathcal{X}_0, \ldots, \mathcal{X}_{d-1}$ [see Eq. (3.21)] such that the Markov chain allows only for cyclic transitions between the states of the sets $\mathcal{X}_i$: $\mathcal{X}_0 \to \mathcal{X}_1 \to \ldots \to \mathcal{X}_{d-1} \to \mathcal{X}_0$. Let $0 \leq k,\ m \leq d-1$ be arbitrarily fixed integers; then, starting from a state $i \in \mathcal{X}_k$, the process arrives at a state of $\mathcal{X}_k$ in exactly

$$\ell = \begin{cases} m-k, & \text{if } k < m, \\ m-k+d, & \text{if } m \leq k, \end{cases}$$

steps. From this follows $p_{ij}(s) = 0$ if $s-1$ is divisible by d.

Theorem 3.33. *Let X be a homogeneous, irreducible, and recurrent Markov chain with period $d > 1$ and $i \in \mathcal{X}_k$, $j \in \mathcal{X}_m$ arbitrarily fixed states. Then*

$$\lim_{n\to\infty} p_{ij}(\ell + nd) = \frac{d}{\mu_j},$$

where $\mu_j = \sum_{k=1}^{\infty} k\ f_{jj}(k) = \sum_{r=1}^{\infty} rd\ f_{jj}(rd)$.

Proof. First assume $k = m$, and consider the transition probabilities $p_{ij}(nd)$ for $i, j \in \mathcal{X}_k$. This is equivalent to the case (see Conclusion 3.22 according to

the cyclic transitions of a Markov chain) where we investigate the Markov chain $\overline{X}$ with the state space $\mathcal{X}_k$ and it has the (one-step) transition probability matrix $\overline{\Pi} = [\overline{p}_{ij}]_{i,j\in\mathcal{X}_k}$, $\overline{p}_{ij} = p_{ij}(d)$, $i, j \in \mathcal{X}_k$. Obviously, the Markov chain $\overline{X}$ that originated from X is a homogeneous, irreducible, recurrent, and aperiodic Markov chain. Using the limit theorem 3.31 we obtain

$$\lim_{n\to\infty} \overline{p}_{ii}(n) = \lim_{n\to\infty} p_{ii}(nd) = \frac{1}{\sum\limits_{k=1}^{\infty} k f_{ii}(kd)} = \frac{d}{\sum\limits_{k=1}^{\infty} kd f_{ii}(kd)} = \frac{d}{\sum\limits_{k=1}^{\infty} k f_{ii}(k)} = \frac{d}{\mu_i},$$

where $f_{ii}(r) = 0$ if $r \neq d, 2d, \ldots$.

Assume now that $k \neq m$. Then $f_{ij}(k) = 0$ and $p_{ij}(k) = 0$ if $k \neq \ell + nd$, $n \geq 0$; moreover, the Markov chain $\overline{X}$ is recurrent because

$$f_{ij} = \sum_{s=1}^{\infty} f_{ij}(s) = \sum_{k=1}^{\infty} f_{ij}(\ell + rd) = 1;$$

then

$$p_{ij}(\ell + nd) = \sum_{k=1}^{\ell+nd} f_{ij}(k) p_{jj}(\ell + nd - k) = \sum_{r=1}^{\ell+nd} f_{ij}(\ell + rd) p_{jj}(rd) \to \frac{d}{\mu_j}, \quad n \to \infty.$$

□

Theorem 3.34. *If the homogeneous Markov chain X is irreducible and has a positive recurrent state $i \in \mathcal{X}$, then all its states are positive recurrent.*

Proof. Let $j \in \mathcal{X}$ be arbitrary. Since the Markov chain is irreducible, there exist integers $s, t > 0$ such that $p_{ij}(s) > 0$, $p_{ji}(t) > 0$. Denote by d the period of the Markov chain. It is clear that $d > 0$ because $p_{ii}(s + t) \geq p_{ij}(s) p_{ji}(t) > 0$. Moreover,

$$p_{ii}(s + nd + t) \geq p_{ij}(s) p_{jj}(nd) p_{ji}(t),$$
$$p_{jj}(s + nd + t) \geq p_{ji}(t) p_{ii}(nd) p_{ij}(s).$$

Applying Theorem 3.33 and taking the limit as $n \to \infty$ we have

$$\frac{1}{\mu_i} \geq p_{ij}(s) \frac{1}{\mu_j} p_{ji}(t), \quad \frac{1}{\mu_j} \geq p_{ij}(s) \frac{1}{\mu_i} p_{ji}(t);$$

thus

$$\frac{1}{\mu_i} \geq p_{ij}(s) p_{ji}(t) \frac{1}{\mu_j} \geq [p_{ij}(s) p_{ji}(t)]^2 \frac{1}{\mu_i}.$$

From the last inequality it immediately follows that when the state i is recurrent, at the same time j is also recurrent. □

Summing up the results derived previously, we can state the following theorem.

Theorem 3.35. *Let X be a homogeneous irreducible Markov chain; then*

1. *All states are aperiodic or all states are periodic with the same period,*
2. *All states are transient or all states are recurrent, and in the latter case*
   - *All are positive recurrent or all are null recurrent.*

3.2.2 Stationary Distribution of Markov Chains

Retaining the notations introduced previously, $P(n) = (P_i(n) = \mathbf{P}(X_n = i),\ i \in \mathcal{X})$ denotes the distribution of a Markov chain depending on the time $n \geq 0$. Then $P(0) = (P_i(0) = p_i,\ i \in \mathcal{X})$ is the initial distribution.

Definition 3.36. Let $\pi = (\pi_i,\ i \in \mathcal{X})$ be a distribution, i.e., $\pi_i \geq 0$ and $\sum_{i \in \mathcal{X}} \pi_i = 1$. π is called a **stationary distribution** of the Markov chain X if by choosing $P(0) = \pi$ as the initial distribution, the distribution of the process does not depend on time, that is,

$$P(n) = \pi,\ n \geq 0.$$

A stationary distribution is also called an **equilibrium distribution** of a chain.

With Markov chains, the main problem is the existence and determination of stationary distributions. Theorem 3.30 deals with the convergence of the sequence of n-step transition probabilities $P(n)$ as $n \to \infty$, and if it converges, then the limit gives the stationary distribution of the chain. The proofs of these results are not too difficult but consist of many technical steps [35, 36], and so we omit them here.

Theorem 3.37. *Let X be a homogeneous, irreducible, recurrent, and aperiodic Markov chain. Then the following assertions hold:*

(A) The limit

$$\pi_i = \lim_{n \to \infty} P_i(n) = \frac{1}{\mu_i},\ i \in \mathcal{X},$$

exists and does not depend on the initial distribution.

(B) If all states are recurrent null states, then the stationary distribution does not exist and $\pi_i = 0$ for all $i \in \mathcal{X}$.

(C) If all states are positive recurrent, then the stationary distribution $\pi = (\pi_i,\ i \in \mathcal{X})$ does exist and $\pi_i = 1/\mu_i > 0$ for all $i \in \mathcal{X}$ and $P(n) \to \pi$, as $n \to \infty$. The stationary distribution is unique and satisfies the system of linear equations

$$\sum_{i \in \mathcal{X}} \pi_i = 1, \tag{3.7}$$

$$\pi_i = \sum_{j \in \mathcal{X}} \pi_j p_{ji},\ i \in \mathcal{X}. \tag{3.8}$$

Comment 3.38. *Since the Markov chain is irreducible, it is enough to require in part (B) the existence of a positive recurrent state because from the existence of a single positive recurrent state and the fact that the Markov chain is irreducible it follows that all states are positive recurrent.*

Equation (3.8) of Theorem 3.37 can be rewritten in the more concise form $\pi = \pi\Pi$, *where* Π *is the one-step transition probability matrix of the chain.*

The initial distribution does not play a role in Eqs. (3.7) and (3.8); therefore, when the stationary distribution π *exists, it does not depend on the initial distribution, only on the transition probability matrix* Π.

Given that the stationary distribution π exists, it can be easily proven that π satisfies the system of linear Eq. (3.8), and at the same time, these circumstances lead to an iterative method of solution [see Eq. (3.9) below]. This iterative procedure to determine the stationary distribution can be applied to chains with finite state spaces.

The time-dependent distribution $P(n) = (P_0(n), P_1(n), \ldots)$ satisfies the equation for all $n = 0, 1, \ldots,$

$$P(n) = P(n-1)\Pi. \tag{3.9}$$

Repeating this equation n times, we have

$$P(n) = P(0)\Pi^n, \; n = 0, 1, \ldots.$$

Since it is assumed that the stationary distribution π exists, we can write

$$\pi = \lim_{n\to\infty} P(n);$$

thus from the equation

$$\lim_{n\to\infty} P(n) = \lim_{n\to\infty} P(n-1)\Pi$$

it follows that

$$\pi = \pi\Pi.$$

Definition 3.39. A state i of an irreducible homogeneous Markov chain X is called **ergodic** if the state i is aperiodic and positive recurrent, i.e., $d(i) = 1, \; \mu_i < \infty$. If all states of the chain are ergodic, then the Markov chain is called ergodic.

Here we define the ergodic property only of Markov chains. This property can be defined for much more complex stochastic processes as well.

By Theorem 3.37, a homogeneous, aperiodic, positive recurrent Markov chain is always ergodic. Since an irreducible Markov chain with finite state space is positive recurrent, the following statement is also true.

Theorem 3.40. *A homogeneous, irreducible, aperiodic Markov chain with finite state space is ergodic.*

In practical applications, the equilibrium distributions of Markov chains play an essential role. In what follows, we give two theorems without proofs whose conditions ensure the existence of the stationary distribution of a homogeneous, irreducible, aperiodic Markov chain X with state space $\mathcal{X} = \{0, 1, \ldots\}$. The third theorem gives an upper bound for the convergence rate to the stationary distribution of the iterative procedure (3.9).

Theorem 3.41 (*Klimov [56]). If there exists a function $g(i)$, $i \in \mathcal{X}$, a state $i_0 \in \mathcal{X}$, and a positive constant ε such that the relations*

$$\mathbf{E}\,(g(X_{n+1}) \mid X_n = i) \le g(i) - \varepsilon, \;\; i \ge i_0, \; n \ge 0,$$
$$\mathbf{E}\,(g(X_{n+1}) \mid X_n = i) < \infty, \;\; i \ge 0, \; n \ge 0,$$

hold, then the chain X is ergodic.

Theorem 3.42 (*Foster [33]). Assume that there exist constants $a, b > 0$ and $\ell \ge 0$ such that the inequalities*

$$\mathbf{E}\,(X_{n+1} \mid X_n = i) \le a, \;\; i \le \ell,$$
$$\mathbf{E}\,(X_{n+1} \mid X_n = i) \le i - b, \;\; i > \ell,$$

are valid. Then the Markov chain X is ergodic.

Theorem 3.43 (*Bernstein [10]). Assume that there exist a state $i_0 \in \mathcal{X}$ and a constant $\lambda > 0$ such that for all $i \in \mathcal{X}$ the inequality $p_{i i_0} \ge \lambda$ holds. Then*

$$\lim_{n\to\infty} p_{ij}(n) = \pi_j, \; i, j \in \mathcal{X},$$

where $\pi = (\pi_i, i \in \mathcal{X})$ denotes the stationary distribution of the Markov chain; moreover,

$$\sum_{i\in\mathcal{X}} \left| p_{ij}(n) - \pi_j \right| \le 2(1-\lambda)^n, \; n \ge 1.$$

3.2.3 Ergodic Theorems for Markov Chains

Let X be a homogeneous irreducible and positive recurrent Markov chain with state space $\mathcal{X} = \{0, 1, \ldots\}$ and i a fixed state. Compute the time and the relative frequencies when the process stays in the state i on the time interval $[0, T]$ as follows:

$$S_i(T) = \sum_{n=0}^{T} \mathcal{I}_{\{X_n = i\}},$$

$$\overline{S}_i(T) = \frac{1}{T} \sum_{n=0}^{T} \mathcal{I}_{\{X_n = i\}} = \frac{1}{T} S_i(T).$$

Let us consider when and in what sense there exists a limit of the relative frequencies $\overline{S}_i(T)$ as $T \to \infty$ and, if it exists, how it can be determined. This problem has, in particular, practical importance when applying simulation methods. To clarify the stochastic background of the problem, we introduce the following notations.

Assume that a process starts at time 0 from the state i. Let $0 = T_0^{(i)} < T_1^{(i)} < T_2^{(i)} < \ldots$ be the sequence of the consecutive random time points when a Markov chain arrives at the state i, that is, $T_k^{(i)}$, $k = 1, 2, \ldots$, are the return time points to the state i of the chain. This means that

$$X(T_n^{(i)}) = i,\ n = 0, 1, \ldots \text{ and } X(k) \neq i, \text{ if } k \neq T_0^{(i)}, T_1^{(i)}, \ldots.$$

Denote by

$$\tau_k^{(i)} = T_k^{(i)} - T_{k-1}^{(i)},\ k = 1, 2, \ldots,$$

the time length between the return time points. Since the Markov chain has the memoryless property, these random variables are independent; moreover, from the homogeneity of the Markov chain it follows that $\tau_n^{(i)}, n \geq 1$, are also identically distributed. The common distribution of these random variables $\tau_n^{(i)}$ is the distribution of the return times from the state i to i, namely, $(f_{ii}(n), n \geq 1)$.

Heuristically, it is clear that when the return time has a finite expected value μ_i, then during the time T the process returns to the state i on average T/μ_i times. This means that the quantity $\overline{S}_i(T)$ fluctuates around the value $1/\mu_i$ and has the same limit as $T \to \infty$. This result can be given in exact mathematical form on the basis of the law of large numbers as follows.

Theorem 3.44. *If X is an ergodic Markov chain, then, with probability* 1,

$$\lim_{T\to\infty} \overline{S}_i(T) = \frac{1}{\mu_i}, \quad i \in \mathcal{X}. \tag{3.10}$$

If the Markov property is satisfied, then not only are the return times independent and identically distributed, but the stochastic behaviors of the processes on the return periods are identical as well. This fact allows us to prove more general results for an ergodic Markov chain as Eq. (3.10).

Theorem 3.45. *Let X be an ergodic Markov chain and $g(i)$, $i \in \mathcal{X}$, be a real-valued function such that $\sum_{i\in\mathcal{X}} \pi_i |g(i)| < \infty$. Then the convergence*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T} g(X_n) = \sum_{i\in\mathcal{X}} \pi_i g(i)$$

is true with probability 1, *where $\pi_i, i \in \mathcal{X}$, denotes the stationary distribution of the Markov chain, which exists under the given condition.*

3.2.4 Estimation of Transition Probabilities

In modeling ergodic Markov chains an important question is to estimate the transition probabilities by the observation of the chain. The relative frequencies give corresponding estimates of the probabilities because by Theorem 3.44 they tend to them with probability 1 under the given conditions. Note that from the heuristic approach discussed previously it follows under quite general conditions that not only can the law of large numbers be derived for the relative frequencies, but the central limit theorems can as well.

Consider now the estimate of transition probabilities with the maximum likelihood method. Let X be an ergodic Markov chain with finite state space $\mathcal{X} = \{0, 1, \dots, N\}$ and with the (one-step) transition probability matrix $\Pi = (p_{ij})_{i,j\in\mathcal{X}}$. Assume that we have an observation of n elements $X_1 = i_1, \dots, X_n = i_n$ starting from the initial state $X_0 = i_0$, and we will estimate the entries of the matrix Π. By the Markov property, the conditional likelihood function can be given in the form

$$\mathbf{P}(X_1 = i_1, \dots, X_n = i_n \mid X_0 = i_0) = p_{i_0 i_1} \dots p_{i_{n-1} i_n}.$$

Denote by n_{ij}, $i, j \in \mathcal{X}$, the number of one-step transitions from the state i to j in the sample path $i_0, i_1, \dots, i_n$, and let $0^0 = 1$, $0/0 = 0$. Then the conditional likelihood function given the $X_0 = i_0$ initial state is

$$L(i_1, \dots, i_n; \Pi \mid i_0) = \prod_{i=0}^{N} \left(\prod_{j=0}^{N} p_{ij}^{n_{ij}} \right). \tag{3.11}$$

Applying the maximum likelihood method, maximize the expression in p_{ij} under the conditions

$$p_{ij} \geq 0,\ i, j \in \mathcal{X},\ \sum_{j\in\mathcal{X}} p_{ij} = 1,\ i \in \mathcal{X}.$$

It is clear that there are no relations between the products playing a role in the parentheses of Eq. (3.11) for different i; therefore, the maximization problem can be solved by means of $N + 1$ different, but similar, optimization problems:

$$\max \left\{ \prod_{j=0}^{N} p_{ij}^{n_{ij}} : p_{ij} \geq 0,\ \sum_{j\in\mathcal{X}} p_{ij} = 1 \right\},\ i = 0, 1, \dots, N.$$

Obviously it is enough to solve it only for one state i since the others can be derived analogously to that one.

Let $i \in \mathcal{X}$ be a fixed state, and denote $n_i = \sum_{j\in\mathcal{X}} n_{ij}$. Apply the Lagrange multiplier method; then for every $m = 0, \dots, N$,

$$\frac{\partial}{\partial p_{im}}\left(\prod_{j=0}^{N} p_{ij}^{n_{ij}} + \lambda(p_{i0} + p_{11} + \ldots + p_{iN} - 1)\right) = \frac{n_{im}}{p_{im}}\prod_{j=0}^{N} p_{ij}^{n_{ij}} + \lambda = 0;$$

consequently, for a constant λ_i we have

$$\frac{n_{im}}{p_{im}} = \lambda \prod_{j=0}^{N} p_{ij}^{-n_{ij}} = \lambda_i,\ m = 0, \ldots, N.$$

From this it follows that the equations

$$n_{im} = \lambda_i p_{im},\ \ m = 0, \ldots, N,$$

hold; then

$$\sum_{m=0}^{N} n_{im} = n_i = \lambda_i \sum_{m=0}^{N} p_{im} = \lambda_i.$$

These relations lead to the conditional maximum likelihood estimates for the transition probabilities p_{im} as follows:

$$\widehat{p}_{im} = \frac{n_{im}}{\lambda_i} = \frac{n_{im}}{n_i},\ 0 \le i, m \le N.$$

It can be verified that these estimates $\widehat{p}_{im}$ converge to p_{im} with probability 1 as $\to \infty$.

3.3 Continuous-Time Markov Chains

Like the case of the DTMCs, we assume that the state space $\mathcal{X}$ is a finite $\{0, 1, \ldots, N\}$ or countably infinite set $\{0, 1, \ldots\}$ and assume that the time parameter varies in $\mathcal{T} = [0, \infty)$. According to the general definition (3.1), a process $X = (X_t,\ t \ge 0)$ is said to be CTMC with state space $\mathcal{X}$ if for every positive integer n and $0 \le t_0 < t_1 < \ldots < t_n, i_0, \ldots, i_n \in \mathcal{X}$, the equation

$$\begin{aligned}&\mathbf{P}\left(X_{t_n} = i_n \mid X_{t_{n-1}} = i_{n-1}, \ldots, X_{t_0} = i_0\right)\\ &= \mathbf{P}\left(X_{t_n} = i_n \mid X_{t_{n-1}} = i_{n-1}\right) = p_{i_{n-1}, i_n}(t_{n-1}, t_n)\end{aligned}$$

holds, provided that a conditional probability exists. The Markov chain X is (time) homogeneous if the transition probability function $p_{ij}(s, t)$ satisfies the condition $p_{ij}(s, t) = p_{ij}(t - s)$ for all $i, j \in \mathcal{X}$, $0 \le s \le t$. Denote by $\Pi(s, t) = \left[p_{ij}(s, t), i, j \in \mathcal{X}\right]$ the transition probability matrix.

In the case of a CTMC the time index $t \in [a,b]$ can take uncountably many values for arbitrary $0 \le a < b < \infty$; therefore, the collection of random variables $X_t,\ t \in (a,b]$, is also uncountable. If we consider the questions in accordance with the sample paths of the chain, then these circumstances can lead to measurability problems (discussed later). However, the Markov processes that will be investigated later are the so-called stepwise processes, and they ensure the necessary measurability property.

We will deal mainly with the part of the theory that is relevant to queueing theory, and we touch upon only some questions in general cases showing the root of the measurability problems. A discussion of jumping processes, which is more general than the investigation of stepwise Markov chains, can be found in [36, Chap. III].

If the Markov chain $\{X_t,\ t \ge 0\}$, is homogeneous, then the transition probability functions $p_{ij}(s,t)$ can be given in a simpler form:

$$p_{ij}(s,s+t) = p_{ij}(t),\ i,j \in \mathcal{X},\ s,t \ge 0,$$

and thus the matrix form of transition probabilities is

$$\Pi(s,s+t) = \Pi(t),\ s,t \ge 0.$$

As was done previously, denote by

$$P(t) = (P_0(t), P_1(t), \ldots),\ t \ge 0,$$

the time-dependent distribution of the chain, where $P_i(t) = \mathbf{P}(X_t = i),\ i \in \mathcal{X}$; then $P(0)$ means the initial distribution, while if there exists a state $k \in \mathcal{X}$ such that $\mathbf{P}(X_0 = k) = 1$, then k is the initial state.

3.3.1 Characterization of Homogeneous Continuous-Time Markov Chains

We now deal with the main properties of homogeneous CTMCs. Similarly to the discrete-time case, the transition probabilities satisfy the following conditions.

(A) $p_{ij}(s) \ge 0,\ s \ge 0,\quad p_{ij}(0) = \delta_{ij},\ i,j \in \mathcal{X}$, where δ_{ij} is the Kronecker δ-function (which equals 1 if $i = j$ and 0 if $i \ne j$).

(B) $\sum\limits_{j \in \mathcal{X}} p_{ij}(s) = 1,\ s \ge 0,\ i \in \mathcal{X}.$

(C) $p_{ij}(s+t) = \sum\limits_{k \in \mathcal{X}} p_{ik}(s) p_{kj}(t),\ s,t \ge 0,\ i,j \in \mathcal{X}.$

An additional condition is needed for our considerations.

(D) The transition probabilities of the Markov chain X satisfy the conditions

$$\lim_{h \to 0+} p_{ij}(h) = p_{ij}(0) = \delta_{ij},\ i,j \in \mathcal{X}. \tag{3.12}$$

Comment 3.46. *Condition (B) expresses that* $\Pi(s),\ s \geq 0$*, is a stochastic matrix. We will not consider the so-called* ***killed Markov chains****, where the lifetime* $[0, \tau]$ *of the chain is random (where the process is defined) and with probability* 1 *is finite, i.e.,* $\mathbf{P}\{\tau < \infty\} = 1$*. It should be noted that condition (B) ensures that the chain is defined on the whole interval* $[0, \infty)$ *because the process will be certainly in some state* $i \in \mathcal{X}$ *for any time* $s \geq 0$*.*

Condition (C) is the Chapman–Kolmogorov equation related to the continuous-time case. It can be given in matrix form as follows:

$$\Pi(s+t) = \Pi(s)\Pi(t),\ s, t \geq 0.$$

Similarly to the discrete-time case, the time-dependent distribution of the chain satisfies the equation

$$P(s+t) = P(s)\Pi(t),\ s, t \geq 0,$$

and thus for all $t > 0$

$$P(t) = P(0)\Pi(t).$$

The last relation means that the initial distribution and the transition probabilities uniquely determine the distribution of the chain at all time points $t \geq 0$*.*

Instead of (D) it is enough to assume that the condition

$$\lim_{h \to 0+} p_{ii}(h) = 1,\ i \in \mathcal{X},$$

holds, because for every $i, j \in \mathcal{X}$*,* $i \neq j$*, the relation*

$$0 \leq p_{ij}(h) \leq \sum_{j \neq i} p_{ij}(h) = 1 - p_{ii}(h) \to 0,\ h \to 0+,$$

is true.

Under conditions (A)–(D), the following relations are valid.

Theorem 3.47. *The transition probabilities* $p_{ij}(t),\ 0 \leq t < \infty,\ i \neq j$*, are uniformly continuous.*

Proof. Using conditions (A)–(D) we obtain

$$\begin{aligned} \left| p_{ij}(t+h) - p_{ij}(t) \right| &= \left| \sum_{k \in \mathcal{X}} p_{ik}(h)\, p_{kj}(t) - \sum_{k \in \mathcal{X}} \delta_{ik}\, p_{kj}(t) \right| \\ &\leq \sum_{k \in \mathcal{X}} \left| p_{ik}(h) - \delta_{ik} \right| p_{kj}(t) \\ &\leq 1 - p_{ii}(h) + \sum_{k \neq i} p_{ik}(h) = 2(1 - p_{ii}(h)) \to 0,\ \ h \to 0+. \end{aligned}$$

□

Theorem 3.48 (*[36, p. 200]*). *For all $i, j \in \mathcal{X}$, $i \neq j$, the finite limit*

$$q_{ij} = \lim_{h \to 0+} \frac{p_{ij}(h)}{h}$$

exists.

For every $i \in \mathcal{X}$ there exists a finite or infinite limit

$$q_i = \lim_{h \to 0+} \frac{1 - p_{ii}(h)}{h} = -p'_{ii}(0).$$

The quantities q_{ij} and q_i are the most important characteristics of a homogeneous continuous-time Markov chain. Subsequently we will use the notation $q_{ii} = -q_i$, $i \in \mathcal{X}$, also and interpret the meaning of these quantities.

Definition 3.49. The quantity q_{ij} is called the **transition rate** of intensity from the state i to the state j, while q_i is called the **transition rate** from the state i.

We classify the states in accordance with whether or not the rate q_i is finite. If $q_i < \infty$, then i is called a **stable state**, while if $q_i = +\infty$, then we say that i is an **instantaneous** state. Note that there exists a Markov chain with the property $q_i = +\infty$ [36, pp. 207–210].

Definition 3.50. A stable noninstantaneous state i is called **regular** if

$$\sum_{i \neq j} q_{ij} = -q_{ii} = q_i,$$

and a Markov chain is **locally regular** if all its states are regular.

Corollary 3.51. *As a consequence of Theorem 3.48, we obtain that locally regular Markov chains satisfy the following asymptotic properties as $h \to 0+$:*

$$\begin{aligned}
&\mathbf{P}\,(X_{t+h} \neq i \mid X_t = i) = q_i h + o\,(h),\\
&\mathbf{P}\,(X_{t+h} = i \mid X_t = i) = 1 - q_i h + o\,(h),\\
&\mathbf{P}\,(X_{t+h} = j \mid X_t = i) = q_{ij} h + o\,(h), \quad j \neq i.
\end{aligned}$$

From Theorem 3.48 it also follows that Markov chains with a finite state space are locally regular because all q_{ij}, $i \neq j$, are finite and, consequently, all q_i are also finite.

The condition

$$q = \sup_{i \in \mathcal{X}} q_i < \infty \tag{3.13}$$

will play an important role in our subsequent investigations. We introduce the notation

$$Q = \left[q_{ij}\right]_{i,j\in\mathcal{X}} = \left[p'_{ij}(0)\right]_{i,j\in\mathcal{X}} = \Pi'(0)$$

for locally regular Markov chains. Recall that

$$\lim_{t\to 0+} \Pi(t) = \Pi(0) = I, \tag{3.14}$$

where I is the identity matrix with suitable dimension.

Definition 3.52. The matrix Q is called a **rate** or **infinitesimal matrix** of a continuous-time Markov chain.

The following assertions hold for all locally regular Markov chains under the initial condition (3.14) [36, pp. 204–206].

Theorem 3.53. *The transition probabilities of a locally regular Markov chain satisfy the Kolmogorov backward differential equation*

$$\Pi'(t) = Q\ \Pi(t),\ t \geq 0 \qquad \textit{(I)}.$$

If condition (3.13) is fulfilled, then the Kolmogorov forward differential equation

$$\Pi'(t) = \Pi(t)\ Q,\ t \geq 0 \qquad \textit{(II)}$$

is valid. Under condition (3.13) differential Eqs. (I) and (II), referred to as first- and second-system Kolmogorov equations, have unique solutions.

3.3.2 Stepwise Markov Chains

The results of Theorem 3.53 are related to the analytical properties of transition probabilities and do not deal with the stochastic behavior of sample paths. In this part we investigate the so-called stepwise Markov chains and their sample paths. We introduce the embedded Markov chain and consider the transition probabilities and holding times. In the remaining part of this chapter we assume that the Markov chain is locally regular and condition (3.13) holds.

Definition 3.54. A Markov chain X is a **jump process** if for any $t \geq 0$ there exists a random time $\Delta = \Delta(t,\omega) > 0$ such that

$$X_s = X_t, \ \text{ if } \ s \in [t, t+\Delta).$$

In the definition, Δ can be the remaining time the process stays at state $X(t)$, and the definition requires that this time be positive.

Definition 3.55. We say that a Markov chain has a **jump** at time $t_0 > 0$ if there exists a monotonically increasing sequence $t_1, t_2, \ldots$ such that $t_n \to t_0,\ n \to \infty$ and at the same time $X_{t_n} \neq X_{t_0},\ n = 1, 2, \ldots$. A Markov chain is called a **stepwise** process if it is a jump process and the number of jumps is finite for all sample paths on all finite intervals $[0, t]$.

It should be noted that a stepwise process is continuous from the right and has a limit from the left at all jumping points.

Denote by $(\tau_0 =)0 < \tau_1 < \tau_2 < \ldots$ the sequence of consecutive jumping points; then all finite time intervals consist, at most, of finite jumping points. Between two jumping points the state of the process does not change, and this time is called the **holding time**.

Definition 3.56. A stepwise Markov chain is called **regular** if the sequence of holding times $\zeta_k = \tau_{k+1} - \tau_k,\ k = 0, 1, \ldots$, satisfies the condition

$$\mathbf{P}\left(\sum_{k=0}^{\infty} \zeta_k = \infty\right) = 1.$$

By the definition of stepwise process, we have

$$X_s \equiv X_{\tau_i},\ s \in [\tau_i, \tau_{i+1}),\ i = 0, 1, \ldots.$$

Denote by $Y_k = X_{\tau_k},\ k = 0, 1, \ldots$, the states at time points where the transitions change, and define for $i \neq j$

$$\pi_{ij} = \begin{cases} \frac{q_{ij}}{q_i}, & \text{if } q_i > 0, \\ 0, & \text{if } q_i = 0. \end{cases} \tag{3.15}$$

In addition, let

$$\pi_{ii} = 1 - \sum_{j \neq i} \pi_{ij}\ . \tag{3.16}$$

By the Markov property, the process $(Y_k,\ k \geq 0)$, is a discrete-time homogeneous Markov chain with the state space $\mathcal{X} = \{0, 1, \ldots\}$ and the transition probabilities

$$\mathbf{P}\,(Y_{n+1} = j \mid Y_n = i) = \pi_{ij},\ ij \in \mathcal{X},\ n \geq 0.$$

The process $(Y_k,\ k \geq 0)$ is called an **embedded Markov chain** of the continuous-time stepwise Markov chain X.

Note that the condition $q_i = 0$ corresponds to the case where i is an absorbing state, and in other cases the holding times for arbitrary state i have an exponential distribution with parameter q_i whose density function is $q_i\ \mathrm{e}^{-q_i x},\ x > 0$.

3.3.3 Construction of Stepwise Markov Chains

The construction derived here gives a method for simulating stepwise Markov chains at the same time. Thus, we construct a CTMC $\{X_t,\ t \geq 0\}$, with initial distribution $P(0) = (P_0(0), P_1(0), \ldots)$ and transition probability matrix $\Pi(t) = \left[p_{ij}(t)\right],\ t \geq 0$, satisfying condition (3.13).

Using notations (3.15) and (3.16), define the random time intervals with length $S_0, S_1, \ldots$, nonnegative random variables $K_0, K_1, \ldots$, taking integer numbers and the random jumping points $\tau_m = S_0 + \ldots + S_{m-1}, m = 1, 2, \ldots$, by the following procedure.

(a) Generate a random variable K_0 with distribution $P(0)$ [i.e., $\mathbf{P}(K_0 = k) = P_k(0),\ k \in \mathcal{X}$] and a random variable S_0 distributed exponentially with parameter q_{K_0} conditionally dependent on K_0. Define $X_t = K_0$ if $0 \leq t < S_0$.
(b) In the mth steps ($m = 1, 2, \ldots$) generate a random variable K_m with distribution $P^{(m)} = (\pi_{K_{m-1},j},\ j \in \mathcal{X})$, and a random variable S_m distributed exponentially with the parameter q_{K_m}. Define $X_t = K_m$ if $\tau_m \leq t < \tau_{m+1},\ m = 0, 1, \ldots$.

Then the stochastic process $\{X_t, t \geq 0\}$ is a stepwise Markov chain with initial distribution $P(0)$ and transition probability matrix $\Pi(t),\ t \geq 0$.

3.3.4 Some Properties of the Sample Path of Continuous-Time Markov Chains

By the considerations of the sample paths of CTMCs, there are problems that cannot arise in the case of discrete-time chains. For example, $X_t,\ t \geq 0$, are random variables; therefore, $\{X_t \leq x\} \in \mathcal{A}$ is an event for all $t \geq 0$ and $x \in \mathbb{R}$. But at the same time, for example, the set

$$\bigcap_{a \leq t < b} \{\omega \in \Omega : X_t(\omega) \leq x\}$$

is not necessarily an event (element of $\mathcal{A}$). This question is closely connected to the separability property of the processes (see, for example, [35, Chap. III]). The root essence is whether a countable and everywhere dense subset $\mathcal{S} \subset [0, \infty)$ exists such that the statistical behavior of the process X can be characterized by a countable set of the random variables $X_t,\ t \in \mathcal{S}$. The notion of separability is given in general by the following definition.

Definition 3.57. A process $X = (X_t,\ t \geq 0)$ is called **separable** if there exists an event $N \in \mathcal{A}$ with probability 0 and a countable subset $\mathcal{S} = \{r_i,\ i = 1, 2, \ldots\}$ of $\mathbb{R}_+ = [0, \infty)$ that is always dense in $\mathbb{R}_+$ such that for any open set $G \subset \mathbb{R}_+$ and for any closed set $F \subset \mathcal{X}$ the sets $\{\omega :\ X_{r_i} \in F,\ r_i \in G\}$ and $\{\omega :\ X_t \in F,\ t \in G\}$ can differ only on the subset of N.

With the help of transition probabilities one can easily give a simple condition that ensures the continuity in probability of the process and, at the same time, the separability property.

Definition 3.58. A stochastic process $(X_t,\ t \geq 0)$ is called **continuous in probability** (or **stochastically**) at the point $t_0 \geq 0$ if for all positive numbers ε the convergence

$$\lim_{t \to t_0} \mathbf{P}\left(|X_t - X_{t_0}| > \varepsilon\right) = 0$$

holds. A process is said to be continuous in probability if it is continuous in probability everywhere.

Theorem 3.59. *If a Markov chain X is locally regular and condition (3.13) is satisfied, then it is continuous in probability.*

Proof. First we check that

$$\delta(h) = \sup_{k \in \mathcal{X}} (1 - p_{kk}(h)) \to 0,\ h \to 0+. \tag{3.17}$$

Since by the relation in [36, p. 201]

$$\frac{1 - p_{kk}(h)}{h} \leq \lim_{h \to 0+} \frac{1 - p_{kk}(h)}{h} = q_k \leq q,$$

then

$$\sup_{k \in \mathcal{X}} (1 - p_{kk}(h)) \leq qh \to 0,\ h \to 0+.$$

It is not difficult to see that for arbitrary $u, h \geq 0$ and $\varepsilon > 0$ we have

$$\begin{aligned}
\mathbf{P}\left(|X_{u+h} - X_u| > \varepsilon\right) &\leq \mathbf{P}\left(|X_{u+h} - X_u| > 0\right) \\
&= \sum_{k \in \mathcal{X}} \mathbf{P}\left(|X_{u+h} - X_u| > 0 \mid X_u = k\right) \mathbf{P}\left(X_u = k\right) \\
&= \sum_{k \in \mathcal{X}} [1 - \mathbf{P}\left(|X_{u+h} - X_u| = 0 \mid X_u = k\right) \mathbf{P}\left(X_u = k\right) \\
&= \sum_{k \in \mathcal{X}} (1 - p_{kk}(h)) \mathbf{P}\left(X_u = k\right) \leq \delta(h) \to 0,\ h \to 0+,
\end{aligned}$$

which means actually the continuity in probability of the chain X. □

Definition 3.60. The stochastic processes $(X_t,\ t \geq 0)$ and $(X'_t,\ t \geq 0)$, given on the same probability space, are said to be equivalent if

$$\mathbf{P}\left(X_t = X'_t\right) = 1,\ t \geq 0.$$

The following theorem ensures that under the condition of continuity in probability, one can consider the separable version of the original process.

Theorem 3.61. *If a process* $(X_t,\ t \geq 0)$ *is continuous in probability, then there exists a continuous-in-probability separable version* $(X'_t,\ t \geq 0)$ *that is stochastically equivalent to* $(X_t,\ t \geq 0)$.

Theorem 3.62. *If a Markov chain satisfies condition (3.13), then there exists a separable and stochastically equivalent version of this Markov chain.*

Proof. From Theorem 3.59 it follows that the Markov chain is continuous in probability; therefore, as a consequence of Theorem 3.61, we have the assertion of the present theorem. □

We assume later on that condition (3.13) is fulfilled because this condition with Theorem 3.62 guarantees that the Markov chain has a stochastically equivalent separable version. Assuming that condition (3.13) holds, one can bypass the measurability problems that can arise in the case of CTMCs, and the holding times are positive for all states.

Theorem 3.63. *If a homogeneous Markov chain* X *satisfies condition (3.13), then* X *has an equivalent stepwise version.*

Proof. Since from condition (3.13) follows Eq. (3.17), then by the use of the theorem of [36, p. 281], we obtain that there exists a stepwise version of the Markov chain that is equivalent to the original Markov chain. □

3.3.5 Poisson Process as Continuous-Time Markov Chain

Theorem 3.64. *Let* $(N_t,\ t \geq 0)$ *be a homogeneous Poisson process with intensity rate* λ, $N_0 = 0$. *Then the process* N_t *is a homogeneous Markov chain.*

Proof. Choose arbitrarily a positive integer n, integers $0 \leq i_1 \leq \ldots \leq i_{n+1}$, and real numbers $t_0 = 0 < t_1 < \ldots < t_{n+1}$. It can be seen that

$$\begin{aligned}
&\mathbf{P}\left(N_{t_{n+1}} = i_{n+1} \mid N_{t_n} = i_n, \ldots, N_{t_1} = i_1\right) \\
&\quad = \frac{\mathbf{P}\left(N_{t_{n+1}} = i_{n+1}, N_{t_n} = i_n, \ldots, N_{t_1} = i_1\right)}{\mathbf{P}\left(N_{t_n} = i_n, \ldots, N_{t_1} = i_1\right)} \\
&\quad = \frac{\mathbf{P}\left(N_{t_{n+1}} - N_{t_n} = i_{n+1} - i_n, \ldots, N_{t_2} - N_{t_1} = i_2 - i_1, N_{t_1} = i_1\right)}{\mathbf{P}\left(N_{t_n} - N_{t_{n-1}} = i_n - i_{n-1}, \ldots, N_{t_2} - N_{t_1} = i_2 - i_1, N_{t_1} = i_1\right)}.
\end{aligned}$$

Since the increments of the Poisson process are independent, the last fraction can be written in the form

$$\begin{aligned}
&\frac{\mathbf{P}\left(N_{t_{n+1}} - N_{t_n} = i_{n+1} - i_n\right) \cdot \ldots \cdot \mathbf{P}\left(N_{t_2} - N_{t_1} = i_2 - i_1\right)\mathbf{P}\left(N_{t_1} = i_1\right)}{\mathbf{P}\left(N_{t_n} - N_{t_{n-1}} = i_n - i_{n-1}\right) \cdot \ldots \cdot \mathbf{P}\left(N_{t_2} - N_{t_1} = i_2 - i_1\right)\mathbf{P}\left(N_{t_1} = i_1\right)} \\
&\quad = \mathbf{P}\left(N_{t_{n+1}} - N_{t_n} = i_{n+1} - i_n\right).
\end{aligned}$$

From the independence of the increments $N_{t_{n+1}} - N_{t_n}$ and $N_{t_n} = N_{t_n} - N_0$ it follows that the events $\{N_{t_{n+1}} - N_{t_n} = i_{n+1} - i_n\}$ and $\{N_{t_n} = i_n\}$ are also independent, and thus

$$\begin{aligned}\mathbf{P}\left(N_{t_{n+1}} - N_{t_n} = i_{n+1} - i_n\right) &= \mathbf{P}\left(N_{t_{n+1}} - N_{t_n} = i_{n+1} - i_n \,|N_{t_n} = i_n\right)\\ &= \mathbf{P}\left(N_{t_{n+1}} = i_{n+1}|N_{t_n} = i_n\right),\end{aligned}$$

and finally we have

$$\mathbf{P}\left(N_{t_{n+1}} = i_{n+1}|N_{t_n} = i_n, \ldots, N_{t_1} = i_1\right) = \mathbf{P}\left(N_{t_{n+1}} = i_{n+1}|N_{t_n} = i_n\right).$$

□

It is easy to determine the rate matrix of a homogeneous Poisson process with intensity λ. Clearly, the transition probability of the process is

$$p_{ij}(h) = \mathbf{P}\left(N_{t+h} = j \mid N_t = i\right) = \mathbf{P}\left(N_h = j - i\right) = \frac{(\lambda h)^{j-i}}{(j-i)!}\mathrm{e}^{-\lambda h}, \quad j \geq i,$$

and

$$p_{ij}(h) \equiv 0, \quad j < i.$$

If $j < i$, then obviously $q_{ij} \equiv 0$. Let now $i \leq j$; then

$$q_{ij} = \lim_{h\to 0+}\frac{p_{ij}(h)}{h} = \lim_{h\to 0+}\frac{1}{h}\frac{(\lambda h)^{j-i}}{(j-i)!}\mathrm{e}^{-\lambda h} = \begin{cases}\lambda, & \text{if } j = i+1,\\ 0, & \text{if } j > i+1.\end{cases}$$

Finally, let $i = j$. By the use of the L'Hospital rule

$$q_i = \lim_{t\to 0+}\frac{1 - p_{ii}(h)}{h} = \lim_{t\to 0+}\frac{1 - \mathrm{e}^{-\lambda h}}{h} = \lambda.$$

Thus, summing up the obtained results, we have the rate matrix

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdot\\ 0 & -\lambda & \lambda & 0 & \cdot\\ 0 & 0 & -\lambda & \lambda & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}. \tag{3.18}$$

The Poisson process is regular because for all $i \in \mathcal{X}$

$$\sum_{j\neq i} q_{ij} = \lambda = q_i < \infty.$$

3.3.6 *Reversible Markov Chains*

Definition 3.65. A discrete-time Markov process is called **reversible** if for every state i, j the equation

$$\pi_i p_{ij} = \pi_j p_{ji}$$

holds, where π_i is the equilibrium probability of the states $i \in \mathcal{X}$.

The equation of the definition is usually called a **local** (or **detailed**) **balance condition** because of its similarity to the (global) balance Eq. (3.8) or, more precisely, to its form

$$\sum_{j \in \mathcal{X}} \pi_i p_{ij} = \sum_{j \in \mathcal{X}} \pi_j p_{ji}, \; i \in \mathcal{X}.$$

The notation of reversibility of Markov chains originates from the fact that if the initial distribution of the chain equals the stationary one, then the forward and reverse conditional transition probabilities are identical, that is,

$$\mathbf{P}(X_n = i \mid X_{n+1} = j) = \mathbf{P}(X_{n+1} = i \mid X_n = j).$$

Indeed,

$$\begin{aligned}
\mathbf{P}(X_n = i \mid X_{n+1} = j) &= \frac{\mathbf{P}(X_n = i, X_{n+1} = j)}{\mathbf{P}(X_{n+1} = j)} \\
&= \frac{\mathbf{P}(X_n = i)\mathbf{P}(X_{n+1} = j \mid X_n = i)}{\mathbf{P}(X_{n+1} = j)} \\
&= \frac{\pi_i p_{ij}}{\pi_j} = \frac{\pi_j p_{ji}}{\pi_j} = p_{ji} \\
&= \mathbf{P}(X_{n+1} = i \mid X_n = j).
\end{aligned}$$

In the case of CTMCs, a definition can be applied analogously to the discrete-time case.

Definition 3.66. A CTMC is called **reversible** if for all pairs i, j of states the equation

$$\pi_i q_{ij} = \pi_j q_{ji}$$

holds, where π_i is the equilibrium probability of the state $i \in \mathcal{X}$.

The reversibility property and the local balance equations are often valid for Markov chains describing the processes in queueing networks (Sect. 10.1); in consequence the equilibrium probabilities can be computed in a simple, so-called **product form**.

3.4 Birth-Death Processes

Definition 3.67. The right-continuous stochastic process $\{\nu(t),\ t \geq 0\}$ is a *birth-death process* if

1. Its set of states is $I = \{0, 1, 2, \ldots\}$ [that is, $\nu(t) \in I$];
2. The sojourn time in the state $k \in I, k > 0$, is exponentially distributed with the parameter
$$\alpha_k = a_k + b_k, \quad a_k, b_k \geq 0, k > 0,$$
and it is independent of the trajectory before arriving at the state k;
3. After the state $k \in I$, $k \geq 1$, the process visits the state $k+1$ with probability $p_k = \dfrac{a_k}{\alpha_k}$ and state $k-1$ with probability $q_k = 1 - p_k = \dfrac{b_k}{\alpha_k}$;
4. For the state 0 we consider the following two cases:
   - The process stays an exponentially distributed amount of time in state 0 with parameter $\alpha_0 = a_0 > 0$ and after that visits state 1 (with probability $p_0 = 1$).
   - Once the process arrives at state 0 it remains there forever ($q_0 = 1$, $p_0 = 0$).

$P_k(0) = \mathbf{P}(\nu(0) = k) = \varphi_k,\ k \in I$, denotes the initial distribution of the process.

If $\{\nu(t),\ t \geq 0\}$ is a birth-death process, then it is an infinite-state continuous-time (time-homogeneous) Markov chain. The parameters a_k and b_k are referred to as the *birth rate* and the *death rate* in the state k, respectively, and k is referred to as the *population*. The special case where $b_k \equiv 0$ is referred to as the *birth process* and where $a_k \equiv 0$ as the *death process*.

Let $T_0 = 0 < T_1 < T_2 < \ldots$ denote the time instants of the population changes (birth and death). The discrete-time $\{\nu_n,\ n \geq 0\}$ process, where $\nu_n = \nu(T_n)$ is the population after the nth change in population [nth jump of $\nu(t)$], is referred to as the Markov chain embedded in the population changes of $\{\nu(t),\ t \geq 0\}$. The state-transition probability matrix of the embedded Markov chain is

$$\begin{bmatrix} q_0 & p_0 & 0 & 0 & 0 & \cdots \\ q_1 & 0 & p_1 & 0 & 0 & \cdots \\ 0 & q_2 & 0 & p_2 & 0 & \cdots \\ 0 & 0 & q_3 & 0 & p_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

3.4.1 Some Properties of Birth-Death Processes

The transient state probability, its Laplace transform, and the initial probabilities for $k \geq 0$, $t \geq 0$, and $Re\ s > 0$ are denoted by

$$P_k(t) = \mathbf{P}(\nu(t) = k), \quad p_k^*(s) = \int_0^\infty \mathrm{e}^{-st} P_k(t)\,\mathrm{d}t, \quad P_k(0) = \mathbf{P}(\nu(0) = k) = \varphi_k.$$

In special cases, the following theorems are true. ([69])

Theorem 3.68. *If $p_0 = 1$, $0 < p_k < 1$, $k \geq 1$, then the following statements hold:*

1. *$P_k(t)$ satisfies the following ordinary differential equations:*

$$P_0'(t) = -a_0 P_0(t) + b_1 P_1(t),$$
$$P_k'(t) = a_{k-1} P_{k-1}(t) - (a_k + b_k) P_k(t) + b_{k+1}(t) P_{k+1}(t), \quad k \geq 1.$$

2. *For φ_k, $k \geq 0$, and $\mathrm{Re}\, s > 0$ the following linear system defines $p_k^*(s)$:*

$$sp_0^*(s) - \varphi_0 = -a_0 p_0^*(s) + b_1 p_1^*(s),$$
$$sp_k^*(s) - \varphi_k = a_{k-1} p_{k-1}^*(s) - (a_k + b_k) p_k^*(s) + b_{k+1} p_{k+1}^*(s), \quad k \geq 1.$$

3. *For $k \geq 0$ the limits*

$$\lim_{t \to \infty} P_k(t) = \pi_k$$

exist and are independent of the initial distribution of the process.

$$\pi_k = 0, \ k \geq 0,$$

if

$$\sum_{k=0}^{\infty} \rho_k < \infty, \tag{3.19}$$

where $\rho_0 = 1$ and $\rho_k = \dfrac{a_0 a_1 \cdots a_{k-1}}{b_1 b_2 \cdots b_k}$, $k \geq 1$. Otherwise, $\pi_k > 0$, $k \geq 0$, and

$$\pi_0 = \left(\sum_{j=0}^{\infty} \rho_j \right)^{-1}, \tag{3.20}$$

$$\pi_k = \rho_k \pi_0. \tag{3.21}$$

Theorem 3.69 (*Finite birth-death process*). *Let the state space of $\nu(t)$ be $\{0, 1, 2, \ldots, n\}$, $p_0 = 1$, $0 < p_k < 1$, for $1 \leq k \leq n-1$ and $p_n = 0$; then the following statements hold:*

1. *$P_k(t)$ satisfies the following ordinary differential equations:*

$$P_0'(t) = -a_0 P_0(t) + b_1 P_1(t),$$
$$P_k'(t) = a_{k-1} P_{k-1}(t) - (a_k + b_k) P_k(t) + b_{k+1} P_{k+1}(t), \quad 1 \leq k \leq n-1,$$
$$P_n'(t) = a_{n-1} P_{n-1}(t) - b_n P_n(t).$$

2. *If the initial distribution of the process is* $\varphi_k = \mathbf{P}(\nu(0) = k),\ 0 \le k \le n$, *then for* $\operatorname{Re} s > 0$ *the Laplace transforms of the transient state probabilities* $p_k^*(s)$ *satisfy*

$$
\begin{aligned}
&sp_0^*(s) - \varphi_0 = -a_0 p_0^*(s) + b_1 p_1^*(s),\\
&sp_k^*(s) - \varphi_k = a_{k-1} p_{k-1}^*(s) - (a_k + b_k) p_k^*(s) + b_{k+1} p_{k+1}^*(s),\ 1 \le k \le n-1,\\
&sp_n^*(s) - \varphi_n = a_{n-1} p_{n-1}^*(s) - b_n p_n^*(s).
\end{aligned}
$$

3. *For* $0 \le k \le n$ *the*

$$\lim_{t\to\infty} P_k(t) = \pi_k > 0$$

limit exists and is independent of the initial distribution:

$$\pi_j = \rho_j \pi_0, \quad \pi_0 = \left(\sum_{j=0}^{\infty} \rho_j \right)^{-1},$$

where

$$\rho_0 = 1, \quad \rho_j = \frac{a_0 a_1 \cdots a_{j-1}}{b_1 b_2 \cdots b_j},\ 1 \le j \le n.$$

Theorem 3.70. *The following equations hold.*

1. *Let* $p_0 = 0,\ 0 < p_k < 1,\ k \ge 1$; *then for* $P_k(t)$ *we have*

$$
\begin{aligned}
&P_0'(t) = b_1 P_1(t),\\
&P_1'(t) = -(a_1 + b_1) P_1(t) + b_2 P_2(t),\\
&P_k'(t) = a_{k-1} P_{k-1}(t) - (a_k + b_k) P_k(t) + b_{k+1} P_{k+1}(t),\ k \ge 2,
\end{aligned}
$$

and for $\operatorname{Re} s > 0$ *and the initial distribution* $\varphi_k,\ k \ge 0$, *we have*

$$
\begin{aligned}
&sp_0^*(s) - \varphi_0 = b_1 p_1^*(s),\\
&sp_1^*(s) - \varphi_1 = -(a_1 + b_1) p_1^*(s) + b_2 p_2^*(s),\\
&sp_k^*(s) - \varphi_k = a_{k-1} p_{k-1}^*(s) - (a_k + b_k) p_k^*(s) + b_{k+1} p_{k+1}^*(s),\ k \ge 2.
\end{aligned}
$$

2. *Let* $\nu(t) \in \{0, 1, 2, \ldots, n\},\ p_0 = 0,\ 0 < p_k < 1$ *if* $1 \le k \le n-1$, *and* $p_n = 0$; *then for* $P_k(t)$ *we have*

$$
\begin{aligned}
&P_0'(t) = b_1 P_1(t),\\
&P_1'(t) = -(a_1 + b_1) P_1(t) + b_2 P_2(t),
\end{aligned}
$$

$$P_k'(t) = a_{k-1}P_{k-1}(t) - (a_k + b_k)P_k(t) + b_{k+1}P_{k+1}(t), \quad 2 \le k \le n-1,$$
$$P_n'(t) = a_{n-1}P_{n-1}(t) - b_n P_n(t),$$

and for $p_k^*(s)$, *Re* $s > 0$, *we have* $[\varphi_k = \mathbf{P}(\nu(0) = k),\ 0 \le k \le n]$

$$sp_0^*(s) - \varphi_0 = b_1 p_1^*(s),$$
$$sp_1^*(s) - \varphi_1 = -(a_1 + b_1)p_1^*(s) + b_2 p_2^*(s),$$
$$sp_k^*(s) - \varphi_k = a_{k-1}p_{k-1}^*(s) - (a_k + b_k)p_k^*(s) + b_{k+1}p_{k+1}^*(s), \quad 2 \le k \le n-1,$$
$$sp_n^*(s) - \varphi_n = a_{n-1}p_{n-1}^*(s) - b_n p_n^*(s).$$

Comment 3.71. *In Theorems 3.68–3.70 the differential equations for* $P_j(t)$ *are indeed the Kolmogorov (forward) differential equations for the given systems. The equations for* $p_j^*(s)$ *can be obtained from the related differential equations for* $P_j(t)$ *using*

$$\int_0^\infty \mathrm{e}^{-st} P_j'(t)\,\mathrm{d}t = sp_j^*(s) - P_j'(0)\,.$$

In Theorem 3.70 state 0 is an absorbing state. In this way, the theorem allows one to compute the parameters of the busy period of birth-death Markov chains starting from state k ($\varphi_k = 1$), where the busy period is the time to reach state 0 (which commonly represents the idle state of a system, where the server is not working, in contrast to the $i > 0$ states, where the server is commonly busy). Let Π_k denote the length of the busy period starting from state k; then

$$\Pi_k(t) = \mathbf{P}(\Pi_k \le t) = \mathbf{P}(\nu(t) = 0) = P_0(t)$$

defines the distribution of the length of the busy period, and from Theorem 3.70.1 we have

$$\Pi_k'(t) = P_0'(t) = b_1 P_1(t),$$

from which the Laplace–Stieltjes transform of the distribution of $\Pi_k(t)$, $\pi_k(s)$, is

$$\pi_k(s) = \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}\Pi_k(t) = \int_0^\infty \mathrm{e}^{-st}\,\Pi_k'(t)\,\mathrm{d}t$$
$$= \int_0^\infty \mathrm{e}^{-st} b_1 P_1(t)\,\mathrm{d}t = b_1 p_1^*(s).$$

If the arrival intensity is constant in all states, i.e., $a_k = \lambda > 0$ ($\forall k \ge 0$), then the arrival process is a Poisson process at rate λ. Further results on the properties of special birth-death processes can be obtained, e.g., in [48].

3.5 Exercises

Exercise 3.1. Compute the probability that a CTMC with the generator matrix $\begin{pmatrix} -1 & 0.5 & 0.5 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{pmatrix}$ stays in state 1 after the second state transition if the initial distribution is $(0.5, 0.5, 0)$.

Exercise 3.2. Compute the stationary distribution of a CTMC with the generator matrix $\begin{pmatrix} -3 & 3 & 0 \\ 4 & -4 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ if the initial distribution is $(0.5, 0, 0.5)$.

Exercise 3.3. Z_n and $Y_n, n = 1, 2, \ldots$, are discrete independent random variables. $\mathbf{P}(Z_n = 0) = 1-p$, $\mathbf{P}(Z_n = 1) = p$ and $\mathbf{P}(Y_n = 0) = 1-q$, $\mathbf{P}(Y_n = 1) = q$. Define the transition probability matrix of the DTMC X_n if

$$X_{n+1} = (X_n - Y_n)^+ + Z_n,$$

where $(x)^+ = \max(x, 0)$. This equation is commonly referred to as the evolution equation of a DTMC.

Exercise 3.4. X_n, $n = 1, 2, \ldots$, is a DTMC with the transition probability matrix $P = \begin{pmatrix} 3/6 & 1/6 & 2/6 \\ 3/4 & 0 & 1/4 \\ 0 & 1/3 & 2/3 \end{pmatrix}$. Compute $\mathbf{E}(X_0 X_1)$ and $corr(X_0, X_1)$ if the initial distribution is $(0.5, 0, 0.5)$ and the state space is $S = \{0, 1, 2\}$.

Exercise 3.5. The generator of a CTMC is defined by

$$q_{0j} = \begin{cases} \frac{1}{3} & \text{if } j = 1, \\ \frac{1}{3} & \text{if } j = 2, \\ -\frac{2}{3} & \text{if } j = 0, \\ 0 & \text{otherwise;} \end{cases} \qquad q_{ij} = \begin{cases} \frac{1}{3i} & \text{if } j = i+1, \\ \frac{1}{3i} & \text{if } j = i+2, \\ -\frac{2}{3i} - \mu & \text{if } j = i, \\ \mu & \text{if } j = i-1, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } i = 1, 2, \ldots.$$

Evaluate the properties of this Markov chain using, e.g., the Foster theorem.

Exercise 3.6. Show examples of

- Reducible
- Periodic (and irreducible)
- Transient (and irreducible)

DTMCs. Evaluate $\lim_{n\to\infty} \mathbf{P}(X_n = i)$ for these DTMCs, where i is a state of the Markov chain.

Exercise 3.7. Two players, A and B, play with dice according to the following rule. They throw the dice, and if the number is 1, then A gets £2 from B; if the number is 2 or 3, then A gets £1 from B; and if the number is greater than 3, then B gets £1 from A. At the beginning of the game both A and B have £3. The game lasts until one of the players can no longer pay. What is the probability that A wins?

Exercise 3.8. Two players, A and B, play with dice according to the following rule. They throw the dice, and if the number is 1, then A gets £2 from B; if the number is 2 or 3, then A gets £1 from B; and if the number is greater than 3, then B gets £1 from A. At the beginning of the game both A and B have £3. If one of them cannot pay the required amount, then he must give all his money to the other player and the game goes on. What is the expected amount of money A will have after a very long run? What is the probability that B will not be able to pay the required amount in the next step of the game after a very long run?

Exercise 3.9. There are two machines, A and B, at a production site. Their failure times are exponentially distributed with the parameters λ_A and λ_B, respectively. Their repair times are also exponentially distributed with the parameters μ_A and μ_B, respectively. A single repairman is associated with the two machines; he can work on only one machine at a time. Compute the probability that at least one of the machines works.

Exercise 3.10. Let $X = (X_0, X_1, \ldots)$ be a two-state Markov chain with the state space $\mathcal{X} = \{0, 1\}$ and with the probability transition matrix $P = \begin{bmatrix} a & 1-a \\ 1-b & b \end{bmatrix}$, where $0 < a, b < 1$. Prove that $P^n = \frac{1}{2-a-b}\Pi + \frac{(a+b-1)^n}{2-a-b}(I - P)$, where $\Pi = \begin{bmatrix} 1-b & 1-a \\ 1-b & 1-a \end{bmatrix}$ and $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Chapter 4
Renewal and Regenerative Processes

4.1 Basic Theory of Renewal Processes

Let $\{N(t),\ t \geq 0\}$ be a nonnegative-integer-valued stochastic process that counts the occurrences of a given event. That is, $N(t)$ is the number of events in the time interval $[0, t)$. For example, $N(t)$ can be the number of bulb replacements in a lamp that is continuously on, and the dead bulbs are immediately replaced (Fig. 4.1).

Let $0 \leq t_1 \leq t_2 \leq \ldots$ be the times of the occurrences of consecutive events and $t_0 = 0$ and $T_i = t_i - t_{i-1}$, $i = 1, 2, 3, \ldots$ be the time intervals between consecutive events.

Definition 4.1. $t_1 \leq t_2 \leq \ldots$ is a **renewal process** if the time intervals between consecutive events $T_i = t_i - t_{i-1}$, $i = 2, 3, \ldots$, are independent and identically distributed (i.i.d.) random variables with CDF

$$F(x) = \mathbf{P}(T_k \leq x), \quad k = 1, 2, \ldots .$$

The nth event time, t_n, $n = 1, 2, \ldots$, is referred to as the n**th renewal point** or renewal time. According to the definition, the first time interval might have a different distribution.

We assume that $F(0) = 0$ and $F(+0) = \mathbf{P}(T_k = 0) < 1$. In this case

$$t_0 = 0, \quad t_n = T_1 + \ldots + T_n, \quad n = 1, 2, \ldots,$$

$$N(0) = 0, \quad N(t) = \sup\{n :\ t_n \leq t,\ n \geq 0\} = \sum_{i=1}^{\infty} \mathcal{I}_{\{t_i \leq t\}}, \quad t > 0.$$

Remark 4.2. $\{N(t),\ t \geq 0\}$ and $\{t_n,\ n \geq 1\}$ mutually and univocally determine each other because for arbitrary $t \geq 0$ and $k \geq 1$ we have

$$N(t) \geq k \quad \Leftrightarrow \quad t_k \leq t.$$

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, DOI 10.1007/978-1-4614-5317-8_4, © Springer Science+Business Media, LLC 2013

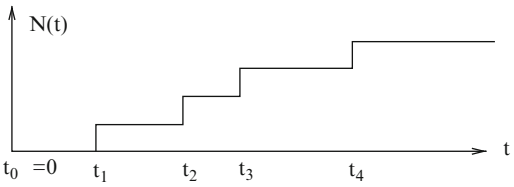


Fig. 4.1 Renewal process

Definition 4.3. When $\mathbf{P}(T_k \le x) = F(x), \quad k = 2, 3, \ldots,$ but $F_1(x) = \mathbf{P}(T_1 \le x) \not\equiv F(x)$, the process is referred to as a **delayed renewal process**.

Remark 4.4. $T_1, T_2, \ldots$ are i.i.d. random variables and from $t_n = T_1 + \ldots + T_n$, and we can compute the distribution of the time of the nth event $F^{(n)}(x) = \mathbf{P}(t_n \le x)$ using the convolution formula

$$F^{(n)}(x) = \int_0^\infty F^{(n-1)}(x-y)\mathrm{d}F(y) = \int_0^x F^{(n-1)}(x-y)\mathrm{d}F(y), \quad n \ge 2, \quad x \ge 0,$$

$$F^{(n)}(x) \equiv 0, \quad \text{if } x \le 0 \text{ and } n \ge 1.$$

Starting from $F^{(1)}(x) = F_1(x)$ the same formula applies in the delayed case.

Definition 4.5. The function $H(t) = \mathbf{E}(N(t)), \quad t \ge 0$, is referred to as a **renewal function**.

One of the main goals of renewal theory is the analysis of the renewal function $H(t)$ and the description of its asymptotic behavior. Below we discuss the related results for regular renewal processes. The properties of delayed renewal processes are similar, and we do not provide details on them here. We will show that the law of large numbers and the central limit theorem hold for the renewal process (see also Ch. 5. in [48]).

Theorem 4.6. *If* $\{T_n, \quad n = 1, 2, \ldots\}$ *is a series of nonnegative i.i.d. random variables and* $\mathbf{P}(T_1 = 0) < 1$, *then there exists* $\rho_0 > 0$ *such that for all* $0 < \rho < \rho_0$ *and* $t \ge 0$

$$\mathbf{E}\left(\mathrm{e}^{\rho N(t)}\right) < \infty$$

holds.

Proof (Proof 1 of Theorem 4.6). From the Markov inequality (Theorem 1.35) we have

$$\begin{aligned}\mathbf{E}\left(\mathrm{e}^{\rho N(t)}\right) &= \sum_{k=0}^{\infty} \mathrm{e}^{\rho k}\mathbf{P}(N(t) = k) \le \sum_{k=0}^{\infty} \mathrm{e}^{\rho k}\mathbf{P}(N(t) \ge k) \\ &= \sum_{k=0}^{\infty} \mathrm{e}^{\rho k}\mathbf{P}(t_k \le t) \le \sum_{k=0}^{\infty} \mathrm{e}^{\rho k}\mathrm{e}^{t-k\kappa} = \mathrm{e}^{t}(1 - \mathrm{e}^{-(\kappa-\rho)})^{-1},\end{aligned}$$

where $\rho < \rho_0 = \kappa$, $\kappa = \log \frac{1}{h}$, and $h = \mathbf{E}\left(\mathrm{e}^{-T_1}\right)$. Additionally, $h < 1$ because $F(0) = 0$ and $\mathbf{P}(T_1 = 0) < 1$. □

Proof (Proof 2 of Theorem 4.6). According to the condition of the theorem, there exist ϵ and δ positive numbers such that $\mathbf{P}(T_k \geq \delta) > \epsilon$. Introducing $\{T'_k = \delta \mathcal{I}_{\{T_k \geq \delta\}},\ k = 1, 2, \ldots\}$ (where T'_n, $n = 1, 2, \ldots$, is a series of i.i.d. random variables) and the related $\{N'(t),\ t \geq 0\}$ renewal process we have that $\mathbf{P}\left(T'_k \leq T_k\right) = 1$, $k \geq 1$, and consequently $\mathbf{P}(N'(t) \geq N(t)) = 1$, $t \geq 0$. The distribution of $N'(t)$ is negative binomial with the parameter $p = \mathbf{P}\left(T'_k \geq \delta\right)$ and order $r = \lfloor t/\delta \rfloor$,

$$\mathbf{P}\left(N'(t) = k + r\right) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k, \quad k = 0, 1, 2, \ldots,$$

from which the statement of the theorem follows. □

Corollary 4.7. *All moments of $N(t)$ $(t \geq 0)$ are finite, and the renewal function $H(t)$ is also finite for all $t \geq 0$.*

Proof. The corollary comes from Theorem 4.6 and the inequality $x^n \leq n!\mathrm{e}^x$, $n \geq 1$, $x \geq 0$. □

Before conducting an analysis of the renewal function we recall some properties of convolution.

Let $A(t)$ and $B(t)$ be monotonically nondecreasing right-continuous functions such that $A(0) = B(0) = 0$.

Definition 4.8. The **convolution** of $A(t)$ and $B(t)$ [denoted by $A * B(t)$] is

$$A * B(t) = \int_0^t B(t - y)\mathrm{d}A(y), \quad t \geq 0.$$

Lemma 4.9. $A * B(t) = B * A(t)$.

Proof. From $B(0) = 0$ we have $B(t - y) = \int_0^{t-y} \mathrm{d}B(s)$, and consequently

$$\begin{aligned} A * B(t) &= \int_0^t \left\{ \int_0^{t-y} \mathrm{d}B(s) \right\} \mathrm{d}A(y) = \int_0^t \int_0^t \mathcal{I}_{\{s < t-y\}} \mathrm{d}A(y)\mathrm{d}B(s) \\ &= \int_0^t \int_0^t \mathcal{I}_{\{y < t-s\}} \mathrm{d}A(y)\mathrm{d}B(s) = \int_0^t \left\{ \int_0^{t-s} \mathrm{d}A(y) \right\} \mathrm{d}B(s) \\ &= B * A(t). \end{aligned}$$

□

Remark 4.10. The definition of the renewal function $H(t)$

$$H(t) = \mathbf{E}\left(N(t)\right) = \mathbf{E}\left(\sum_{i=1}^{\infty} \mathcal{I}_{\{t_i \le t\}}\right) = \sum_{i=1}^{\infty} \mathbf{P}\left(T_1 + \ldots + T_i \le t\right)$$

immediately determines the relation between the renewal function and the order k of the convolutions of the event time distribution

$$H(t) = \sum_{k=1}^{\infty} F^{(k)}(t).$$

Theorem 4.11. *If* $\{T_n,\ n = 1, 2, \ldots\}$ *is a series of i.i.d. random variables and* $\mathbf{P}\left(T_1 < 0\right) = 0,\ \mathbf{P}\left(T_1 = 0\right) < 1$, *then* $H(t)$ *satisfies the* ***renewal equation***

$$H(t) = F(t) + \int_0^t H(t-y)\mathrm{d}F(y),\ \ t \ge 0.$$

Proof. According to Remarks 4.4 and 4.10, the renewal function can be written as

$$\begin{aligned} H(t) &= F^{(1)}(t) + \sum_{k=1}^{\infty} \int_0^t F^{(k)}(t-y)\mathrm{d}F(y) \\ &= F(t) + \int_0^t \left(\sum_{k=1}^{\infty} F^{(k)}(t-y)\right)\mathrm{d}F(y) \\ &= F(t) + \int_0^t H(t-y)\mathrm{d}F(y)\,, \end{aligned}$$

where the order of the summation and the integration are interchanged based on Corollary 4.7. □

In the case of a delayed renewal process, the renewal function is denoted by $H_1(t)$, and the same composition holds as for the regular renewal process (Remark 4.10)

$$H_1(t) = \sum_{k=1}^{\infty} F^{(k)}(t),\ t \ge 0, \qquad (F^{(k)}(t) = \mathbf{P}\left(t_k \le t\right)),$$

but in this case $F_1 \not\equiv F$.

Theorem 4.12. *The renewal function can be written in the following forms:*

$$H_1(t) = F_1(t) + H_1 * F(t) = F_1(t) + F * H_1(t),$$
$$H_1(t) = F_1(t) + H * F_1(t) = F_1(t) + F_1 * H(t),$$
$$H(t) = F(t) + H * F(t) = F(t) + F * H(t).$$

Renewal Equations

Definition 4.13. An integral equation of the type

$$A(t) = a(t) + \int_0^t A(t-x)\mathrm{d}F(x),\ t \geq 0,$$

where $a(t)$ and $F(t)$ are known functions and $A(t)$ is unknown, is referred to as a **renewal equation** (see also Theorem 4.1 of Ch. 5. in [48]).

Theorem 4.14. *If $a(t),\ t \geq 0$, is a bounded real function that is Riemann–Stieltjes integrable according to $H(t)$ over any finite interval, then there uniquely exists the function $A(t),\ t \geq 0$, which is finite over any finite interval and satisfies the renewal equation*

$$(i) \qquad A(t) = a(t) + \int_0^t A(t-x)\mathrm{d}F(x), \quad t \geq 0,$$

and furthermore it satisfies

$$(ii) \qquad A(t) = a(t) + \int_0^t a(t-x)\mathrm{d}H(x), \quad t \geq 0,$$

where $H(t) = \sum_{k=1}^{\infty} F^{(k)}(t),\ t \geq 0$, is the renewal function.

Proof. First we show that the function $A(t)$, $t \geq 0$, defined by equation (ii), is (a) bounded on the $[0, T]$ interval for all $T > 0$ and (b) satisfies (i). Next we prove that (c) all solutions of (i) that are bounded on $[0, T]$ can be given in form (ii), i.e., the solution is unique.

(a) Since $a(t)$ is bounded and $H(t)$ is monotonically nondecreasing, we have

$$\sup_{0\le t\le T}|A(t)| \le \sup_{0\le t\le T}|a(t)| + \int_0^T [\sup_{0\le y\le T}|a(y)|\mathrm{d}H(x)$$
$$\le \sup_{0\le t\le T}|a(t)|(1+H(T)) < \infty.$$

(b) Furthermore, we have

$$A(t) = a(t) + H * a(t) = a(t) + \left(\sum_{k=1}^{\infty} F^{(k)}\right) * a(t)$$
$$= a(t) + F * a(t) + \left(\sum_{k=2}^{\infty} F^{(k)}\right) * a(t)$$
$$= a(t) + F * [a(t) + \left(\sum_{k=1}^{\infty} F^{(k)}\right) * a(t)]$$
$$= a(t) + F * A(t).$$

(c) We prove this by successive approximation. According to equation (i), $A = a + F * A$. Substituting this into (i) we have

$$A = a(t) + F * (a + F * A) = a + F * a + F * (F * A)$$
$$= a + F * a + F^{(2)} * A.$$

Continuously substituting equation (i) we obtain for $n \ge 1$ that

$$A = a + F * a + F^{(2)} * (a + F * A) = \ldots = a + \sum_{k=1}^{n-1}(F^{(k)} * a) + F^{(n)} * A.$$

Since $A(t)$ is bounded on every finite interval according to (a), $F^{(n)}(0-) = 0$, $F^{(n)}(y)$ is monotonically nondecreasing, and $F^{(n)}(t) \to 0,\ n \to \infty$, for all fixed t, we have that for a fixed t

$$|F^{(n)} * A(t)| = \left|\int_0^t A(t-y)\mathrm{d}F^{(n)}(y)\right| \le \sup_{0\le y\le t}|A(t-y)|F^{(n)}(t) \to 0,\ n \to \infty.$$

From the fact that $a(t)$ is bounded it follows that

$$\lim_{n\to\infty}\left(\sum_{k=1}^{n-1} F^{(k)}\right) * a(t) = \left(\sum_{k=1}^{\infty} F^{(k)}\right) * a(t) = H * a(t),$$

and consequently

$$A(t) = a(t) + \lim_{n\to\infty}\left[\left(\sum_{k=1}^{n-1} F^{(k)}\right) * a(t) + F^{(n)}A(t)\right] = a(t) + H * a(t).$$

This means that if A is a bounded solution of (i), then it is identical with (ii).

□

Analysis of the Renewal Function

One of the main goals of the renewal theorem is the analysis of the renewal function. According to Theorem 4.12, in the case of delayed renewal processes the renewal function $H_1(t)$ can be obtained from $F_1(t)$ and $H(t)$. In the rest of this section we focus on the analysis of the renewal function of an ordinary renewal process, $H(t)$, that is, $F_k = F,\ k \geq 1$. During the subsequent analysis we assume that $F(t)$ is such that $F(0-) = 0$ and $F(0+) < 1$.

Theorem 4.15 (*Elementary renewal theorem). There exists the limit*

$$\lim_{t\to\infty} \frac{H(t)}{t} = \frac{1}{\mathbf{E}(T_1)},$$

and it is 0 *if* $\mathbf{E}(T_1) = \infty$.

Definition 4.16. The random variable X has a lattice distribution if there exists $d > 0$ and $r \in R$ such that the random variable $\frac{1}{d}(X - r)$ is distributed on the integer numbers, that is, $\mathbf{P}\left(\frac{1}{d}(X - r) \in \mathbf{Z}\right) = 1$. The largest d with that property is referred to as the step size of the distribution.

Remark 4.17. If X has a lattice distribution with step size d, then

$$d = \min\{s :\ |\psi(2\pi/s)| = 1\},$$

where $\psi(u) = \mathbf{E}\left(\mathrm{e}^{iuX}\right),\ u \in R$, denotes the characteristic function of X. In this case, $|\psi(u)| < 1$ if $0 < |u| < 2\pi/d$. If the distribution of X is not lattice, then $|\psi(u)| < 1$ if $u \neq 0$.

Theorem 4.18 (*Blackwell's theorem). If $F(t)$ is a lattice distribution with step size d, then*

$$\lim_{n\to\infty} q_n = \frac{d}{\mathbf{E}(T_1)},$$

where $q_n = H(nd) - H((n-1)d)$. *If* $F(t)$ *is not a lattice distribution, then for all* $h > 0$

$$\lim_{t\to\infty}(H(t+h) - H(t)) = \frac{h}{\mathbf{E}(T_1)}$$

holds.

The following theorems require the introduction of *direct Riemann integrability*, which is more strict than Riemann integrability.

Let g be a nonnegative function on the interval $[0, \infty)$ and

$$s(\delta) = \delta \sum_{n=1}^{\infty} \inf\{g(x) : \ (n-1)\delta \le x \le n\delta\},$$

$$S(\delta) = \delta \sum_{n=1}^{\infty} \sup\{g(x) : \ (n-1)\delta \le x \le n\delta\}.$$

Definition 4.19. The function g is *directly Riemann integrable* if $s(\delta)$ and $S(\delta)$ are finite for all $\delta > 0$ and

$$\lim_{\delta\to 0}[S(\delta) - s(\delta)] = 0.$$

Remark 4.20. If the function g is directly Riemann integrable, then g is bounded, and the limit of $s(\delta)$ and $S(\delta)$ at $\delta \to 0$ is equal to the infinite Riemann integral, that is,

$$\lim_{\delta\to 0} s(\delta) = \lim_{\delta\to 0} S(\delta) = \int_0^{\infty} g(x)\mathrm{d}x = \lim_{y\to\infty} \int_0^{y} g(x)\mathrm{d}x.$$

Sufficient and necessary conditions for direct Riemann integrability:

(a) There exists $\delta > 0$ such that $S(\delta) < \infty$.
(b) g is almost everywhere continuous along the real axes according to the Lebesgue measure (that is, equivalent to Riemann integrability on every finite interval).

Sufficient conditions for direct Riemann integrability:

g is bounded and has a countable number of discontinuities, and at least either condition (a) or (b) holds:

(a) g equals 0 apart from a finite interval.
(b) g is monotonically decreasing and $\int_0^{\infty} g(x)\mathrm{d}x < \infty$.

Theorem 4.21 (*Smith's renewal theorem). If* $g(x) \ge 0$, $x \ge 0$, *is a nonincreasing directly Riemann integrable function on the interval* $[0, \infty)$, *then for* $t \to \infty$ *one of the following identities holds:*

(a) If F is a nonlattice distribution, then

$$\lim_{t\to\infty} H * g(t) = \lim_{t\to\infty} \int_0^t g(t-u)\mathrm{d}H(u) = \frac{1}{\mathbf{E}(T_1)} \int_0^\infty g(u)\mathrm{d}u.$$

(b) If F is a lattice distribution with step size d, then

$$\lim_{n\to\infty} H*g(x+nd) = \lim_{n\to\infty} \int_0^{x+nd} g(x+nd-u)\mathrm{d}H(u) = \frac{d}{\mathbf{E}(T_1)} \sum_{k=0}^{\infty} g(x+kd).$$

Remark 4.22. Blackwell's theorem (Theorem 4.18) follows from Smith's renewal theorem (Theorem 4.21) assuming that $g(u) = \mathcal{I}_{\{0<u\le h\}}$. The reverse direction is an implicit consequence of the proof of Blackwell's theorem provided by Feller in [31].

Before proving Theorem 4.21 we collect some simple properties of the renewal function $H(t)$.

Lemma 4.23. *H is monotonically nondecreasing and continuous from the right.*

Proof. $F^{(k)}(t)$ is monotonically nondecreasing and continuous from the right for all $k \ge 1$, and the series $\sum_{k=1}^{\infty} F^{(k)}(t)$ is uniformly convergent on every finite interval, from which the lemma follows. □

Lemma 4.24. *The function H is subadditive, that is,*

$$H(t+h) \le H(t) + H(h) \tag{4.1}$$

for $t, h \ge 0$.

Proof. Since $H(0) = 0$, it is enough to consider the case where $t, h > 0$. Let $n(t) = \inf\{n : t_n \ge t,\ n \ge 0\}$. If $t_n \le t$ for all $n \ge 0$, then let $n(t) = \infty$. This case can occur only on a set with measure 0.

Due to the fact that $\mathbf{P}(T_1 = 0)$ might be positive, the relation of $n(t)$ and $N(t)$ is not deterministic. It holds that $n(t) \ge N(t) + 1$ and the right continuity of $N(t)$ implies $N(t_{n(t)}) = N(t),\ t \ge 0$. Using that we have

$$N(t+h) - N(t) = N(t+h) - N(t_{n(t)}) \le N(t_{n(t)} + h) - N(t_{n(t)}),$$

and using the total probability theorem, we obtain

$$\begin{aligned}
\mathbf{E}(N(t+h) - N(t)) &\le \mathbf{E}\left(N(t_{n(t)} + h) - N(t_{n(t)})\right) \\
&= \sum_{k=1}^{\infty} \mathbf{E}\left(N(t_{n(t)} + h) - N(t_{n(t)})|n(t) = k\right)\mathbf{P}(n(t) = k) \\
&= \sum_{k=1}^{\infty} \mathbf{E}(N(t_k + h) - N(t_k)|n(t) = k)\mathbf{P}(n(t) = k).
\end{aligned}$$

Since t_k is a renewal point, the conditional expected value in the last summation does not depend on the condition

$$\mathbf{E}\,(N(t_k+h)-N(t_k)|n(t)=k)=\mathbf{E}\,(N(h)-N(0))=\mathbf{E}\,(N(h)),$$

and in this way we have

$$\begin{aligned}\mathbf{E}\,(N(t+h)-N(t)) &\le \sum_{k=1}^{\infty}\mathbf{E}\,(N(h))\mathbf{P}\,(n(t)=k)\\ &= \mathbf{E}\,(N(h))\sum_{k=1}^{\infty}\mathbf{P}\,(n(t)=k)=\mathbf{E}\,(N(h)),\end{aligned}$$

from which the lemma follows. □

Lemma 4.25. *For the renewal function H the following inequality holds:*

$$H(t)\le H(1)(1+t),\quad t\ge 0. \tag{4.2}$$

Proof. From the previous statement and the monotonicity of H

$$\begin{aligned}H(t)&\le H(\lfloor t\rfloor+1)\le H(1)+H(\lfloor t\rfloor)\le H(1)+(H(1)+H(\lfloor t\rfloor-1))\le\\ &\le\ldots\le H(1)+\lfloor t\rfloor H(1)\le H(1)+tH(1)=H(1)(1+t).\end{aligned}$$

□

Remark 4.26. The nonnegative subadditive functions can be estimated from the preceding expression by a linear function.

Lemma 4.27. *For arbitrary $\lambda>0$ the Laplace–Stieltjes transform $H^{\sim}(\lambda)=\int_0^{\infty}\mathrm{e}^{-\lambda t}\mathrm{d}H(t)$, $\lambda\ge 0$, of the function H can be represented in the Laplace–Stieltjes transform as*

$$H^{\sim}(\lambda)=(1-\varphi^{\sim}(\lambda))^{-1},$$

where $\varphi^{\sim}(\lambda)=\mathbf{E}\left(\mathrm{e}^{-\lambda T_1}\right)$ is the Laplace–Stieltjes transform of the distribution function F.

Proof. For $\lambda>0$ there obviously exists $H^{\sim}(\lambda)$ since, according to Eqs. (1.3) and (4.2),

$$H^{\sim}(\lambda)=\lambda\int_0^{\infty}\mathrm{e}^{-\lambda t}H(t)\mathrm{d}t\le\lambda H(1)\int_0^{\infty}\mathrm{e}^{-\lambda t}(1+t)\mathrm{d}t<\infty.$$

It is clear that

$$\int_0^{\infty}\mathrm{e}^{-\lambda t}\mathrm{d}N(t)=\sum_{k=0}^{\infty}\mathrm{e}^{-\lambda t_k}=1+\sum_{k=1}^{\infty}\prod_{i=1}^{k}\mathrm{e}^{-\lambda T_i}.$$

Using this equality we obtain

$$\mathbf{E}\left(\int_0^\infty \mathrm{e}^{-\lambda t}\mathrm{d}N(t)\right) = \mathbf{E}\left(\lambda\int_0^\infty N(t)\mathrm{e}^{-\lambda t}\mathrm{d}t\right)$$
$$= \lambda\int_0^\infty H(t)\mathrm{e}^{-\lambda t}\mathrm{d}t = \int_0^\infty \mathrm{e}^{-\lambda t}\mathrm{d}H(t) = (h(\lambda) =)$$
$$= \mathbf{E}\left(1 + \sum_{k=1}^\infty\prod_{i=1}^k \mathrm{e}^{-\lambda T_i}\right) = 1 + \sum_{k=1}^\infty(\varphi(\lambda))^k = \frac{1}{1-\varphi(\lambda)},$$

where $0 < \varphi(\lambda) < 1$ if $\lambda > 0$. □

Proof of Elementary Renewal Theorem. First we prove that the limit exists. If $t \geq 1$, then we have that $0 \leq \frac{H(t)}{t} \leq \frac{1+t}{t}H(1) \leq 2H(1)$ is bounded. Let $c = \inf_{t\geq 1}\frac{H(t)}{t}$. Then for arbitrary $\epsilon > 0$ there exists a number $t_0 > 0$ such that

$$\frac{H(t_0)}{t_0} < c + \epsilon.$$

Moreover, for all integers $k \geq 1$ and $\tau \geq 0$

$$\frac{H(kt_0+\tau)}{kt_0+\tau} \leq \frac{kH(t_0)+H(\tau)}{kt_0} \leq c + \epsilon + \frac{H(\tau)}{kt_0},$$

and consequently

$$\limsup_{t\to\infty}\frac{H(t)}{t} \leq c + \epsilon,$$

and

$$c \leq \liminf_{t\to\infty}\frac{H(t)}{t} \leq \limsup_{t\to\infty}\frac{H(t)}{t} \leq c$$

follows. We have proved the existence of the limit.

Using the preceding expression for the Laplace–Stieltjes transform $h(\lambda)$,

$$\int_0^\infty \mathrm{e}^{-\lambda t}H(t)\mathrm{d}t = \frac{1}{\lambda}\int_0^\infty \mathrm{e}^{-\lambda t}\mathrm{d}H(t) = \frac{1}{\lambda}h(\lambda) = \frac{1}{\lambda}\frac{1}{1-\varphi(\lambda)},$$

and we obtain

$$\frac{\lambda}{1-\varphi(\lambda)} = \lambda^2\int_0^\infty \mathrm{e}^{-\lambda t}H(t)\mathrm{d}t = \int_0^\infty \mathrm{e}^{-t}\lambda H\left(\frac{t}{\lambda}\right)\mathrm{d}t. \tag{4.3}$$

By means of the relation for the derivative of the Laplace–Stieltjes transform

$$\lim_{\lambda\to+0}\frac{\lambda}{1-\varphi(\lambda)}=\lim_{\lambda\to+0}\left(\mathbf{E}\left(\frac{1-\mathrm{e}^{-\lambda T_1}}{\lambda}\right)\right)^{-1}=\begin{cases}0, & if\ \ \mathbf{E}\,(T_1)=\infty,\\ \frac{1}{\mathbf{E}(T_1)}, & if\ \ \mathbf{E}\,(T_1)<\infty.\end{cases}$$

On the other hand, in the case $0<\lambda\le 1$, we can give a uniform upper estimation for the integrand in Eq. (4.3):

$$\mathrm{e}^{-t}\lambda H\left(\frac{t}{\lambda}\right)\le \mathrm{e}^{-t}\lambda\left(1+\frac{t}{\lambda}\right)H(1)\le \mathrm{e}^{-t}(1+t)H(1);$$

furthermore,

$$\lim_{\lambda\to+0}\lambda H\left(\frac{t}{\lambda}\right)=t\lim_{\lambda\to+0}\frac{H\left(\frac{t}{\lambda}\right)}{\frac{t}{\lambda}}=tc,$$

so from the Lebesgue majorated convergence theorem

$$\lim_{\lambda\to+0}\int_0^\infty \mathrm{e}^{-t}\lambda H\left(\frac{t}{\lambda}\right)\mathrm{d}t=\int_0^\infty \mathrm{e}^{-t}ct\,\mathrm{d}t=c.$$

Summing up the previous results we obtain

$$c=\lim_{\lambda\to+0}\frac{\lambda}{1-\varphi(\lambda)}=\begin{cases}0 & if\ \ \mathbf{E}\,(T_1)=\infty,\\ \frac{1}{\mathbf{E}(T_1)} & if\ \ \mathbf{E}\,(T_1)<\infty.\end{cases}$$

□

4.1.1 Limit Theorems for Renewal Processes

Theorem 4.28. *Let* $0<\mathbf{E}\,(T_1)=\mu<\infty$*; then the following stochastic convergence holds:*

$$\frac{N(t)}{t}\xrightarrow{P}\frac{1}{\mu},\qquad t\to\infty.$$

Proof. The proof of Theorem 4.28 is based on the relation

$$\{N(t)>k\}=\{t_k\le t\}$$

from Comment 4.2. Let us estimate the probability $\mathbf{P}\,(|N(t)/t-1/\mu|>\epsilon)$ for arbitrary $\epsilon>0$. Let $n=n(t)=\lfloor t/\mu+\epsilon t\rfloor$; then

$$\begin{aligned}\mathbf{P}\left(\frac{N(t)}{t}-\frac{1}{\mu}>\epsilon\right)&=\mathbf{P}\left(N(t)>\frac{t}{\mu}+\epsilon t\right)\le\mathbf{P}\left(N(t)>n\right)\\&=\mathbf{P}\left(t_n\le t\right)=\mathbf{P}\left(\frac{t_n}{n}\le\frac{t}{\lfloor t/\mu+\epsilon t\rfloor}\right)\\&\le\mathbf{P}\left(\frac{t_n}{n}\le\frac{t}{t/\mu+\epsilon t-1}\right)\\&=\mathbf{P}\left(\frac{t_n}{n}\le\frac{1}{1/\mu+\epsilon-1/t}\right)\\&\le\mathbf{P}\left(\frac{t_n}{n}\le\frac{\mu}{1+\mu\epsilon/2}\right)\quad\text{if } t\ge 2/\epsilon,\end{aligned}$$

which by Bernoulli's law of large numbers tends to 0 for the sequence t_n, $n=1,2,\ldots$, as $t\to\infty$. The probability $\mathbf{P}\left(N(t)/t-1/\mu<-\epsilon\right)$ is estimated in a similar way. □

Remark 4.29. By the strong law of large numbers, $\frac{t_k}{k}\to\mu$, $k\to\infty$, with probability 1. Using this fact one can prove that with probability 1

$$\frac{N(t)}{t}\to\frac{1}{\mu},\ \ t\to\infty.$$

The convergence with probability 1 remains valid for delayed renewal processes if the first time interval is finite with probability 1.

Theorem 4.30. *If* $\mathbf{E}\left(T_1\right)=\mu>0$, $\mathbf{D}^2\left(T_1\right)=\sigma^2<\infty$, *then as* $t\to\infty$

$$\lim_{t\to\infty}\mathbf{P}\left(\frac{N(t)-t/\mu}{\sqrt{t\sigma^2/\mu^3}}\le x\right)=\Phi(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x}\mathrm{e}^{-u^2/2}\mathrm{d}u.$$

Proof. Let x be a real number and denote

$$r(t)=\lfloor t/\mu+x\sqrt{t\sigma^2/\mu^3}\rfloor.$$

Note that $r(t)\ge 1$ if $\sqrt{t}+x\sigma/\sqrt{\mu}-\mu/\sqrt{t}\ge 0$. Since $r(t)\to\infty$ as $t\to\infty$, then from the central limit theorem it follows that for all $x\in\mathbb{R}$

$$\mathbf{P}\left(\frac{t_{r(t)}-\mu r(t)}{\sigma\sqrt{r(t)}}\le x\right)\to\Phi(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x}\mathrm{e}^{-u^2/2}du,\ \ t\to\infty.\tag{4.4}$$

Using the relation $\{N(t) \le r(t)\} = \{t_{r(t)} > t\}$ we have

$$\begin{aligned}\mathbf{P}\left(\frac{N(t)-t/\mu}{\sqrt{t\sigma^2/\mu^3}} \le x\right) &= \mathbf{P}\left(N(t) \le t/\mu + x\sqrt{t\sigma^2/\mu^3}\right)\\ &= \mathbf{P}\left(N(t) \le r(t)\right) = \mathbf{P}\left(t_{r(t)} > t\right)\\ &= \mathbf{P}\left(\frac{t_{r(t)}-\mu r(t)}{\sigma\sqrt{r(t)}} > \frac{t-\mu r(t)}{\sigma\sqrt{r(t)}}\right)\\ &= 1 - \mathbf{P}\left(\frac{t_{r(t)}-\mu r(t)}{\sigma\sqrt{r(t)}} \le \frac{t-\mu r(t)}{\sigma\sqrt{r(t)}}\right).\end{aligned}$$

It can be easily checked that

$$\frac{t-\mu r(t)}{\sigma\sqrt{r(t)}} \to -x, \quad t \to \infty,$$

and the continuity of the standard normal distribution function implies the convergence

$$\mathbf{P}\left(\frac{N(t)-t/\mu}{\sqrt{t\sigma^2/\mu^3}} \le x\right) \to 1 - \Phi(-x) = \Phi(x), \quad t \to \infty.$$

The equation $1-\Phi(-x) = \Phi(x)$ follows from the symmetry of the standard normal distribution. □

The following results (without proof) concerning the mean value and variance of the renewal process $N(t)$ are a generalization of previous results and are valid for the renewal processes with delay, too.

Theorem 4.31. *If $\mu_2 = \mathbf{E}\left(T_1^2\right) < \infty$ and T_1 has a nonlattice distribution, then as $t \to \infty$ [31, XIII-12§]*

$$\mathbf{E}\left(N(t)\right) - \frac{t}{\mu} = H(t) - \frac{t}{\mu} \to \frac{\mu_2}{2\mu^2} - 1,$$

$$\mathbf{D}^2\left(N(t)\right) = \frac{\mu_2 - \mu^2}{\mu^3}t + o(t).$$

If, additionally, $\mu_3 = \mathbf{E}\left(T_1^3\right) < \infty$, then [31]

$$\mathbf{D}^2\left(N(t)\right) = \frac{\mu_2 - \mu^2}{\mu^3}t + \left(\frac{5\mu_2^2}{4\mu^4} - \frac{2\mu_3}{3\mu^3} - \frac{\mu_2}{2\mu^2}\right) + o(1).$$

4.2 Regenerative Processes

Many queueing systems can be described by means of regenerative processes. This property makes it possible to prove the limit and stability theorems in order to use the method of simulation.

Definition 4.32. Let T be a nonnegative random variable and $Z(t),\ t \in [0, T)$ be a stochastic process. The pair $(T, Z(t))$, taking on values in the measurable space $(\mathcal{Z}, \mathcal{B})$, is called a cycle of length T.

Definition 4.33. The stochastic process $Z(t),\ t \geq 0$, taking on values in the measurable space $(\mathcal{Z}, \mathcal{B})$, is called a **regenerative process** with moments of regeneration $t_0 = 0 < t_1 < t_2 < \ldots$ if there exists a sequence of independent cycles $(T_k, Z_k(t)),\ k \geq 1$, such that

(1) $T_k = t_k - t_{k-1},\ k \geq 1$;
(2) $\mathbf{P}(T_k > 0) = 1,\ \mathbf{P}(T_k < \infty) = 1$;
(3) All cycles are stochastically equivalent.
(4) $Z(t) = Z_k(t - t_{k-1})$ if $t \in [t_{k-1}, t_k),\ k \geq 1$.

Definition 4.34. If property (3) is fulfilled only starting with the second cycle (analogously to the renewal processes), then we have a **delayed regenerative process**.

Remark 4.35. $t_k,\ k \geq 1$, is a renewal process.

In the case of regenerative processes, an important task is to find conditions assuring the existence and possibility of determining the limit

$$\lim_{t\to\infty} \mathbf{P}(Z(t) \in B),\quad B \in \mathcal{B}.$$

It is also important to estimate the rate of convergence (especially upon examination of the stability problems of queueing systems and simulation procedures).

Let $\{Z(t),\ t \geq 0\}$ be a regenerative process taking on values in the measurable space $(\mathcal{Z}, \mathcal{B})$ with regeneration points $t_0 = 0 < t_1 < t_2 < \ldots, T_n = t_n - t_{n-1},\ n = 1, 2, \ldots$. Assume that $Z(t)$ is right continuous and there exists a limit from the left. Then the cycles $\{T_n,\ \{Z(t_{n-1} + u) :\ 0 \leq u < T_n\}\},\ n = 1, 2, \ldots$, are independent and stochastically equivalent; $\{t_n,\ n \geq 1\}$; and the corresponding counting process $\{N(t),\ t \geq 0\}$ is a renewal process. Let F denote the common distribution of random variables $\{T_n,\ n \geq 1\}$.

The most important application of Smith's theorem is the determination of limit values $\lim_{t\to\infty} \mathbf{E}(W(t))$ for the renewal and regenerative processes, where $W(t) = \Psi(t, N, Z)$ is the function of t, the renewal process N, and the regenerative process Z. The determination of the limit value is based on a more general theorem.

Theorem 4.36. *Let $\{V(t),\ t \geq 0\}$ be a real-valued stochastic process on the same probability space as the process $\{N(t),\ t \geq 0\}$, and for which the mean value $f(t) = \mathbf{E}(V(t))$ is bounded on each finite interval. Let*

$$g(t) = \mathbf{E}\left(V(t)\mathcal{I}_{\{T_1>t\}}\right) + \int_0^t \left[\mathbf{E}(V(t)|T_1 = s) - \mathbf{E}(V(t-s))\right] \mathrm{d}F(s), \quad t \geq 0.$$

Assume that the positive and negative parts of g are directly Riemann integrable. If F is a nonlattice distribution, then

$$\lim_{t\to\infty} f(t) = \lim_{t\to\infty} \mathbf{E}(V(t)) = \frac{1}{\mu}\int_0^\infty g(x)\mathrm{d}x.$$

A similar result is valid if F is a lattice distribution.

Remark 4.37. In the theorem, the property of direct Riemann integrability was required separately for the positive and negative parts of the function g. The reason is that the property is defined only for nonnegative functions.

Proof. It is clear that

$$\begin{aligned} f(t) &= \mathbf{E}\left(V(t)\mathcal{I}_{\{T_1>t\}}\right) + \mathbf{E}\left(V(t)\mathcal{I}_{\{T_1\leq t\}}\right) \\ &= \mathbf{E}\left(V(t)\mathcal{I}_{\{T_1>t\}}\right) + \int_0^t \mathbf{E}(V(t)|T_1 = s)\mathrm{d}F(s). \end{aligned}$$

Let us add and subtract $F * f(t)$; then we get the renewal equation

$$f = g + F * f.$$

The solution of the equation is $f(t) = g + H * g(t)$, which because of the convergence $g(t) \to 0,\ t \to \infty$, and the elementary renewal theorem as a simple consequence of direct Riemann integrability tends to $\frac{1}{\mu}\int_0^\infty g(x)\mathrm{d}x$ as $t \to \infty$. □

Remark 4.38. From the proof it is clear that under the condition of Theorem 4.36 for an arbitrary process $V(t)$ there exists the representation $\mathbf{E}(V(t)) = H * g(t)$ and for the existence of the limit the direct Riemann integrability is required. This representation is interesting if $V(t)$ depends on $Z(t)$.

Special Case Let $h : \mathcal{Z} \to R$ be a measurable function for which, for all t, $\mathbf{E}(|h(Z(t))|) < \infty$. $Z(t)$ is a regenerative process, and the part starting with the second cycle is independent of the first cycle of length T_1, so for arbitrary $0 < s < t$

$$\mathbf{E}((h(Z(t))|T_1 = s)) = \mathbf{E}(h(Z(t-s))).$$

Using the previous notation

$$g(t) = \mathbf{E}\left(h(Z(t))\mathcal{I}_{\{T_1>t\}}\right).$$

Theorem 4.39. *If g_+ and g_- are directly Riemann integrable, then*

$$\begin{aligned}\lim_{t\to\infty} \mathbf{E}\,(h(Z(t))) &= \mu^{-1}\int_0^\infty g(s)\,\mathrm{d}s \\ &= \mu^{-1}\int_0^\infty \mathbf{E}\left(h(Z(s)\mathcal{I}_{\{T_1>s\}}\right)\mathrm{d}s \\ &= \mu^{-1}\mathbf{E}\left(\int_0^{T_1} h(Z(s))\right)\mathrm{d}s.\end{aligned}$$

For arbitrary $A \in \mathcal{B}$ the following equality holds:

$$\begin{aligned}\lim_{t\to\infty} \mathbf{P}\,(Z(t)\in A) &= \mu^{-1}\int_0^\infty \mathbf{P}\,(Z(s)\in A,\ T_1 > s)\,\mathrm{d}s \\ &= \mu^{-1}\mathbf{E}\left(\int_0^{T_1} \mathcal{I}_{\{Z(s)\in A\}}\,\mathrm{d}s\right).\end{aligned}$$

Proof. The first relation follows from the previous theorem, and for the second one it is necessary to mention that, since the trajectories of Z are right continuous and have left limits, the (integrable, bounded) function $\mathbf{P}\,(Z(s)\in A,\ T_1 > s)$ has a countable number of discontinuities and, consequently, is directly Riemann integrable. □

We give one more limit theorem (without proof) that is often useful in practice.

Theorem 4.40. *Let F be a nonlattice distribution, and let at least one of the following conditions be fulfilled:*

(a) $\mathbf{P}\,(Z(t)\in A)$ *is Riemann integrable on an arbitrary finite interval, and* $\mu = \int_0^\infty x\,\mathrm{d}F(x) < \infty$ *holds.*

(b) Starting with a certain integer $n \geq 1$ *the distribution functions defined by* $F^{(1)} = F$, $F^{(n+1)} = F^{(n)} * F$, *are absolute continuous and* $\mu = \int_0^\infty x\,\mathrm{d}F(x) < \infty$.

Then the following relation holds:

$$\lim_{t\to\infty} \mathbf{P}\,(Z(t) \in A) = \mu^{-1} \int_0^\infty \mathbf{P}\,(Z(s) \in A, T_1 > s)\,\mathrm{d}s$$

$$= \mu^{-1}\mathbf{E}\left(\int_0^{T_1} \mathcal{I}_{\{Z(s)\in A\}}\,\mathrm{d}s\right).$$

Example 4.41. Let us consider the renewal process $\{N(t),\ t \geq 0\}$; the renewal moments are

$$t_0 = 0,\ \ t_n = T_1 + T_2 + \ldots + T_n,\ \ n \geq 1,$$

and, furthermore, $\mathbf{P}\,(T_k \leq x) = F(x),\ k \geq 1$, $\mu = \int_0^\infty x\,\mathrm{d}F(x)$. For arbitrary $t > 0$ we define

$$\begin{array}{ll} \delta(t) = t - t_{N(t)}, & \text{the age,} \\ \gamma(t) = t_{N(t)+1} - t, & \text{the residual lifetime,} \\ \beta(t) = \gamma(t) - \delta(t) = t_{N(t)+1} - t_{N(t)}, & \text{the total lifetime.} \end{array}$$

(For example, at instant t, $\delta(t)$ indicates how much time passed without a car arriving at the station, and $\gamma(t)$ indicates how long it was necessary to wait till the arrival of the next car, on the condition that the interarrival times are i.i.d. random variables with the common distribution function F.)

Theorem 4.42. *$\{\delta(t),\ t \geq 0\}$ and $\{\gamma(t),\ t \geq 0\}$ are regenerative processes, and in the case of the nonlattice distribution F,*

$$\lim_{t\to\infty} \mathbf{P}\,(\delta(t) \leq x) = \lim_{t\to\infty} \mathbf{P}\,(\gamma(t) \leq x) = \frac{1}{\mu}\int_0^x (1 - F(u))\,\mathrm{d}u,$$

$$\lim_{t\to\infty} \mathbf{P}\,(\beta(t) \leq x) = \frac{1}{\mu}\int_0^x s\,\mathrm{d}F(s).$$

Proof. Both processes are obviously regenerative with common regeneration points $t_n,\ n \geq 1$. By our previous theorem,

$$\lim_{t\to\infty} \mathbf{P}\,(\delta(t) \leq x) = \frac{1}{\mu}\int_0^\infty \mathbf{P}\,(\delta(s) \leq x, T_1 > s)\,\mathrm{d}s;$$

furthermore,

$$\mathbf{P}\,(\delta(s) \leq x, T_1 > s) = \mathbf{P}\,(s \leq x, T_1 > s) = \begin{cases} 1 - F(s), & \text{if } s < x, \\ 0, & \text{if } s \geq x, \end{cases}$$

so

$$\lim_{t\to\infty} \mathbf{P}\left(\delta(t) \le x\right) = \frac{1}{\mu}\int_0^x (1 - F(s))\mathrm{d}s = \frac{1}{\mu}\int_0^x (1 - F(s))\mathrm{d}s$$

using the identity $\mu = \int_0^\infty (1 - F(s))\mathrm{d}s$ (Exercise 1.5). Similarly, for the process $\{\gamma(t),\ t \ge 0\}$ we obtain

$$\begin{aligned}
&\lim_{t\to\infty} \mathbf{P}\left(\gamma(t) \le x\right) \\
&= \frac{1}{\mu}\int_0^\infty \mathbf{P}\left(\gamma(s) \le x, T_1 > s\right)\mathrm{d}s = \frac{1}{\mu}\int_0^\infty \mathbf{P}\left(T_1 - s \le x, T_1 > s\right)\mathrm{d}s \\
&= \frac{1}{\mu}\int_0^\infty \mathbf{P}\left(s \le T_1 < s + x\right)\mathrm{d}s = \frac{1}{\mu}\int_0^\infty (F(s + x) - F(s))\mathrm{d}s \\
&= -\frac{1}{\mu}\left(\int_x^\infty (1 - F(s))\mathrm{d}s - \int_0^\infty (1 - F(s))\mathrm{d}s\right) = \frac{1}{\mu}\int_0^x (1 - F(s))\mathrm{d}s.
\end{aligned}$$

The statement for $\{\gamma(t),\ t \ge 0\}$ can be obtained analogously. □

Similarly to the renewal processes, the law of large numbers and the central limit theorem can be proved for the regenerative processes, too. Here we will not deal with these questions.

4.2.1 Estimation of Convergence Rate for Regenerative Processes

For a wide class of regenerative processes (e.g., stochastic processes describing queueing systems) one can estimate the rate of convergence of distributions of certain parameters to a stationary distribution by means of the so-called coupling method [65].

Lemma 4.43 (*Coupling lemma*). *For the arbitrary random variables X and Y and an arbitrary Borel set A of the real line the following statements hold:*

(i) $|\mathbf{P}\left(X \in A\right) - \mathbf{P}\left(Y \in A\right)| \le \mathbf{P}\left(X \ne Y\right)$.

(ii) If $X = X_1 + \ldots + X_n$ *and* $Y = Y_1 + \ldots + Y_n$, *then* $|\mathbf{P}\left(X \in A\right) - \mathbf{P}\left(Y \in A\right)| \le \sum_{k=1}^{n} \mathbf{P}\left(X_k \ne Y_k\right)$.

Proof. If $\mathbf{P}\left(X \in A\right) = \mathbf{P}\left(Y \in A\right)$, then (i) is obviously true.

Suppose that $\mathbf{P}(X \in A) > \mathbf{P}(Y \in A)$ (if one changes the notation, then this can always be done if the two probabilities differ). Then

$$\begin{aligned}|\mathbf{P}(X \in A) - \mathbf{P}(Y \in A)| &= \mathbf{P}(X \in A) - \mathbf{P}(Y \in A) \\ &\le \mathbf{P}(X \in A) - \mathbf{P}(Y \in A, X \in A) \\ &= \mathbf{P}(X \in A, Y \in A^c) \le \mathbf{P}(X \ne Y).\end{aligned}$$

Proof of relation (ii). Since $\{X \ne Y\} \subset \bigcup_{k=1}^{n} \{X_k \ne Y_k\}$, we have

$$\mathbf{P}(X \ne Y) \le \mathbf{P}\left(\bigcup_{k=1}^{n} \{X_k \ne Y_k\}\right) \le \sum_{k=1}^{n} \mathbf{P}(X_k \ne Y_k).$$

□

Application of Coupling Lemma Let $Z = \{Z(j),\ j \ge 1\}$ be the discrete-time, real-valued regenerative process under consideration. Assume that there exists the weak stationary limit of the process $\tilde{Z} = \{Z(j+n),\ j \ge 1\}$ as $n \to \infty$ (its finite-dimensional distributions weakly converge to the finite-dimensional distributions of a stationary process), which is also regenerative, and let $Y = \{Y(j),\ j \ge 1\}$ be its realization, not necessarily different from Z on the same probability space. Let τ denote the first instant when the processes Z and Y are regenerated at the same time (in many concrete cases the distribution of τ can be easily estimated). Then the convergence rate of the distribution of $Z(j)$ can be estimated by means of the distribution of τ as follows: if after the regeneration point τ the process Z is replaced by the next part of process Y following the common regeneration point τ, then the finite-dimensional distributions of process Z do not change. It is clear that $\{\tau < j\} \subseteq \{Z(j) = Y(j)\}$, i.e., $\{Z(j) \ne Y(j)\} \subseteq \{\tau \ge j\}$, from which, using the coupling lemma for the arbitrary Borel set A of the real line, the estimation

$$|\mathbf{P}(Z(j) \in A) - \mathbf{P}(Y(j) \in A)| \le \mathbf{P}(Z(j) \ne Y(j)) \le \mathbf{P}(\tau \ge j)$$

holds.

4.3 Analysis Methods Based on Markov Property

Definition 4.44. A discrete-state, continuous-time stochastic process, $X(t)$, possesses the **Markov propety** at time t_n if for all $n, m \ge 1$, $0 \le t_0 < t_1 < \ldots < t_n < t_{n+1} < \ldots < t_{n+m}$, and $x_0, x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m} \in S$ we have

$$\begin{aligned}&\mathbf{P}(X(t_{n+m}) = x_{n+m}, \ldots, X(t_{n+1}) = x_{n+1} | X(t_n) = x_n, \ldots, X(t_0) = x_0) \\ &\quad = \mathbf{P}(X(t_{n+m}) = x_{n+m}, \ldots, X(t_{n+1}) = x_{n+1} | X(t_n) = x_n). \qquad (4.5)\end{aligned}$$

In this case t_n is referred to as a regenerative point.

A commonly applied interpretation of the Markov property is as follows. Assuming that the current time is t_n (present), which is a regenerative point, and we know the current state of the process $X(t_n)$, then the future of the stochastic process $X(t)$ for $t_n \leq t$ is independent of the past history of the process $X(t)$ for $0 \leq t < t_n$, and it only depends on the current state of the process $X(t_n)$. That is, if one knows the present state, the future is independent of the past.

In the case of discrete-time processes, it is enough to check if the one-step state transitions are independent of the past, i.e., it is enough to check the condition for $m = 1$.

Usually, we restrict our attention to stochastic processes with nonnegative parameters (positive half of the time axes), and in these cases we assume that $t = 0$ is a regenerative point.

4.3.1 Time-Homogeneous Behavior

Definition 4.45. The stochastic process $X(t)$ is *time homogeneous* if the stochastic behavior of $X(t)$ is invariant for time shifting, that is, the stochastic behavior of $X(t)$ and $X'(t) = X(t + s)$ are identical in distribution $X(t) \stackrel{d}{=} X'(t)$.

Corollary 4.46. *If the time-homogeneous stochastic process $X(t)$ possesses the Markov property at time T and $X(T) = i$, then $X(t) \stackrel{d}{=} X(t - T)$ if $X(0) = i$.*

The corollary states that starting from two different Markov points with the same state results in stochastically identical processes.

4.4 Analysis of Continuous-Time Markov Chains

Definition 4.47. The discrete-state, continuous-time stochastic process $X(t)$ is a *continuous-time Markov chain* (CTMC) if it possesses the Markov property for all $t \geq 0$.

Based on this definition and assuming time-homogeneous behavior we obtain the following properties.

Corollary 4.48. *An arbitrary finite-dimensional joint distribution of a CTMC is composed of the product of transition probabilities multiplied by an initial probability.*

Corollary 4.49. *For the time points $t < u < v$ the following Chapman–Kolmogorov equation holds:*

$$\hat{p}_{ij}(t, v) = \sum_{l \in S} \hat{p}_{il}(t, u)\hat{p}_{lj}(u, v); \quad \hat{\Pi}(t, v) = \hat{\Pi}(t, u)\hat{\Pi}(u, v), \tag{4.6}$$

where $\hat{p}_{ij}(t,u) = \mathbf{P}(X(u) = j \mid X(t) = i)$ *for all* $i,\ j \in S, 0 \le t \le u$, $\hat{\Pi}(t,u) = \left[\hat{p}_{ij}(t,u)\right]$. *In the case of time-homogeneous processes the time shifts* $u - t = \tau_1$ *and* $v - u = \tau_2$ *play a role:*

$$p_{ij}(\tau_1 + \tau_2) = \sum_{l \in S} p_{il}(\tau_1) p_{lj}(\tau_2); \quad \Pi(\tau_1 + \tau_2) = \Pi(\tau_1)\Pi(\tau_2), \tag{4.7}$$

where $\Pi(\tau) = \left[p_{ij}(\tau)\right]$, $p_{ij}(\tau) = \mathbf{P}(X(\tau) = j \mid X(0) = i)$, *for all* $i,\ j \in S, 0 \le \tau$.

Definition 4.50. The stochastic evolution of a CTMC is commonly characterized by an *infinitesimal generator* matrix (commonly denoted by Q) that can be obtained from the derivative of the state-transition probabilities as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Pi(t) = \lim_{\delta \to 0} \frac{\Pi(t+\delta) - \Pi(t)}{\delta} = \Pi(t) \underbrace{\lim_{\delta \to 0} \frac{\Pi(\delta) - I}{\delta}}_{Q} = \Pi(t) Q. \tag{4.8}$$

Corollary 4.51. *The sojourn time of a CTMC in a given state* i *is exponentially distributed with the parameter* $q_i = -q_{ii}$. *The probability that after state* i *the next visited state will be state* j *is* q_{ij}/q_i, *and it is independent of the sojourn time in state* i.

Remark 4.52. Based on Corollary 4.51 and the properties of the exponential distribution, the state transitions of a CTMC can also be interpreted in the following way. When the CTMC moves to state i, several exponentially distributed activities start, exactly one for each nonzero transition rate. The time of the activity associated with the state transition from state i to state j is exponentially distributed with the parameter q_{ij}. The CTMC leaves state i and moves to the next state when the first one of these activities completes. The next visited state is the state whose associated activity finishes first.

Corollary 4.53 (*Short-term behavior of CTMCs). During a short time period* Δ, *the behavior of a CTMC is characterized by the following transition probabilities:*

(a) $\mathbf{P}(X(t+\Delta) = i | X(t) = i) = 1 - q_i\Delta + o(\Delta)$;
(b) $\mathbf{P}(X(t+\Delta) = j | X(t) = i) = q_{ij}\Delta + o(\Delta)$ *for* $i \ne j$;
(c) $\mathbf{P}(X(t+\Delta) = j, X(u) = k | X(t) = i) = o(\Delta)$ *for* $i \ne k,\ j \ne k$, *and* $t < u < t + \Delta$,

where $o(x)$ *denotes the set of functions with the property* $\lim_{x \to 0} o(x)/x = 0$.

According to the corollary, two main events can happen with significant probability during a short time period:

- The CTMC stays in the initial state during the whole period [(a)].
- It moves from state i to j [(b)].

The event that more than one state transition happens during a short time period [(c)] has a negligible probability as $\Delta \to 0$.

Corollaries 4.51 and 4.53 allow different analytical approaches for the description of the transient behavior of CTMCs.

4.4.1 Analysis Based on Short-Term Behavior

Let $X(t)$ be a CTMC with state space S, and let us consider the change in state probability $P_i(t+\Delta) = \mathbf{P}(X(t+\Delta) = i)$ $(i \in S)$ considering the possible events during the interval $(t, t+\Delta)$. The following cases must be considered:

- There is no state transition during the interval $(t, t+\Delta)$. In this case $P_i(t+\Delta) = P_i(t)$, and the probability of this event is $1 - q_i\Delta + o(\Delta)$.
- There is one state transition during the $(t, t+\Delta)$ interval from state k to state i. In this case $P_i(t+\Delta) = P_k(t)$, and the probability of this event is $q_{ki}\Delta + o(\Delta)$.
- The process stays in state i at time $t+\Delta$ such that there is more than one state transition during the interval $(t, t+\Delta)$. The probability of this event is $o(\Delta)$.

Considering these cases we can compute $P_i(t+\Delta)$ from $P_k(t)$, $k \in S$, as follows:

$$\begin{aligned} P_i(t+\Delta) &= (1 - q_i\Delta + o(\Delta))P_i(t) + \sum_{k\in S, k\neq i} (q_{ki}\Delta + o(\Delta))P_k(t) + o(\Delta) \\ &= (1 - q_i\Delta)P_i(t) + \sum_{k\in S, k\neq i} (q_{ki}\Delta)P_k(t) + o(\Delta), \end{aligned}$$

from which

$$\frac{P_i(t+\Delta) - P_i(t)}{\Delta} = -q_i P_i(t) + \sum_{k\in S, k\neq i} q_{ki} P_k(t) + \frac{o(\Delta)}{\Delta} = \sum_{k\in S} q_{ki} P_k(t) + \frac{o(\Delta)}{\Delta}.$$

Finally, setting the limit $\Delta \to 0$ we obtain that

$$\frac{\mathrm{d}P_i(t)}{\mathrm{d}t} = \sum_{k\in S} q_{ki} P_k(t).$$

Introducing the row vector of state probabilities $P(t) = \{P_i(t)\}, i \in S$, we obtain the vector-matrix form of the previous equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} P(t) = P(t)\mathbf{Q}. \tag{4.9}$$

A differential equation describes the evolution of a transient state probability vector. To define the state probabilities, we additionally need to have an initial

condition. In practical applications, the initial condition is most often the state probability distribution at time 0, i.e., $P(0)$. The solution of Eq. (4.9) with initial condition $P(0)$ is [55]

$$P(t) = P(0)\mathrm{e}^{\mathbf{Q}t} = P(0)\sum_{n=0}^{\infty}\frac{\mathbf{Q}^n\, t^n}{n!}.$$

Transform Domain Description The Laplace transform of the two sides of Eq. (4.9) gives

$$s\ P^*(s) - P(0) = P^*(s)\mathbf{Q},$$

from which we can express $P^*(s)$ in the following form:

$$P^*(s) = P(0)[s\mathbf{I} - \mathbf{Q}]^{-1}.$$

Comparing the time and transform domain expressions we have that $\mathrm{e}^{\mathbf{Q}t}$ and $[s\mathbf{I} - \mathbf{Q}]^{-1}$ are Laplace transform pairs of each other.

Stationary Behavior If $\lim_{t\to\infty} P_i(t)$ exists, then we say that $\lim_{t\to\infty} P_i(t) = P_i$ is the stationary probability of state i. In this case, $\lim_{t\to\infty} \mathrm{d}P_i(t)/\mathrm{d}t = 0$, and the stationary probability satisfies the system of linear equations $\sum_{k\in S} q_{ki} P_k(t) = 0$ for all $k \in S$.

4.4.2 Analysis Based on First State Transition

Let $X(t)$ be a CTMC with state space S, and let $T_1, T_2, T_3, \ldots$ denote the time of the first, second, etc. state transitions of the CTMC. We assume that $T_0 = 0$, and $\tau_1, \tau_2, \tau_3, \ldots$ are the sojourn times spent in the consecutively visited states ($\tau_i = T_i - T_{i-1}$). We compute the state-transition probability $\pi_{ij}(t) = \mathbf{P}(X(t) = j \mid X(0) = i)$ assuming that $T_1 = h$, i.e., we are interested in

$$\pi_{ij}(t|T_1 = h) = \mathbf{P}(X(t) = j \mid X(0) = i, T_1 = h).$$

We have

$$\pi_{ij}(t|T_1 = h) = \begin{cases} \delta_{ij}, & h \geq t, \\ \displaystyle\sum_{k\in S, k\neq i} \frac{q_{ik}}{-q_{ii}}\pi_{kj}(t-h), & h < t, \end{cases} \tag{4.10}$$

where δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$), and $\frac{q_{ik}}{-q_{ii}}$ is the probability that after visiting state i the Markov chain moves to state k. In the case of general stochastic processes, this probability might depend on the sojourn time in state i, but in the case of CTMCs, it is independent.

Equation (4.10) has two cases:

- If the time point of interest, t, is before the first state transition of the CTMC, $h \geq t$, then the conditional state-transition probability is either 1 (if the initial and final states are identical $i = j$) or 0 (if $i \neq j$).
- If the time point of interest, t, is after the first state transition of the CTMC, $T_1 < t$, then we can analyze the evolution of the process from T_1 to t using the fact that the process possesses the Markov property at time T_1. In this case we need to consider all possible states that might be visited at time T_1, $k \in S, k \neq i$, with the associated probability $\frac{q_{ik}}{-q_{ii}}$. The state-transition probabilities from T_1 to t are identical with the state-transition probabilities of the original process from 0 to $T_1 - t$, assuming that the original process starts from state k.

The distribution of T_1 is known. It is exponentially distributed with the parameter $-q_{ii}$. Its cumulated and probability density functions are $F_{T_1}(x) = 1 - \mathrm{e}^{q_{ii}x}$ and $f_{T_1}(x) = -q_{ii}\mathrm{e}^{q_{ii}x}$, respectively. With that we can apply the total probability theorem to compute the (unconditional) state-transition probability $\pi_{ij}(t)$:

$$\begin{aligned}
\pi_{ij}(t) &= \int_{h=0}^{\infty} \pi_{ij}(t|T_1 = h)\ f_{T_1}(h)\ \mathrm{d}h \\
&= \int_{h=t}^{\infty} \delta_{ij}\ f_{T_1}(h)\ \mathrm{d}h + \int_{h=0}^{t} \sum_{k \in S, k \neq i} \frac{q_{ik}}{-q_{ii}}\ \pi_{kj}(t-h)\ f_{T_1}(h)\ \mathrm{d}h \\
&= \delta_{ij}\ (1 - F_{T_1}(t)) + \int_{h=0}^{t} \sum_{k \in S, k \neq i} \frac{q_{ik}}{-q_{ii}}\ \pi_{kj}(t-h)\ f_{T_1}(h)\ \mathrm{d}h \\
&= \delta_{ij}\ \mathrm{e}^{q_{ii}t} + \sum_{k \in S, k \neq i} q_{ik} \int_{h=0}^{t} \pi_{kj}(t-h)\ \mathrm{e}^{q_{ii}h}\ \mathrm{d}h. \qquad (4.11)
\end{aligned}$$

The obtained integral equation is commonly referred to as a Volterra integral equation. Its only unknown is the state-transition probability function $\pi_{ij}(t)$. The numerical methods developed for the numerical analysis of Volterra integral equations can be used to compute the state-transition probabilities of a CTMC.

Relation of Analysis Methods We can rewrite Eq. (4.11) in the following form:

$$\begin{aligned}
\pi_{ij}(t) &= \delta_{ij}\ \mathrm{e}^{q_{ii}t} + \sum_{k \in S, k \neq i} q_{ik} \int_{h=0}^{t} \pi_{kj}(t-h)\ \mathrm{e}^{q_{ii}h}\ \mathrm{d}h \\
&= \delta_{ij}\ \mathrm{e}^{q_{ii}t} + \sum_{k \in S, k \neq i} q_{ik} \int_{h=0}^{t} \pi_{kj}(h)\ \mathrm{e}^{q_{ii}(t-h)}\ \mathrm{d}h \\
&= \delta_{ij}\ \mathrm{e}^{q_{ii}t} + \sum_{k \in S, k \neq i} q_{ik}\ \mathrm{e}^{q_{ii}t} \int_{h=0}^{t} \pi_{kj}(h)\ \mathrm{e}^{-q_{ii}h}\ \mathrm{d}h. \qquad (4.12)
\end{aligned}$$

The derivation of the two sides of Eq. (4.12) according to t is as follows:

$$\begin{aligned}\pi'_{ij}(t) &= \delta_{ij}\, q_{ii}\, \mathrm{e}^{q_{ii}t} + \sum_{k\in S, k\neq i} q_{ik}\left(q_{ii}\, \mathrm{e}^{q_{ii}t}\int_{h=0}^{t} \pi_{kj}(h)\, \mathrm{e}^{-q_{ii}h}\, \mathrm{d}h + \mathrm{e}^{q_{ii}t}\, \pi_{kj}(t)\, \mathrm{e}^{-q_{ii}t}\right)\\ &= \sum_{k\in S, k\neq i} q_{ik}\, \pi_{kj}(t) + q_{ii}\underbrace{\left(\delta_{ij}\, \mathrm{e}^{q_{ii}t} + \sum_{k\in S, k\neq i} q_{ik}\, \mathrm{e}^{q_{ii}t}\int_{h=0}^{t} \pi_{kj}(h)\, \mathrm{e}^{-q_{ii}h}\, \mathrm{d}h\right)}_{\pi_{ij}(t)}\\ &= \sum_{k\in S} q_{ik}\pi_{kj}(t),\end{aligned}$$

where we used Eq. (4.11) for the substitution of the integral expression. The obtained differential equation is similar to that provided by the analysis of the short-term behavior.

Transform Domain Description To relate the two transient descriptions of the CTMC, one with a differential equation and one with an integral equation, we transform these descriptions into a Laplace transform domain. It is easy to take the Laplace transform from the last line of Eq. (4.11) because the second term of the right-hand side is a convolution integral. That is,

$$\pi^*_{ij}(s) = \delta_{ij}\, \frac{1}{s - q_{ii}} + \sum_{k\in S, k\neq i} q_{ik}\, \pi^*_{kj}(s)\, \frac{1}{s - q_{ii}}.$$

Multiplying by the denominator and using that $-q_{ii} = \sum_{k\in S, k\neq i} q_{ik}$ we obtain

$$s\, \pi^*_{ij}(s) = \delta_{ij} + \sum_{k\in S} q_{ik}\, \pi^*_{kj}(s),$$

which can be written in the matrix form

$$s\, \Pi^*(s) = \mathbf{I} + \mathbf{Q}\Pi^*(s).$$

Finally, we have

$$\Pi^*(s) = [s\mathbf{I} - \mathbf{Q}]^{-1},$$

which is identical to the Laplace transform expression obtained from the differential equation.

Embedded Markov Chain at State Transitions Let $X_i \in S, i = 0, 1, \ldots,$ denote the ith visited state of the Markov chain $X(t)$, which is the state of the Markov chain in the interval (T_i, T_{i+1}) (Fig. 4.2). The $X_0, X_1, \ldots$ series of random variables is a discrete-time Markov chain (DTMC) due to the Markov property of $X(t)$. This DTMC is commonly referred to as a Markov chain embedded at the state transitions or simply an *embedded Markov chain* (EMC). The state-transition probability matrix of the EMC is

$$\Pi_{ij} = \begin{cases} \dfrac{q_{ij}}{-q_{ii}}, & i \neq j, \\ 0, & i = j. \end{cases}$$

Stationary Analysis Based on the EMC The stationary distribution of the EMC $\hat{P}$ (which is the solution of $\hat{P} = \hat{P}\Pi, \sum_i \hat{P}_i = 1$) defines the relative frequency of the visits to the state of the Markov chain. The higher the stationary probability is, the more frequently the state is visited. The stationary behavior of the CTMC $X(t)$ is characterized by two main factors: how often the state is visited (represented by $\hat{P}_i$) and how long a visit lasts. If state i is visited twice as frequently as state j but the mean time of a visit to state i is half the mean time of a visit to j, then the stationary probabilities of states i and j are identical. This intuitive behavior is summarized in the following general rule of renewal theory [58]:

$$P_i = \frac{\hat{P}_i \hat{\tau}_i}{\sum_j \hat{P}_j \hat{\tau}_j},$$

where $\hat{\tau}_j$ is the mean time spent in state j, which is known from the diagonal element of the infinitesimal generator, $\hat{\tau}_j = -1/q_{jj}$.

Discrete-Event Simulation of CTMCs There are at least two possible approaches.

- When the CTMC is in state i, first draw an exponentially distributed random sample with parameter $-q_{ii}$ for the sojourn time in state i, then draw a discrete random sample for deciding the next visited state with distribution Π_{ij}, $j \in S$.
- When the CTMC is in state i, draw an exponentially distributed random sample with parameter q_{ij}, say τ_{ij}, for all positive transition rates of row i of the infinitesimal generator matrix. Find the minimum of these samples, $\min_j \tau_{ij}$. The sojourn time in state i is this minimum, and the next state is the one whose associated random sample is minimal.

4.5 Semi-Markov Process

Definition 4.54. The discrete-state, continuous-time random process $X(t)$ is a semi-Markov process if it is time homogeneous and it possesses the Markov property at the state-transition instances (Fig. 4.2).

The name semi-Markov process comes from the fact that such processes do not always possess the Markov property (during its sojourn in a state), but there are particular instances (state-transition instances) when they do.

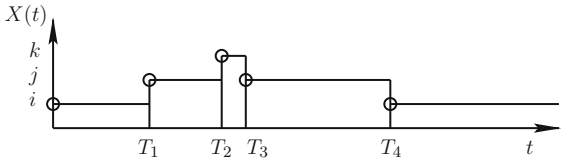
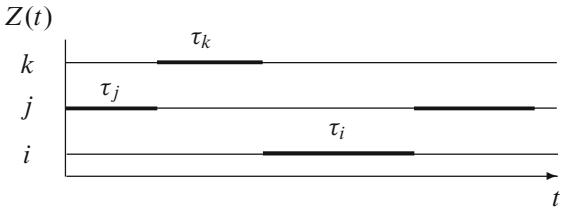


Fig. 4.2 Semi-Markov process that possesses the Markov property at the indicated time points

Corollary 4.55. *The sojourn time in state i can be any general real-valued positive random variable. During a sojourn in state i, both the remaining time in both that state and the next visited state depend on the elapsed time since the process entered state i.*



Example 4.56. A two-state (up/down) system fails at a rate λ (the up time of the system is exponentially distributed with parameter λ) and gets repaired at a rate μ. To avoid long down periods, the repair process is stopped and a replacement process is initialized after a deterministic time limit d. The time of the replacement is a random variable with a distribution $G(t)$. Define a system model and check if it is a semi-Markov process.

Because a CTMC always possesses the Markov property, it follows that the sojourn time in a state is exponentially distributed and that the distribution of the next state is independent of the sojourn time. For example, considering the first state transition and the sojourn time in the first state we have

$$\mathbf{P}\,(X_1 = j, T_1 = c|X_0 = i) = \mathbf{P}\,(X_1 = j\,|X_0 = i)\mathbf{P}\,(T_1 = c|X_0 = i).$$

This property does not hold for semi-Markov processes in general. The most important consequences of the definition of semi-Markov processes are the following ones. The sojourn time in a state can have any positive distribution, and the distribution of the next state and the time spent in a state are not independent in general. Consequently, to define a semi-Markov process, this joint distribution must be given. This is usually done by defining the kernel matrix of a process whose i, j element is

$$Q_{ij}\,(t) = \mathbf{P}\,(X(T_{i+1}) = j, \tau_{i+1} \leq t \mid X(T_i) = i).$$

Utilizing the time homogeneity of the process we further have for T_i that

$$Q_{ij}\,(t) = \mathbf{P}\,(X(T_{i+1})=j, \tau_{i+1} \leq t \mid X(T_i)=i)=\mathbf{P}\,(X(T_1)=j, T_1 \leq t \mid X(0) = i).$$

The analysis of semi-Markov processes is based on the results of renewal theory and the analysis of an EMC (of state-transition instances). The definition of a semi-Markov process requires knowledge of the kernel matrix $Q(t) = \{Q_{ij}(t)\}$ (for $t \geq 0$) and an initial distribution. It is commonly assumed that $X(t)$ possesses the Markov property at time $t = 0$.

4.5.1 Analysis Based on State Transitions

Let $X(t) \in S$ be a continuous-time semi-Markov process, $T_1, T_2, T_3, \ldots$ the state-transition instances, and $\tau_1, \tau_2, \tau_3, \ldots$ the consecutive sojourn times ($\tau_i = T_i - T_{i-1}$). We assume $T_0 = 0$. We intend to compute the state-transition probability $\pi_{ij}(t) = \mathbf{P}(X(t) = j \mid X(0) = i)$ assuming that the sojourn in the first state finishes at time h ($T_1 = h$), that is,

$$\pi_{ij}(t|T_1 = h) = \mathbf{P}(X(t) = j \mid X(0) = i, T_1 = h).$$

In this case

$$\pi_{ij}(t|T_1 = h) = \begin{cases} \delta_{ij}, & h \geq t, \\ \sum_{k \in S} \mathbf{P}(X(T_1) = k \mid X(0) = i, T_1 = h)\, \pi_{kj}(t-h), & h < t, \end{cases} \tag{4.13}$$

where $\mathbf{P}(X(T_1) = j \mid X(0) = i, T_1 = h)$ is the probability that the process will start from state i at time 0 and is in state j right after the state transition at time T_1 assuming $T_1 = h$. In contrast with CTMCs, this probability depends on the sojourn time in state i:

$$\begin{aligned} &\mathbf{P}(X(T_1) = j \mid X(0) = i, T_1 = h) \\ &\quad = \lim_{\Delta \to 0} \frac{\mathbf{P}(X(T_1) = j, h < T_1 \leq h + \Delta \mid X(0) = i)}{\mathbf{P}(h < T_1 \leq h + \Delta \mid X(0) = i)} \\ &\quad = \lim_{\Delta \to 0} \frac{Q_{ij}(h + \Delta) - Q_{ij}(h)}{Q_i(h + \Delta) - Q_i(h)} = \frac{dQ_{ij}(h)}{dQ_i(h)} \ , \end{aligned} \tag{4.14}$$

where $Q_i(h)$ denotes the distribution of time spent in state i,

$$Q_i(t) = \mathbf{P}(T_1 \leq t \mid Z(0){=}i) = \sum_j \mathbf{P}(Z(T_1){=}j, T_1 \leq t \mid Z(0){=}i) = \sum_j Q_{ij}(t).$$

It is commonly assumed that state transitions are real, which means that after staying in state i a state transition moves the process to a different state. This means that $Q_{ii}(t) = 0, \ \forall i \in S$. It is also possible to consider virtual state transitions from state i to state i, but this does not expand the set of semi-Markov

processes and we do not consider it here. Note that the meaning of a diagonal element of a semi-Markov kernel matrix is completely different from that of a diagonal element of an infinitesimal generator of a CTMC. One of the technical consequences of this difference is the fact that we do not need to exclude the diagonal element from the summations over the set of states.

Two cases are considered in Eq. (4.13):

- If the time point of interest, t, is before the first state transition of the process ($h \geq t$), then the conditional state-transition probability is either 0 or 1 depending on the initial and final states. If the initial state i is identical with the final state j, then the transition probability is 1 because there is no state transition up to time t, otherwise it is 0.
- If the time point of interest, t, is after the first state transition of the process ($h < t$), then we need to evaluate the distribution of the next state k, assuming that the state transition occurs at time h, and after that the state-transition probability from the new state k to the final state j during time $t-h$, using the Markov property of the process at time h. The probability that the process moves to state k assuming it occurs at time h is $\frac{dQ_{ij}(h)}{dQ_i(h)}$, and the probability of its moving from state k to state j during an interval of length $t-h$ is $\pi_{ij}(t-h)$.

The distribution of the condition of Eq. (4.13) is known. The distribution of the sojourn time in state i is $Q_i(h)$. Using the law of total probability we obtain

$$\begin{aligned}\pi_{ij}(t) &= \int_{h=0}^{\infty} \pi_{ij}(t|T_1 = h)\, \mathrm{d}F_{T_1}(h) \\ &= \int_{h=t}^{\infty} \delta_{ij}\, \mathrm{d}Q_i(t) + \int_{h=0}^{t} \sum_{k\in S} \frac{dQ_{ik}(h)}{dQ_i(h)}\, \pi_{kj}(t-h)\, \mathrm{d}Q_i(h) \\ &= \delta_{ij}\,(1 - Q_i(t)) + \int_{h=0}^{t} \sum_{k\in S} \pi_{kj}(t-h)\, \mathrm{d}Q_{ik}(h). \end{aligned} \tag{4.15}$$

Similar to the case of CTMCs, analysis based on the first state transition resulted in a Volterra integral equation also in the case of semi-Markov processes. The transient behavior of semi-Markov processes can be computed using the same numerical procedures.

Transform Domain Description We take the Laplace transform of both sides of the Volterra integral Eq. (4.15). The only nontrivial term is a convolution integral on the right-hand side:

$$\pi_{ij}^*(s) = \delta_{ij}\,(1 - Q_i^*(s)) + \sum_{k\in S} q_{ik}^*(s)\, \pi_{kj}^*(s),$$

where $q_{ik}(t) = \mathrm{d}Q_{ik}(t)/\mathrm{d}t$ and the transform domain functions are defined as $f^*(s) = \int_0^\infty f(t)\mathrm{e}^{-st}\mathrm{d}t$.

Introducing the diagonal matrix $D^*(s)$ composed of the elements $1 - Q_i^*(s)$, that is , $D^*(s) = \text{diag}\langle 1 - Q_i^*(s)\rangle$, the Laplace transforms of the state transition probabilities are obtained in matrix form,

$$\Pi^*(s) = D^*(s) + q^*(s)\Pi^*(s),$$

from which

$$\Pi^*(s) = [\mathbf{I} - q^*(s)]^{-1} D^*(s).$$

Stationary Behavior The stationary analysis of a semi-Markov process is very similar to the stationary analysis of a CTMC based on an EMC. Let the transition probability matrix of the EMC be Π. It is obtained from the kernel matrix through the following relation:

$$\Pi_{ij} = \mathbf{P}(Z(T_1) = j \mid Z(0) = i) = \lim_{t\to\infty} \mathbf{P}(Z(T_1) = j, T_1 \le t \mid Z(0) = i) = \lim_{t\to\infty} Q_{ij}(t).$$

The stationary distribution of the EMC $\hat{P}$ is the solution of the linear system $\hat{P} = \hat{P}\,\Pi, \sum_i \hat{P}_i = 1$. The stationary distribution of the semi-Markov process is

$$P_i = \frac{\hat{P}_i \hat{\tau}_i}{\sum_j \hat{P}_j \hat{\tau}_j}, \tag{4.16}$$

where $\hat{\tau}_i$ is the mean time spent in state i. It can be computed from a kernel matrix using $\hat{\tau}_i = \int_0^\infty (1 - Q_i(t))\mathrm{d}t$.

Discrete-Event Simulation of Semi-Markov Processes The initial distribution and the $Q_{ij}(t)$ kernel completely define the stochastic behavior of a semi-Markov process. As a consequence, it is possible to simulate the process behavior based on them.

The key step of the simulation is to draw dependent samples for the sojourn time and the next visited state. This can be done based on the marginal distribution of one of the two random variables and a conditional distribution of the other one. Depending on which random variable is sampled first, there are two ways to simulate a semi-Markov process:

- When the process is in state i, first draw a $Q_i(t)$ distributed sample for the sojourn time, denoted by τ, then draw a sample for the next state assuming that the sojourn is τ based on the discrete probability distribution $\mathbf{P}(X(T_1) = j \mid X(0) = i, T_1 = \tau)$ $(\forall j \in S)$ given in Eq. (4.14).
- When the process is in state i, first draw a sample for the next visited state based on the discrete probability distribution $\Pi_{ij} = \mathbf{P}(X(T_1) = j \mid X(0) = i)$ $(\forall j \in S)$, then draw a sample for the sojourn time given in the next state with a distribution

$$\mathbf{P}(T_1 \le t \mid Z(0) = i, Z(T_1) = j) = \frac{Q_{ij}(t)}{\Pi_{ij}}. \tag{4.17}$$

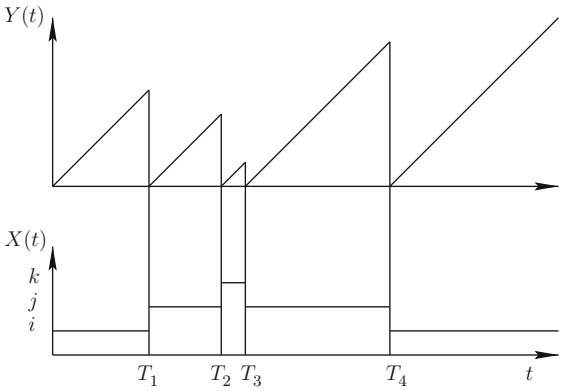


Fig. 4.3 Analysis of semi-Markov process with supplementary variable

4.5.2 *Transient Analysis Using the Method of Supplementary Variables*

A semi-Markov process does not possess the Markov property during its sojourn in a state. For example, the distribution of the time till the next state transition may depend on the amount of time that has passed since the last state transition. It is possible to extend the analysis of semi-Markov processes so that all information that makes the future evolution of the process conditionally independent of its past history is involved in the process description for $\forall t \geq 0$. It is indeed the Markov property for $\forall t \geq 0$. In the case of semi-Markov processes, this means that the discrete state of the process $X(t)$ and the time passed since the last state transition $Y(t) = t - \max(T_i \leq t)$ need to be considered together because the vector-valued stochastic process $\{X(t), Y(t)\}$ is already such that the future behavior of this vector process is conditionally independent of its past given the current value of the vector. That is, the $\{X(t), Y(t)\}$ process possesses the Markov property for $\forall t \geq 0$. The behavior of the $\{X(t), Y(t)\}$ process is depicted in Fig. 4.3.

This extension of a random process with an additional variable such that the obtained vector-valued process possesses the Markov property is referred to as the method of supplementary variables [24].

With $X(t)$ and $Y(t)$ and the kernel matrix of the process we can compute the distribution of time till the next state transition at any time instant; this is commonly referred to as the remaining sojourn time in the given state. If at time t the process stays in state i for a period of τ $[X(t) = i, Y(t) = \tau]$ and the distribution of the total sojourn time in state i is $Q_i(t)$, then the distribution of the remaining sojourn time in state i, denoted by γ, is

$$\mathbf{P}(\gamma \leq t) = \mathbf{P}(\gamma_t \leq t + \tau \mid \gamma_t > \tau) = \frac{Q_i(t+\tau) - Q_i(\tau)}{1 - Q_i(\tau)},$$

where γ_t denotes the total time spent in state i during this visit in state i.

To analyze the $\{X(t), Y(t)\}$ process, we need to characterize the joint distribution of the following two quantities:

$$h_i(t,x) = \frac{\mathbf{P}\left(X(t) = i, x \le Y(t) < x + \Delta\right)}{\Delta}.$$

It is possible to obtain $h_i(t,x)$ based on the analysis of the short-term behavior of CTMCs:

$$\begin{aligned}
&h_i(t+\Delta, x)\\
&\quad = \mathbf{P}[\text{there is no state transition in the interval } (t, t+\Delta)]\\
&\qquad \times h_i(t+\Delta, x \mid \text{there is no state transition})\\
&\qquad + \mathbf{P}[\text{there is one state transition in the interval } (t, t+\Delta)]\\
&\qquad \times h_i(t+\Delta, x \mid \text{there is one state transition}) + o(\Delta),
\end{aligned}$$

where $h_i(t+\Delta, x \mid \text{condition})$ denotes $\dfrac{\mathbf{P}\left(X(t) = i, x \le Y(t) < x + \Delta \mid \text{condition}\right)}{\Delta}$. The probability of the state transition can be computed based on the distribution of the remaining sojourn time:

$$\begin{aligned}
&\mathbf{P}[\text{there is one state transition in the interval } (t, t+\Delta)]\\
&\quad = \mathbf{P}\left(\text{remaining sojourn time} \le \Delta\right) = \frac{Q_i(x+\Delta) - Q_i(x)}{1 - Q_i(x)},
\end{aligned}$$

from which

$$\mathbf{P}[\text{there is no state transition in the interval } (t, t+\Delta)] = \frac{1 - Q_i(x+\Delta)}{1 - Q_i(x)}.$$

Immediately following a state transition $Y(t)$ is reset to zero. Consequently, the probability that $Y(t+\Delta) = x$ for a fixed $x > 0$ is zero when Δ is sufficiently small. That is,

$$h_i[t+\Delta, x \mid \text{there is one state transition in the interval } (t, t+\Delta)] = 0 \ \text{ if } \ x > 0.$$

It follows that

$$\begin{aligned}
&h_i(t+\Delta, x)\\
&\quad = \mathbf{P}[\text{there is no state transition in the interval } (t, t+\Delta)]\\
&\qquad \times h_i[t+\Delta, x \mid \text{there is no state transition in the interval } (t, t+\Delta)]\\
&\quad = \frac{1 - Q_i(x+\Delta)}{1 - Q_i(x)} \cdot h_i(t, x-\Delta).
\end{aligned}$$

Analysis of the process $\{X(t), Y(t)\}$ is made much simpler by the use of the transition rate of α_i instead of its distribution $Q_i(t)$. The transition rate is defined by

$$\lambda_i(t) = \lim_{\Delta \to 0} \frac{\mathbf{P}(\alpha_i \leq t + \Delta \mid \alpha_i > t)}{\Delta} = \lim_{\Delta \to 0} \frac{Q_i(t+\Delta) - Q_i(t)}{\Delta\,(1 - Q_i(t))} = \frac{Q_i'(t)}{1 - Q_i(t)}\,.$$

It is also referred to as the hazard rate in probability theory. The probability of a state transition can be written in the following form:

$$\begin{aligned}&\mathbf{P}[\text{there is one state transition in the interval } (t, t+\Delta)]\\ &= \frac{Q_i(x+\Delta) - Q_i(x)}{1 - Q_i(x)} = \lambda_i(x)\Delta + o(\Delta)\,,\end{aligned}$$

from which

$$\mathbf{P}[\text{there is no state transition in the interval } (t, t+\Delta)] = 1 - \lambda_i(x)\Delta + o(\Delta)\,.$$

Based on all of these expressions, $h_i(t, x)$ satisfies

$$h_i(t+\Delta, x) = \Big(1 - \lambda_i(x)\Delta + o(\Delta)\Big)\, h_i(t, x-\Delta).$$

From this difference equation we can go through the usual steps to obtain the partial differential equation for $h_i(t, x)$. First we move $h_i(t, x-\Delta)$ to the other side,

$$h_i(t+\Delta, x) - h_i(t, x-\Delta) = \Big(-\lambda_i(x)\Delta + o(\Delta)\Big)\, h_i(t, x-\Delta),$$

then we add and subtract $h_i(t, x)$,

$$h_i(t+\Delta, x) - h_i(t, x) + h_i(t, x) - h_i(t, x-\Delta) = \Big(-\lambda_i(x)\Delta + o(\Delta)\Big)\, h_i(t, x-\Delta),$$

and reorder the terms,

$$\frac{h_i(t+\Delta, x) - h_i(t, x)}{\Delta} + \frac{h_i(t, x) - h_i(t, x-\Delta)}{\Delta} = \Big(-\lambda_i(x) + \frac{o(\Delta)}{\Delta}\Big)\, h_i(t, x-\Delta).$$

Finally, the $\Delta \to 0$ transition results in

$$\frac{\partial h_i(t, x)}{\partial t} + \frac{\partial h_i(t, x)}{\partial x} = -\lambda_i(x)\, h_i(t, x). \tag{4.18}$$

This partial differential equation describes $h_i(t, x)$ for $x > 0$. The case of $x = 0$ requires a different treatment:

$$\mathbf{P}\left(X(t+\Delta)=i, Y(t)\leq\Delta\right)$$
$$=\sum_{k\in S,k\neq i}\int_{x=0}^{\infty}\mathbf{P}\left(X(t)=k, Y(t)=x, \text{one transition to state } i \text{ in } (t,t+\Delta)\right)\mathrm{d}x.$$

The probability that in the interval $(t,t+\Delta)$ the process moves from state k to state i is

$$\begin{aligned}&\mathbf{P}[\text{there is one state transition in the interval } (t,t+\Delta) \text{ from } k \text{ to } i]\\&\quad=\mathbf{P}[\text{one state transition in the interval } (t,t+\Delta)]\\&\qquad\times\mathbf{P}[\text{state transition from } k \text{ to } i \mid \text{one state transition in the interval } (t,t+\Delta)]\\&\quad=\frac{Q_k(x+\Delta)-Q_k(x)}{1-Q_k(x)}\cdot\frac{Q_{ki}(x+\Delta)-Q_{ki}(x)}{Q_k(x+\Delta)-Q_k(x)},\end{aligned}$$

where the second term is already known from Eq. (4.14). We can also introduce the intensity of transition from k to i:

$$\begin{aligned}\lambda_{ki}(x)&=\lim_{\Delta\to0}\frac{\mathbf{P}[\text{there is a transition in the interval } (t,t+\Delta) \text{ from } k \text{ to } i]}{\Delta}\\&=\lim_{\Delta\to0}\frac{Q_{ki}(x+\Delta)-Q_{ki}(x)}{\Delta(1-Q_k(x))}=\frac{Q'_{ki}(x)}{1-Q_k(x)}.\end{aligned}$$

The transition probability can be written in the form

$$\mathbf{P}[\text{there is a transition in the interval } (t,t+\Delta) \text{ from } k \text{ to } i]=\lambda_{ki}(x)\Delta+o(\Delta).$$

Using this we can write

$$\begin{aligned}\mathbf{P}(X(t+\Delta)=i, Y(t)\leq\Delta)&=h_i(t+\Delta,0)\Delta\\&=\sum_{k\in S,k\neq i}\int_{x=0}^{\infty}\left(\lambda_{ki}(x)\Delta+o(\Delta)\right)h_k(t,x)\,\mathrm{d}x,\end{aligned}$$

from which a multiplication with Δ and the $\Delta\to0$ transition result in

$$h_i(t,0)=\sum_{k\in S,k\neq i}\int_{x=0}^{\infty}\lambda_{ki}(x)\,h_k(t,x)\,\mathrm{d}x. \tag{4.19}$$

In summary, the method of supplementary variable allows for the analysis of the process $\{X(t),Y(t)\}$ through the function $h_i(t,x)$, which is given by a partial differential equation (4.18) for $x>0$ and a boundary equation (4.19) for $x=0$. Based on these equations and the initial distributions of $h_i(0,x)$ for $\forall i\in S$ numerical partial differential solutions methods can be applied to compute the transient behavior of a semi-Markov process.

Stationary Behavior If the limit $\lim_{t\to\infty} h_i(t,x) = h_i(x)$ exists for all states $i \in S$, then we can evaluate the limit $t \to \infty$ of Eqs. (4.18) and (4.19)

$$\frac{\mathrm{d}h_i(x)}{\mathrm{d}x} = -\lambda_i(x)\, h_i(x), \tag{4.20}$$

$$h_i(0) = \sum_{k\in S, k\neq i} \int_{x=0}^{\infty} \lambda_{ki}(x)\, h_k(x)\, \mathrm{d}x. \tag{4.21}$$

The solution of ordinary differential Eq. (4.20) is

$$h_i(x) = h_i(0)\mathrm{e}^{\int_{u=0}^{x} -\lambda_i(u)\, \mathrm{d}u},$$

where the unknown quantity is $h_i(0)$. It can be obtained from Eq. (4.21) as follows:

$$h_i(0) = \sum_{k\in S, k\neq i} \int_{x=0}^{\infty} \lambda_{ki}(x)\, h_k(0)\mathrm{e}^{\int_{u=0}^{x} -\lambda_k(u)\, \mathrm{d}u}\, \mathrm{d}x$$

$$= \sum_{k\in S, k\neq i} h_k(0) \int_{x=0}^{\infty} \lambda_{ki}(x)\mathrm{e}^{\int_{u=0}^{x} -\lambda_k(u)\, \mathrm{d}u}\, \mathrm{d}x,$$

where

$$\int_{x=0}^{\infty} \lambda_{ki}(x)\mathrm{e}^{\int_{u=0}^{x} -\lambda_k(u)\, \mathrm{d}u}\, \mathrm{d}x = \mathbf{P}\,(\text{after state } k \text{ the process moves to state } i) = \Pi_{ki}.$$

That is, we are looking for the solution of the linear system

$$h_i(0) = \sum_{k\in S, k\neq i} h_k(0)\, \Pi_{ki} \quad \forall i \in S$$

with the normalizing condition

$$\sum_{i\in S} \int_{x=0}^{\infty} h_i(x)\, \mathrm{d}x = 1,$$

where the normalizing condition is the sum of the stationary-state probabilities. From

$$\sum_{i\in S} \int_{x=0}^{\infty} h_i(x)\, \mathrm{d}x = \sum_{i\in S} h_i(0) \int_{x=0}^{\infty} \mathrm{e}^{\int_{u=0}^{x} -\lambda_i(u)\, \mathrm{d}u}\, \mathrm{d}x = \sum_{i\in S} h_i(0)\, \hat{\tau}_i = 1$$

and Eq. (4.16) we have that the required solution is

$$h_i(0) = \frac{\hat{P}_i}{\sum_j \hat{P}_j \hat{\tau}_j}.$$

4.6 Markov Regenerative Process

Definition 4.57. The $X(t)$ discrete-state, continuous-time, time-homogeneous stochastic process is a **Markov regenerative process** if there exists a random time series $T_0, T_1, T_2, \ldots$ ($T_0 = 0$) such that the $X(t)$ process possesses the Markov property at time $T_0, T_1, T_2, \ldots$ [23, 58] (Fig. 4.4).

Compared to the properties of semi-Markov processes, where the process possesses the Markov property at all state-transition points, the definition of Markov regenerative processes is less restrictive. It allows that at some state-transition point the process does not possess the Markov property, but the analysis of Markov regenerative processes is still based on the occurrence of time points where the process possesses the Markov property.

Since Definition 4.57 does not address the behavior of the process between the consecutive time points $T_0, T_1, T_2, \ldots$, Markov regenerative processes can be fairly general stochastic processes. In practice, the use of a renewal theorem for the analysis of these processes is meaningful only when the stochastic behavior between the consecutive time points $T_0, T_1, T_2, \ldots$ is easy to analyze.

A common method for analyzing Markov regenerative processes is based on the next time point with the Markov property (T_1).

Definition 4.58. The series of random variables $\{Y_n, T_n; n \geq 0\}$ is a time-homogeneous **Markov renewal series** if

$$\begin{aligned}
&\mathbf{P}\left(Y_{n+1} = y, T_{n+1} - T_n \leq t \mid Y_0, \ldots, Y_n, T_0, \ldots, T_n\right) \\
&= \mathbf{P}\left(Y_{n+1} = y, T_{n+1} - T_n \leq t \mid Y_n\right) \\
&= \mathbf{P}\left(Y_1 = y, T_1 - T_0 \leq t \mid y_0\right)
\end{aligned}$$

for all $n \geq 0$, $y \in S$, and $t \geq 0$.

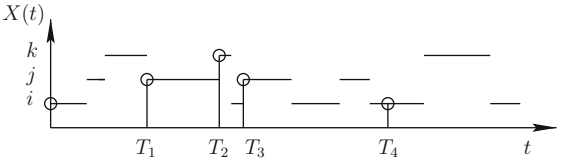


Fig. 4.4 Markov regenerative process; *circles* denote points with Markov property

It can be seen from the definition of Markov renewal series that the series $Y_0, Y_1, \ldots$ is a DTMC. According to Definition 4.57, the sequence of states $X(T_i)$ of a Markov regenerative process at the time sequence T_i instants with the Markov property and the time sequence T_i instants with the Markov property form a Markov renewal sequence $\{X(T_i), T_i\}$ $(i = 0, 1, \ldots)$.

Analysis of Markov regenerative processes is based on this *embedded* Markov renewal series. To this end the joint distribution of the next time point and the state in that time point must be known. In contrast with the similar kernel of semi-Markov processes, in the case of Markov regenerative processes, the kernel is denoted by

$$K_{ij}(t) = \mathbf{P}\left(X_1 = j, T_1 - T_0 \leq t \mid X_0 = i\right), \quad i, j \in S,$$

and the matrix $K(t) = \{K_{ij}(t)\}$ is referred to as the global kernel of a Markov regenerative process. The global kernel of a Markov regenerative process completely defines the stochastic properties of the Markov regenerative process at time points with the Markov property. The description of the process between those time points is complex, but for a transient analysis of the process (more precisely for computing transient-state probabilities) it is enough to know the transient-state probabilities between consecutive time points with the Markov property. This is given by the local kernel matrix of the Markov regenerative process $E(t) = \{E_{ij}(t)\}$ whose elements are

$$E_{ij}(t) = \mathbf{P}\left(X(t) = j,\, T_1 > t, \mid Z(0) = i\right),$$

where $E_{ij}(t)$ is the probability that the process will start in state i, the first point with the Markov property will be later than t, and the process will stay in state j at time t.

4.6.1 Transient Analysis Based on Embedded Markov Renewal Series

Let the transient-state transition probability matrix be $\Pi(t)$ whose elements are

$$\Pi_{ij}(t) = \mathbf{P}\left(X(t) = j \mid X(0) = i\right).$$

Assuming that $T_1 = h$, we can compute the conditional state-transition probability as follows:

$$\Pi_{ij}(t \mid T_1 = h) = \begin{cases} \mathbf{P}\left(X(t) = j \mid T_1 = h,\, X(0) = i\right), & h > t, \\ \sum\limits_{k \in S} \mathbf{P}\left(X(T_1) = k \mid X(0) = i, T_1 = h\right) \cdot \Pi_{kj}(t - h), & h \leq t. \end{cases} \tag{4.22}$$

Similar to the transient analysis of semi-Markov processes, Eq. (4.22) describes two exclusive cases: $h \leq t$ and $h > t$. In the case of semi-Markov processes, the $h > t$ case results in 0 or 1; in the case of a Markov regenerative process, the conditional probability for $h > t$ can be different from 0 or 1 because the process can have state transitions also before T_1.

Using the distribution of T_1 and the formula of total probability we obtain

$$\Pi_{ij}(t) = \int_{h=t}^{\infty} \mathbf{P}\left(X(t) = j \mid T_1 = h,\ X(0) = i\right) \mathrm{d}K_i(h) + \int_{h=0}^{t} \sum_{k \in S} \frac{dK_{ik}(t)}{dK_i(t)} \Pi_{kj}(t-h)\, \mathrm{d}K_i(h)\,. \tag{4.23}$$

Let us consider the first term on the right-hand side:

$$\begin{aligned}
&\int_{h=t}^{\infty} \mathbf{P}\left(X(t) = j \mid T_1 = h,\ X(0) = i\right) \mathrm{d}K_i(h) \\
&= \int_{h=t}^{\infty} \lim_{\Delta \to 0} \mathbf{P}\left(X(t) = j \mid h \leq T_1 < h + \Delta,\ X(0) = i\right)\ \mathrm{d}K_i(h) \\
&= \int_{h=t}^{\infty} \lim_{\Delta \to 0} \frac{\mathbf{P}\left(X(t) = j,\ h \leq T_1 < h + \Delta \mid X(0) = i\right)}{\mathbf{P}\left(h \leq T_1 < h + \Delta,\ \mid X(0) = i\right)}\ \mathrm{d}K_i(h) \\
&= \int_{h=t}^{\infty} \frac{d_h\, \mathbf{P}\left(X(t) = j,\ T_1 < h \mid X(0) = i\right)}{dK_i(h)}\ \mathrm{d}K_i(h) \\
&= \mathbf{P}\left(X(t) = j,\ t < T_1 \mid X(0) = i\right),
\end{aligned}$$

from which

$$\Pi_{ij}(t) \;=\; E_{ij}(t) \;+\; \sum_{k \in S} \int_{h=0}^{t} \Pi_{kj}(t-h)\, \mathrm{d}K_{ik}(h)\,. \tag{4.24}$$

Assuming that $K(t)$ is derivable and $dK(t)/dt = k(t)$ we have

$$\Pi_{ij}(t) \;=\; E_{ij}(t) \;+\; \sum_{k \in S} \int_{h=0}^{t} \Pi_{kj}(t-h)\, k_{ik}(h)\, \mathrm{d}h. \tag{4.25}$$

Similar to the transient analysis of CTMCs and semi-Markov processes we obtain a Volterra equation for the transient analysis of Markov regenerative processes.

Transform Domain Description The Laplace transform of Eq. (4.25) is

$$\Pi^*_{ij}(s) \;=\; E^*_{ij}(s) + \sum_{k \in \Omega} k^*_{ik}(s)\, \Pi^*_{kj}(s), \tag{4.26}$$

which can be written in matrix form:

$$\Pi^*(s) \;=\; E^*(s) \;+\; k^*(s)\,\Pi^*(s). \tag{4.27}$$

The solution of Eq. (4.27) is

$$\Pi^*(s) \;=\; [\mathbf{I} - k^*(s)]^{-1}\, E^*(s). \tag{4.28}$$

Based on Eq. (4.28), numerical inverse Laplace methods can also be used for the transient analysis of Markov regenerative processes.

Stationary Behavior Despite the differences between semi-Markov and Markov regenerative processes, their stationary analysis follows the same steps. The state-transition probability of the DTMC embedded in time points with the Markov property is

$$\Pi_{ij} = \mathbf{P}\left(Z(T_1) = j \mid Z(0) = i\right) = \lim_{t\to\infty} \mathbf{P}\left(Z(T_1) = j, T_1 \le t \mid Z(0) = i\right) = \lim_{t\to\infty} K_{ij}(t).$$

The stationary distribution of the EMC is the solution of $\hat{P} = \hat{P}\,\Pi, \sum_i \hat{P}_i = 1$. Now we need to compute the mean time spent in the different states during the interval (T_0, T_1). Fortunately, the local kernel carries the necessary information. Let τ_{ij} be the mean time the process spends in state j during the interval (T_0, T_1) assuming that it starts from state i $(X(T_0) = i)$. Then

$$\begin{aligned}
\tau_{ij} &= \mathbf{E}\left(\int_{t=0}^{\infty} \mathcal{I}_{\{X(t)=j, T_1>t \;\mid\; X(0)=i\}} \mathrm{d}t\right) \\
&= \int_{t=0}^{\infty} \mathbf{P}\left(X(t) = j, T_1 > t \mid X(0) = i\right) \mathrm{d}t \\
&= \int_{t=0}^{\infty} E_{ij}(t) \mathrm{d}t,
\end{aligned}$$

where $\mathcal{I}_{\{\bullet\}}$ is the indicator of event $\bullet$. The mean length of the interval (T_0, T_1) is

$$\tau_i = \sum_{j\in S} \tau_{ij}.$$

Finally, the stationary distribution of the process can be computed as

$$P_i = \frac{\displaystyle\sum_{j\in S} \hat{P}_j\; \tau_{ji}}{\displaystyle\sum_{j\in S} \hat{P}_j\; \tau_j}.$$

4.7 Exercises

Exercise 4.1. Applying Theorem 4.42, find the limit (stationary) distributions of age, residual lifetime, and total lifetime [$\delta(t) = t - t_{N(t)}$, $\gamma(t) = t_{N(t)+1} - t$, $\beta(t) = t_{N(t)+1} - t_{N(t)}$] if the interarrival times are independent random variables having a joint exponential distribution with the parameter λ. Show the expected values for the limit distributions.

Exercise 4.2 (*Ergodic property of semi-Markov processes).* Consider a system with the finite state space $\mathcal{X} = \{1, \ldots, N\}$. The system begins to work at the moment $T_0 = 0$ in a state $X_0 \in \mathcal{X}$ and changes states at the random moments $0 < T_1 < T_2 < \ldots$. Denote by $X_1, X_2, \ldots$ the sequence of consecutive states of the system, and suppose that it constitutes a homogeneous, irreducible, and aperiodic Markov chain with initial distribution ($p_i = \mathbf{P}(X_0 = i)$, $1 \le i \le N$) and probability transition matrix $\Pi = \left(p_{ij}\right)_{i,j=1}^{n}$. Define the process $X(t) = X_{n-1}$, $T_{n-1} \le t < T_n$, $n = 1, 2, \ldots$, assume that the sequence of holding times $Y_k = T_k - T_{k-1}$, $k = 1, 2, \ldots$, depends only conditionally on the states $X_{k-1} = i$ and $X_k = j$, and denote $F_{ij}(x) = \mathbf{P}(Y_k \le x \mid X_{k-1} = i,\ X_k = j)$ if $p_{ij} > 0$, where $\nu_{ij} = \int_0^\infty x \mathrm{d}F_{ij}(x) < \infty$.

Find the limits for

(a) The average number of transitions/time;
(b) The relative frequencies of states i in the sequence $X_0, X_1, \ldots$;
(c) The limit distribution $\mathbf{P}(X_t = i)$, $i \in X$;
(d) The average time spent in a state $i \in X$.

Chapter 5
Markov Chains with Special Structures

The previous chapter presented methods for analyzing stochastic models where some of the distributions were other than exponential. In these cases the analysis of the models is more complex than the analysis of Markov models. In this chapter we introduce a methodology to extend the set of models that can be analyzed by Markov models while the distributions can be other than exponential.

5.1 Phase Type Distributions

Combination of exponential distributions, such as convolution and probabilistic mixtures, was used for a long time to approximate nonexponential distributions such that the composed model remained a Markov model. The most general class of distributions fulfilling these requirement is the set of phase-type distributions (commonly abbreviated as PH distributions) [73, 74].

Definition 5.1. Time to absorption in a Markov chain with N transient and 1 absorbing state is *phase-type* distributed (cf. Fig. 5.1).

5.1.1 Continuous-Time PH Distributions

Definition 5.1 is valid for both CTMCs and DTMCs. In this section we focus on the case of CTMCs.

It is possible to define a PH distribution by defining the initial probability vector $\boldsymbol{p}$ and the generator matrix $\boldsymbol{Q}$ of a Markov chain with $N+1$ states. Let states of the Markov chain be numbered so that the first N states are transient and the $N+1$th is absorbing and let $X(t)$ be the state of the Markov chain at time t. The distributions of the time to absorption, T, is related to the transient probabilities of the Markov

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, DOI 10.1007/978-1-4614-5317-8_5,

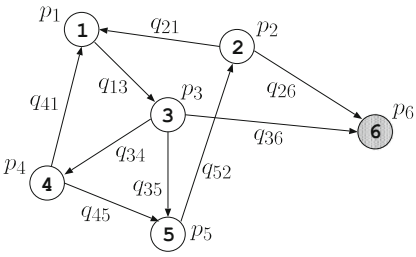


Fig. 5.1 Markov chain with five transient and an absorbing states defines a PH distribution

chain, which can be computed from the initial probability vector and the generator matrix as follows:

$$\mathbf{P}(T < t) = \mathbf{P}(X(t) = N + 1) = \boldsymbol{p}\mathrm{e}^{\boldsymbol{Q}t}e_{N+1}^T,$$

where e_{N+1} is the row vector whose only nonzero element the $N + 1$th is one. A multiplication of a row vector with e_{N+1}^T results in the $N + 1$th element of the row vector.

Analysis of PH distributions based on this expression results in technical difficulties in more complex cases. A more convenient expression can be derived from the partitioned generator matrix, where the set of states is divided into transient states and an absorbing one

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{A} & \mathbf{a} \\ \mathbf{0} & 0 \end{bmatrix},$$

where $\mathbf{a} = -\boldsymbol{A}\mathbb{1}$ and $\mathbb{1}$ is a column vector whose elements equal to one. The size of $\mathbb{1}$ is always assumed to be such that the multiplication is valid. A multiplication of a row vector by $\mathbb{1}$ results in the sum of the elements of the row vector. A column vector $\mathbf{a}$ that contains the transition rates to the absorbing state (Fig. 5.1) can be computed from $\boldsymbol{A}$ due to the fact that the row sum of $\boldsymbol{Q}$ is zero. The last row of $\boldsymbol{Q}$ is zero because the state $N + 1$ is absorbing.

Matrix $\boldsymbol{A}$ is called a transient generator (or PH generator). It inherits its main properties from matrix $\boldsymbol{Q}$. The diagonal elements of $\boldsymbol{A}$ are negative, the nondiagonal elements are nonnegative, and the row sums of $\boldsymbol{A}$ are nonpositive. Due to the fact that the first N states are transient, matrix $\boldsymbol{A}$ is nonsingular, in contrast with matrix $\boldsymbol{Q}$, which is singular because $\boldsymbol{Q}\mathbb{1} = \mathbf{0}$.

In this book we restrict our attention to the case where a Markov chain starts from one of the transient states with probability one. In this case, the partitioned form of vector $\boldsymbol{p}$ is $\boldsymbol{p} = [\boldsymbol{\alpha} \mid 0]$. Based on the partitioned form of $\boldsymbol{p}$ and $\boldsymbol{Q}$, the CDF of the PH distribution is

$$F_T(t) = \mathbf{P}(T \le t) = \mathbf{P}(T < t) = \mathbf{P}(X(t) = N + 1) = 1 - \mathbf{P}(X(t) < N + 1)$$

$$= 1 - [\boldsymbol{\alpha} \mid 0]\,\mathrm{e}^{\boldsymbol{Q}t}\begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix} = 1 - [\boldsymbol{\alpha} \mid 0]\sum_{i=0}^{\infty}\frac{t^i}{i!}\begin{bmatrix} \boldsymbol{A} & \mathbf{a} \\ \mathbf{0} & 0 \end{bmatrix}^i\begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix}$$

$$= 1 - [\boldsymbol{\alpha} \mid 0] \sum_{i=0}^{\infty} \frac{t^i}{i!} \begin{pmatrix} \boldsymbol{A}^i & \mathcal{I}_{\{i>0\}} \boldsymbol{A}^{i-1} \mathbf{a} \\ \mathbf{0} & 0 \end{pmatrix} \begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix}$$

$$= 1 - [\boldsymbol{\alpha} \mid 0] \begin{bmatrix} \mathrm{e}^{\boldsymbol{A}t} & \bullet \\ \mathbf{0} & 0 \end{bmatrix} \begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix} = 1 - \boldsymbol{\alpha}\, \mathrm{e}^{\boldsymbol{A}t} \mathbb{1},$$

where $\bullet$ denotes an irrelevant matrix block. Furthermore the PDF, the Laplace transform, and the moments of the PH distribution can be computed as

$$f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_T(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{e}^{\boldsymbol{A}t} \mathbb{1} = -\boldsymbol{\alpha} \sum_{i=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}t} \frac{t^i}{i!} \boldsymbol{A}^i \mathbb{1}$$

$$= -\boldsymbol{\alpha} \sum_{i=1}^{\infty} \frac{t^{i-1}}{(i-1)!} \boldsymbol{A}^{i-1} \boldsymbol{A} \mathbb{1} = -\boldsymbol{\alpha}\, \mathrm{e}^{\boldsymbol{A}t}\, \boldsymbol{A} \mathbb{1} = \boldsymbol{\alpha}\, \mathrm{e}^{\boldsymbol{A}t} \boldsymbol{a},$$

$$f_T^*(s) = \int_{t=0}^{\infty} \mathrm{e}^{-st} f_T(t) \mathrm{d}t = \boldsymbol{\alpha} \int_{t=0}^{\infty} \mathrm{e}^{-st} \mathrm{e}^{\boldsymbol{A}t} \mathrm{d}t\, \boldsymbol{a}$$

$$= \boldsymbol{\alpha} \int_{t=0}^{\infty} \mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})t} \mathrm{d}t\, \boldsymbol{a} = \boldsymbol{\alpha}\, (s\boldsymbol{I} - \boldsymbol{A})^{-1} \boldsymbol{a},$$

$$\mathbf{E}\,(T^n) = \int_{t=0}^{\infty} t^n f_T(t) \mathrm{d}t = \boldsymbol{\alpha} \int_{t=0}^{\infty} t^n \mathrm{e}^{\boldsymbol{A}t} \mathrm{d}t\, \boldsymbol{a} = \boldsymbol{\alpha}\, n! (-\boldsymbol{A})^{-n-1}\, \boldsymbol{a}$$

$$= \boldsymbol{\alpha}\, n! (-\boldsymbol{A})^{-n-1}\, (-\boldsymbol{A}) \mathbb{1} = \boldsymbol{\alpha}\, n! (-\boldsymbol{A})^{-n} \mathbb{1}.$$

The infinite integrals of the preceding derivations are computed as follows:

$$\int_{t=0}^{\infty} \mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})t} \mathrm{d}t = \lim_{\tau \to \infty} \int_{t=0}^{\tau} \mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})t} \mathrm{d}t = \lim_{\tau \to \infty} \int_{t=0}^{\tau} \sum_{i=0}^{\infty} \frac{t^i}{i!} (-s\boldsymbol{I} + \boldsymbol{A})^i \, \mathrm{d}t$$

$$= \lim_{\tau \to \infty} \sum_{i=0}^{\infty} \frac{\tau^{i+1}}{(i+1)!} (-s\boldsymbol{I} + \boldsymbol{A})^{(i+1)}\, (-s\boldsymbol{I} + \boldsymbol{A})^{-1}$$

$$= \lim_{\tau \to \infty} \left(\mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})\tau} - \boldsymbol{I} \right) (-s\boldsymbol{I} + \boldsymbol{A})^{-1} = (s\boldsymbol{I} - \boldsymbol{A})^{-1}, \quad (5.1)$$

where $\mathrm{e}^{(-s\boldsymbol{I}+\boldsymbol{A})\tau}$ vanishes in the convergence region of $f_T^*(s)$. The moments can also be computed from the Laplace transform

$$\begin{aligned}\mathbf{E}\left(T^n\right) &= (-1)^n \left.\frac{\mathrm{d}^n}{\mathrm{d}s^n} f_T^*(s)\right|_{s=0} = (-1)^n \boldsymbol{\alpha} \left.\frac{\mathrm{d}^n}{\mathrm{d}s^n}(s\boldsymbol{I}-\boldsymbol{A})^{-1}\right|_{s=0} \boldsymbol{a} \\ &= (-1)^n \boldsymbol{\alpha} \left.(-1)^n n!(s\boldsymbol{I}-\boldsymbol{A})^{-n-1}\right|_{s=0} \boldsymbol{a} = \boldsymbol{\alpha}\, n!(-\boldsymbol{A})^{-n-1}\, \boldsymbol{a} \\ &= \boldsymbol{\alpha}\, n!(-\boldsymbol{A})^{-n}\mathbb{1}.\end{aligned}$$

The elements of $(-\boldsymbol{A})^{-1}$ have an important stochastic interpretation. Let T_{ij} be the time spent in state j before moving to the absorbing state when the process starts in state i:

$$\begin{aligned}\mathbf{E}\left(T_{ij}\right) &= \int_{t=0}^{\infty} \mathbf{E}\left(\mathcal{I}_{\{X(t)=j\,|\,X(0)=i\}}\right)\mathrm{d}t = \int_{t=0}^{\infty} \mathbf{P}\left(X(t)=j\,|\,X(0)=i\right)\mathrm{d}t \\ &= \int_{t=0}^{\infty} \left(\mathrm{e}^{\boldsymbol{A}t}\right)_{ij}\mathrm{d}t = \left(\int_{t=0}^{\infty} \mathrm{e}^{\boldsymbol{A}t}\mathrm{d}t\right)_{ij} = \left((-\boldsymbol{A})^{-1}\right)_{ij}. \qquad (5.2)\end{aligned}$$

Consequently, $(-\boldsymbol{A})^{-1}$ is nonnegative. Some characteristics of PH distributions can be seen from these expressions. From

$$f^*(s) = \boldsymbol{\alpha}(s\boldsymbol{I}-\boldsymbol{A})^{-1}\mathbf{a} = \boldsymbol{\alpha}\left[\frac{\det(s\boldsymbol{I}-Ay)_{ji}}{\det(s\boldsymbol{I}-\boldsymbol{A})}\right]\mathbf{a}$$

we have that the Laplace transform is a rational function of s where the degree of the polynomial in the numerator is at most $N-1$ and in the denominator it is at most N, where N is the number of transient states and $\det(s\boldsymbol{I}-\boldsymbol{A})_{ji}$ denotes the subdeterminant associated with element i, j. The related properties of PH distributions in the time domain can be obtained from the spectral decomposition of $\boldsymbol{A}$. Let η be the number of eigenvalues of $\boldsymbol{A}$ and λ_i the ith eigenvalue whose multiplicity is η_i. In this case

$$f_T(t) = \boldsymbol{\alpha}\,\mathrm{e}^{\boldsymbol{A}t}\boldsymbol{a} = \sum_{i=1}^{\eta}\sum_{j=1}^{\eta_i} a_{ij}\, t^{j-1}\mathrm{e}^{\lambda_i t}.$$

This means that in the case of distinct eigenvalues ($\eta = N$, $\eta_i = 1$) $f_T(t)$ is a combination of exponential functions with possibly negative coefficients, and in the case of multiple eigenvalues $f_T(t)$ is a combination of exponential polynomial functions. As a consequence, as t goes to infinity, the exponential function associated with the eigenvalue with maximal real part dominates the density, meaning that PH distributions have asymptotically exponentially decaying tail behavior.

A wide range of positive distributions can be approximated with PH distributions of size N. A set of PH distributions approximating different positive distributions are depicted in Fig. 5.2. The exponentially decaying tail behavior is not visible in the figure, but there is another significant limitation of PH distributions of size N.

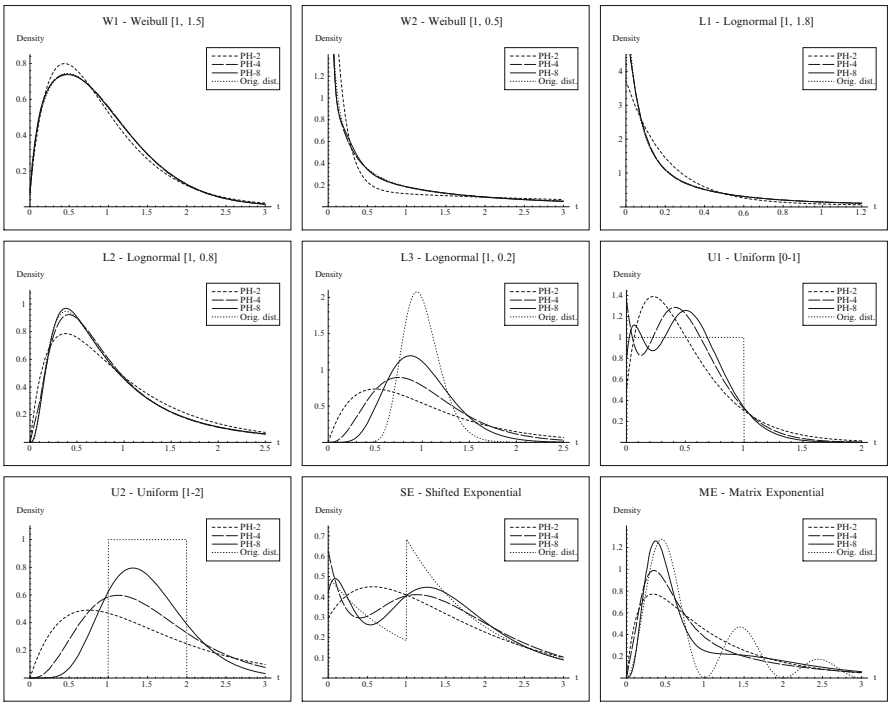


Fig. 5.2 Approximation of different positive distributions with $N = 2, 4, 8$ (figure copied from [13])

Theorem 5.2 ([3]). *The squared coefficient of variation of* T $(cv^2(\tau) = \mathbf{E}\left(T^2\right)/\mathbf{E}\left(T\right)^2)$ *satisfies*

$$cv^2(\tau) \geq \frac{1}{N},$$

and the only CPH distribution that satisfies the equality is the Erlang (N) distribution:

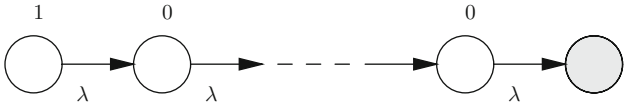


Figure 5.2 shows several distributions with low coefficient of variation whose approximation is poor due to this bound of the coefficient of variation. It is visible that PH distributions with larger N approximate these distributions significantly better. Theoretical results prove that as N tends to infinity, any positive distribution can be approximated arbitrarily closely.

5.1.2 Discrete-Time PH Distributions

The majority of the analysis steps and the properties of discrete-time PH distributions are similar to those of continuous-time PH distributions. Using a similar approach as for continuous-time PH distributions, the state-transition probability matrix can be partitioned as $\boldsymbol{P} = \begin{bmatrix} \boldsymbol{B} & \mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix}$, where $\mathbf{b} = \mathbb{1} - \boldsymbol{B}\mathbb{1}$ and the initial probability vector $\boldsymbol{p}$ as $\boldsymbol{p} = [\boldsymbol{\alpha} \mid 0]$. $\boldsymbol{B}$ is a sub-stochastic matrix, whose elements are nonnegative and row sums are not greater than one. The probability that the chain moves to the absorbing state in the kth step is

$$r_k = Pr(T = k) = \boldsymbol{\alpha}\boldsymbol{B}^{k-1}\mathbf{b},$$

which defines the probability mass function (PMF) of T. The CDF can be obtained as

$$F(k) = Pr(T \leq k) = Pr(X_k = N + 1) = 1 - Pr(X_k < N + 1) = 1 - \boldsymbol{\alpha}\boldsymbol{B}^k\mathbb{1},$$

and the z-transform or generator function of T is

$$\mathcal{F}(z) = \mathbf{E}\left(z^T\right) = \sum_{k=0}^{\infty} z^k r_k = z\,\boldsymbol{\alpha}(\boldsymbol{I} - z\boldsymbol{B})^{-1}\mathbf{b}.$$

The factorial moments are

$$\gamma_n = \mathbf{E}\left(T(T-1)\ldots(T-n+1)\right) = \frac{\mathrm{d}^n}{\mathrm{d}z^n}\mathcal{F}(z)|_{z=1} = n!\,\boldsymbol{\alpha}(\boldsymbol{I} - \boldsymbol{B})^{-n}\boldsymbol{B}^{n-1}\mathbb{1}.$$

Like the continuous-time case the z-transform is a rational function of z

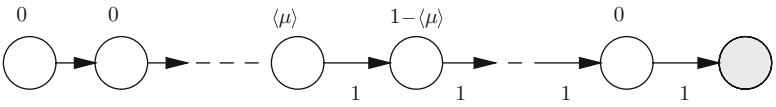
$$\mathcal{F}(z) = \mathbf{E}\left(z^T\right) = z\,\boldsymbol{\alpha}(\boldsymbol{I} - z\boldsymbol{B})^{-1}\mathbf{b} = z\,\boldsymbol{\alpha}\left[\frac{\det(\boldsymbol{I} - z\boldsymbol{B})_{ji}}{\det(\boldsymbol{I} - z\boldsymbol{B})}\right]\mathbf{b},$$

and based on the spectral decomposition of $\boldsymbol{B}$, the PMF is a combination of geometric series. The coefficient of variation of discrete PH (DPH) distributions is also bounded from below, but one of the most significant differences between the continuous and discrete PH distributions is that the bound in this case also depends on the mean of the distribution, $\mu = \mathbf{E}(T)$.

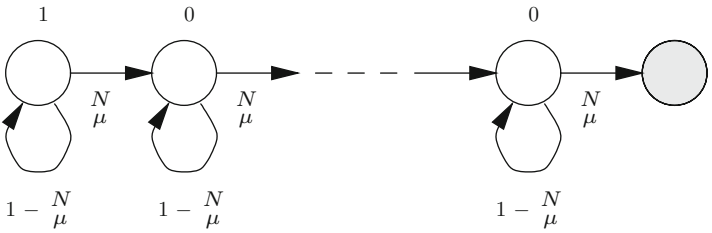
Theorem 5.3 ([92]). *The squared coefficient of variation of T satisfies the inequality*

$$cv^2(\tau) \geq \begin{cases} \dfrac{\langle\mu\rangle(1 - \langle\mu\rangle)}{\mu^2} & \text{if } \mu < N, \\ \dfrac{1}{N} - \dfrac{1}{\mu} & \text{if } \mu \geq N, \end{cases} \tag{5.3}$$

where $\langle x \rangle$ denotes the fraction part of x ($x = \lfloor x \rfloor + \langle x \rangle$). For $\mu \leq N$, $CV_{\min}$ is provided by the mixture of two deterministic distributions. Its DPH representation is



For $\mu > N$, $CV_{\min}$ is provided by the discrete Erlang distribution, whose DPH representation is



5.1.3 Special PH Classes

The set of PH distributions with N transient states is often too complex for particular practical applications (e.g., derivations by hand). There are special subclasses with restricted flexibility whose application is often more convenient. The most often used subclasses are

- Acyclic PH distributions,
- Hyper-Erlang distributions,
- Hyperexponential distributions ("parallel," "$cv > 1$").

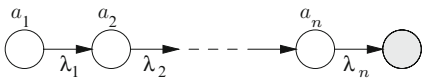
Acyclic PH Distributions

Definition 5.4. *Acyclic PH distributions* are PH distributions whose generator is an upper triangular matrix.

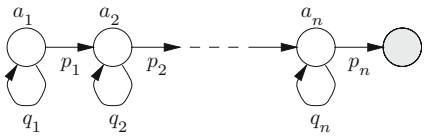
A direct consequence of the structural property of acyclic PH distributions is that the eigenvalues are explicitly given in the diagonal of the generator.

The practical applicability of acyclic PH distributions is due to the following result.

Theorem 5.5 ([25]). *Any acyclic PH distribution can be transformed into the following canonical form. In the case of continuous-time acyclic PH distributions:*



in the case of discrete-time acyclic PH distributions:



where the transition rates and probabilities are ordered such that $\lambda_i \leq \lambda_{i+1}$ *and* $p_i \leq p_{i+1}$.

This essential result allows one to consider only these canonical forms with $2N$ parameters to represent the whole acyclic PH class with N transient states.

Hyper-Erlang Distributions

Definition 5.6. A *hyper-Erlang distribution* is a probabilistic mixture of Erlang distributions.

Hyper-Erlang distributions are special acyclic PH distributions, and even fewer than $2N$ parameters can define them. Let ϑ be the number of Erlang branches, p_i the probability of taking branch i, and λ_i and n_i the parameters of the ith Erlang branch. These 3ϑ parameters completely define the hyper-Erlang distribution

$$f(t) = \sum_{i=1}^{\vartheta} p_i \, \frac{\lambda_i^{n_i} t^{n_i - 1} \mathrm{e}^{-\lambda_i t}}{(n_i - 1)!}.$$

Hyperexponential Distributions

Definition 5.7. A *hyperexponential distribution* is a probabilistic mixture of exponential distributions.

Hyperexponential distributions are special hyper-Erlang distributions where the order parameter of the Erlang distribution is one ($n_i = 1$). The PDF of hyperexponential distributions

$$f(t) = \sum_{i=1}^{\vartheta} p_i \lambda_i \mathrm{e}^{-\lambda_i t}$$

is monotonically decreasing due to the fact that it is the mixture of monotonically decreasing exponential density functions.

5.1.4 *Fitting with PH Distributions*

As was mentioned in the introduction of this chapter, PH distributions are often used to approximate experimental or exactly given but nonexponential positive distributions in order to analyze the obtained system behavior with discrete-state Markov chains. The engineering description of the fitting procedure is rather straightforward: given a nonnegative distribution or a set of experimental data, find a "similar" PH distribution, but for the practical implementation of this approach we need to answer several underlying questions. First we formalize the problem as an optimization problem:

$$\min_{\text{PHparameters}} \left\{ \text{Distance}(F_{PH}(t), \hat{F}_{\text{Original}}(t)) \right\},$$

that is, we optimize the parameters of the PH distribution such that the distance between the original distribution and the PH distribution is minimal. The two main technical problems are finding a proper distance measure and solving the optimization problem. Several solutions to these problems have been proposed in the literature, but there is room for further improvement. Some of the typical distance measures are

- Squared CDF difference: $\int_0^\infty (F(t) - \hat{F}(t))^2 \mathrm{d}t$;
- Density difference: $\int_0^\infty |f(t) - \hat{f}(t)| \mathrm{d}t$;
- Relative entropy: $\int_0^\infty f(t) \log\left(\frac{f(t)}{\hat{f}(t)}\right) \mathrm{d}t$.

The optimization problems according to these distance measures are typically nonlinear and numerically difficult. The close relation of the relative entropy measure with commonly applied statistical parameters (likelihood) makes this measure the most popular one in practice. It is worth mentioning that the complexity of the optimization procedures largely depends on the number of parameters of the PH distributions. That is why we discussed the number of parameters of the aforementioned special PH subclasses. A few implemented fitting procedures are available on the Internet. One fitting procedure that uses acyclic PH distributions is PhFit [43], and one using hyper-Erlang distributions is G-fit [93]. The literature of PH fitting is rather extended. Several other heuristic fitting approaches exist, e.g., combined with moment matching, that are left to the ambitions of interested readers.

5.2 Markov Arrival Process

A continuous-time Markov arrival process (MAP) is a generalization of a Poisson process such that the interarrival times are PH distributed and can be dependent. One of the simplest interpretations of MAPs considers a CTMC, $J(t)$, with N states and

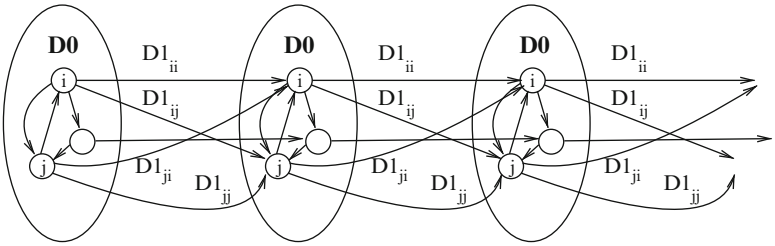


Fig. 5.3 Structure of Markov chain describing arrivals of a MAP

with generator $\boldsymbol{D}$, which determines the arrivals in the following way. While the Markov chain remains in state i, it generates arrivals according to a Poisson process at rate λ_i. When the Markov chain experiences a state transition from state i to j, then an arrival occurs with probability p_{ij} and does not occur with probability $1-p_{ij}$. Based on generator $\boldsymbol{D}$, rates λ_i $(i = 1, \ldots, N)$, and probabilities p_{ij} $(i, j = 1, \ldots, N, i \neq j)$, one can easily simulate the behavior of the MAP. Due to technical convenience MAPs are most commonly defined by a pair of matrices $\boldsymbol{D_0}$, $\boldsymbol{D_1}$, which are obtained from the previously introduced parameters in the following way:

$$\boldsymbol{D}_{0_{ij}} = \begin{cases} \boldsymbol{D}_{ij}(1-p_{ij}) & \text{if } i \neq j, \\ \boldsymbol{D}_{ii} - \lambda_i & \text{if } i = j, \end{cases} \qquad \boldsymbol{D}_{1_{ij}} = \begin{cases} \boldsymbol{D}_{ij}\, p_{ij} & \text{if } i \neq j, \\ \lambda_i & \text{if } i = j. \end{cases}$$

In this description, matrix $\boldsymbol{D}_0$ is associated with events that do not result in an arrival, and matrix $\boldsymbol{D}_1$ is associated with events that result in arrivals. By these definitions we have $\boldsymbol{D}_0 + \boldsymbol{D}_1 = \boldsymbol{D}$.

Based on these two matrices, we can investigate the counting process of arrivals. Let $N(t)$ be the number of arrivals of a MAP and $J(t)$ the state of the background Markov chain at time t. The $(N(t), J(t))$ $(N(t) \in \mathbb{N}, J(t) \in \{1, \ldots, N\})$ process is a CTMC. The transition structure of this Markov chain is depicted in Fig. 5.3. The set of states where the number of arrivals is n is commonly referred to as level n, and the state of the background Markov chain $(J(t))$ is commonly referred to as phase.

If the states are numbered in lexicographical order $((0, 1), \ldots, (0, N), (1, 1), \ldots, (1, N), \ldots)$, then the generator matrix has the form

$$Q = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{D_0} & \boldsymbol{D_1} & & & \\ \hline & \boldsymbol{D_0} & \boldsymbol{D_1} & & \\ \hline & & \boldsymbol{D_0} & \boldsymbol{D_1} & \\ \hline & & & \boldsymbol{D_0} & \boldsymbol{D_1} \\ \hline & & & & \ddots \\ \hline \end{array},$$

where the matrix blocks are of size N. Comparing this with the CTMC describing the number of arrivals of the Poisson process in Eq. (3.18), we have conspicuous similarities: only the diagonal elements/blocks and the first subdiagonal elements/blocks are nonzero, and the transition structure of the arrival process is independent of the number of arrivals.

It is commonly assumed that $N(0) = 0$, and thus the initial probability is 0 for all states (n, j) where $n > 0$. Let vector $\boldsymbol{\pi}_0$ be the initial probability for states with $n = 0$. The arrival instants are determined by $N(t)$ as follows: $\Theta_n = \min(t : N(t) = n)$, and the nth interarrival time is $T_n = \Theta_n - \Theta_{n-1}$. Based on the simple block structure of the CTMC, we can analyze the properties of $N(t)$ and T_n. For example, the distribution of T_1 is

$$\begin{aligned}\mathbf{P}(T_1 \le t) &= 1 - \mathbf{P}(T_1 > t) = 1 - \mathbf{P}(N(t) = 0) \\ &= 1 - \sum_{i=1}^{N} \mathbf{P}(N(t) = 0, J(t) = i) = 1 - \pi_0 \mathrm{e}^{\boldsymbol{D}_0 t} \mathbb{1},\end{aligned}$$

that is, T_1 is PH distributed with initial vector $\boldsymbol{\pi}$ and generator $\boldsymbol{D}_0$. For the analysis of the nth interarrival time we introduce the phase distributions vector after the $n-1$th arrivals, $\boldsymbol{\pi}_{n-1}$. The ith element of this vector is the probability that after the $n-1$th arrivals the background Markov chain is in state i, that is, $(\boldsymbol{\pi}_{n-1})_i = \mathbf{P}(J(\Theta_{n-1}) = i)$. Based on $\boldsymbol{\pi}_{n-1}$ the distribution of T_n is

$$\begin{aligned}\mathbf{P}(T_n \le t) &= 1 - \mathbf{P}(T_n > t) = 1 - \mathbf{P}(N(t + \Theta_{n-1}) = n - 1) \\ &= 1 - \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{P}(J(\Theta_{n-1}) = i)\, \mathbf{P}(N(t + \Theta_{n-1}) \\ &= 1 - \ss_{n-1} \mathrm{e}^{\boldsymbol{D}_0 t} \mathbb{1},\end{aligned}$$

that is, T_n is PH distributed with initial vector $\boldsymbol{\pi}_{n-1}$ and generator $\boldsymbol{D}_0$. The $\boldsymbol{\pi}_n$ vectors can be computed recursively. The ith element of $\boldsymbol{\pi}_1$ has the following stochastic interpretation:

$$\begin{aligned}(\boldsymbol{\pi}_1)_i &= \lim_{\Delta \to 0} \sum_{n=0}^{\infty} \sum_{j=1}^{N} \mathbf{P}(J(n\Delta) = j, T_1 > n\Delta) \\ &\quad \times \mathbf{P}(J((n+1)\Delta) = i, n\Delta < T_1 \le (n+1)\Delta) \\ &= \lim_{\Delta \to 0} \sum_{n=0}^{\infty} \sum_{j=1}^{N} \left(\boldsymbol{\pi} \mathrm{e}^{\boldsymbol{D}_0 n\Delta}\right)_j \left(\boldsymbol{D}_{1_{j,i}} \Delta + \sigma(\Delta)\right) \\ &= \int_{t=0}^{\infty} \sum_{j=1}^{N} \left(\ss \mathrm{e}^{\boldsymbol{D}_0 t}\right)_j \boldsymbol{D}_{1_{j,i}} \, \mathrm{d}t \,,\end{aligned}$$

where the first term on the right-hand side of the first row is the probability that there is no arrival up to time $n\Delta$ and the background Markov chain is in state j at time

$n\Delta$, and the second term on the right-hand side of the first row is the probability that there is an arrival between $n\Delta$ and $(n+1)\Delta$ such that the background Markov chain is in state i at time $(n+1)\Delta$. Using Eq. (5.1), we further have

$$\boldsymbol{ß}_1 = \boldsymbol{ß} \int_{t=0}^{\infty} \mathrm{e}^{\boldsymbol{D}_0 t}\, \mathrm{d}t\ \boldsymbol{D}_1 = \boldsymbol{ß}(-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1. \tag{5.4}$$

According to Eq. (5.4) we can compute the phase distribution after the first arrival from the initial distribution and the phase-transition probability matrix $\boldsymbol{P} = (-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1$. $\boldsymbol{P}$ is a stochastic matrix because from $(\boldsymbol{D}_0 + \boldsymbol{D}_1)\mathbb{1} = 0$ we have $-\boldsymbol{D}_0\mathbb{1} = \boldsymbol{D}_1\mathbb{1}$, from which $\boldsymbol{P}\mathbb{1} = (-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1\mathbb{1} = (-\boldsymbol{D}_0)^{-1}(-\boldsymbol{D}_0)\mathbb{1} = \mathbb{1}$, and $(-\boldsymbol{D}_0)^{-1}$ is nonnegative according to Eq. (5.2). Applying the same analysis for the nth interval starting with initial phase distribution $\boldsymbol{\pi}_{n-1}$ we have $\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1}\boldsymbol{P}$.

5.2.1 Properties of Markov Arrival Processes

The basic properties of MAPs or the $(N(t), J(t))$ CTMC [with level process $N(t) \in \mathbb{N}$ and phase process $J(t) \in \{1, \ldots, N\}$] are as follows.

- The phase distribution at arrival instants form a DTMC with transition probability matrix $\boldsymbol{P} = (-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1$. As a consequence, the phase distributions might be correlated at consecutive arrivals.
- The interarrival times are PH distributed with representation $(\boldsymbol{\pi}_0, \boldsymbol{D}_0)$, $(\boldsymbol{\pi}_1, \boldsymbol{D}_0)$, $(\boldsymbol{\pi}_2, \boldsymbol{D}_0)$, ... The interarrival times can be correlated due to the correlation of the initial phases.
- The phase process $(J(t))$ is a CTMC with generator $\mathbf{D} = \boldsymbol{D}_0 + \boldsymbol{D}_1$, which means that some properties of the phase process can be analyzed independent of the level process.
- The (time) stationary phase distribution $\boldsymbol{\alpha}$ is the solution of $\boldsymbol{\alpha}\mathbf{D} = 0, \boldsymbol{\alpha}\mathbb{1} = 1$.
- The (embedded) stationary phase distribution right after an arrival $\boldsymbol{\pi}$ is the solution of $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}, \boldsymbol{\pi}\mathbb{1} = 1$.
- These stationary distributions are closely related. On the one hand, the row vector of the mean time spent in the different phases during the stationary interarrival interval is $\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}$ [cf. Eq. (5.2)], from which the portion of time spent in the phases is

$$\boldsymbol{\alpha} = \frac{\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}}{\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}\mathbb{1}}.$$

  On the other hand, when the phase process is (time) stationary, the arrival intensities resulting in different initial phases for the next interarrival period are given by $\boldsymbol{\alpha}\boldsymbol{D}_1$, and after normalizing the result we have

$$\boldsymbol{\pi} = \frac{\boldsymbol{\alpha}\boldsymbol{D}_1}{\boldsymbol{\alpha}\boldsymbol{D}_1\mathbb{1}}.$$

- The stationary interarrival time (T) is PH distributed with representation $(\boldsymbol{\pi}, \boldsymbol{D}_0)$, and its nth moment is $\mathbf{E}(T^n) = n!\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-n}\mathbb{1}$.
- The stationary arrival intensity can be computed both from $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ as follows:

$$\lambda = \boldsymbol{\alpha}\boldsymbol{D}_1\mathbb{1} = \frac{1}{\mathbf{E}(X)} = \frac{1}{\boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}\mathbb{1}}.$$

  The first equality is based on the arrival intensities in the (time) stationary phase process. The second equality is based on the mean stationary interarrival time.

Further properties of stationary MAPs can be computed from the joint density functions of consecutive interarrivals:

$$f_{T_0,T_1,\ldots,T_k}(x_0,\ldots,x_k) = \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{D}_0 x_0}\boldsymbol{D}_1\mathrm{e}^{\boldsymbol{D}_0 x_1}\boldsymbol{D}_1\ldots\mathrm{e}^{\boldsymbol{D}_0 x_k}\boldsymbol{D}_1\mathbb{1}.$$

This joint density function describes the probability density that the process starts in phase i with probability $\boldsymbol{\pi}_i$ at time 0, it does not generate an arrival until time x_0, and an arrival occurs at x_0 according to the arrival intensities in $\boldsymbol{D}_1$. This arrival results in the second interarrival period's starting in phase j, and so on. If the MAP starts from a different initial phase distribution, e.g., γ, then the stationary embedded phase distribution vector $\boldsymbol{\pi}$ needs to be replaced by γ and the same joint density function applies. For example, we can compute the joint pdf of T_0 and T_k as

$$\begin{aligned} f_{T_0,T_k}(x_0,x_k) &= \int_{x_1}\ldots\int_{x_{k-1}} f_{T_0,T_1,\ldots,T_k}(x_0,\ldots,x_k)\,\mathrm{d}x_{k-1}\ldots\mathrm{d}x_1 \\ &= \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{D}_0 x_0}\boldsymbol{D}_1\boldsymbol{P}^{k-1}\mathrm{e}^{\boldsymbol{D}_0 x_k}\boldsymbol{D}_1\mathbb{1}, \end{aligned}$$

where we used that $\int_x \mathrm{e}^{\boldsymbol{D}_0 x}\mathrm{d}x = (-\boldsymbol{D}_0)^{-1}$ according to Eq. (5.1). This expression indicates that T_0 and T_k are dependent due to their dependent initial phases. It is also visible that as k tends to infinity, this dependency vanishes according to the speed at which the Markov chain of the initial vectors with transition probability matrix $\boldsymbol{P}$ converges to its stationary distribution $\boldsymbol{\pi}$.

The lag-k correlation of a MAP can be computed based on $f_{T_0,T_k}(x_0,x_k)$ as follows:

$$\begin{aligned} \mathbf{E}(T_0T_k) &= \int_{t=0}^{\infty}\int_{\tau=0}^{\infty} t\ \tau\ \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{D}_0 t}\boldsymbol{D}_1\boldsymbol{P}^{k-1}\mathrm{e}^{\boldsymbol{D}_0\tau}\boldsymbol{D}_1\mathbb{1}\,\mathrm{d}\tau\,\mathrm{d}t \\ &= \boldsymbol{\pi}(-\boldsymbol{D}_0)^{-2}\boldsymbol{D}_1\boldsymbol{P}^{k-1}(-\boldsymbol{D}_0)^{-2}\underbrace{\boldsymbol{D}_1\mathbb{1}}_{-\boldsymbol{D}_0\mathbb{1}} \\ &= \boldsymbol{\pi}(-\boldsymbol{D}_0)^{-1}\boldsymbol{P}^{k}(-\boldsymbol{D}_0)^{-1}\mathbb{1} = \frac{1}{\lambda}\boldsymbol{\alpha}\boldsymbol{P}^{k}(-\boldsymbol{D}_0)^{-1}\mathbb{1}, \end{aligned}$$

since

$$\int_{t=0}^{\infty} t\ \mathrm{e}^{\boldsymbol{D}_0 t}\mathrm{d}t = \underbrace{\left[t\ (\boldsymbol{D}_0)^{-1}\mathrm{e}^{\boldsymbol{D}_0 t}\right]_0^{\infty}}_{0} - \int_{t=0}^{\infty} (\boldsymbol{D}_0)^{-1}\ \mathrm{e}^{\boldsymbol{D}_0 t}\mathrm{d}t$$

and

$$\int_{t=0}^{\infty} \mathrm{e}^{\boldsymbol{D}_0 t}\mathrm{d}t = \lim_{T\to\infty} \sum_{i=0}^{\infty} \frac{\boldsymbol{D}_0^i}{i!} \int_0^T t^i \mathrm{d}t = \lim_{T\to\infty} \sum_{i=0}^{\infty} \frac{\boldsymbol{D}_0^i}{i!} \frac{T^{i+1}}{i+1}$$

$$= \lim_{T\to\infty} (\boldsymbol{D}_0)^{-1} \left(\underbrace{\mathrm{e}^{\boldsymbol{D}_0 T}}_{\to 0} - I \right) = (-\boldsymbol{D}_0)^{-1}\ .$$

Based on $\mathbf{E}\,(T_0 T_k)$ the covariance is

$$\mathrm{Cov}(T_0, T_k) = \mathbf{E}\,(T_0 T_k) - \mathbf{E}\,(T)^2 = \frac{1}{\lambda}\,\boldsymbol{\alpha}\boldsymbol{P}^k(-\boldsymbol{D}_0)^{-1}\mathbb{1} - \frac{1}{\lambda^2},$$

and the coefficient of correlation is

$$\mathrm{Corr}(T_0, T_k) = \frac{\mathrm{Cov}(T_0, T_k)}{\mathbf{E}\,(T^2) - \mathbf{E}\,(T)^2} = \frac{\frac{\mathbf{E}(T_0 T_k)}{\mathbf{E}(T)^2} - 1}{\frac{\mathbf{E}(T^2)}{\mathbf{E}(T)^2} - 1} = \frac{\lambda\,\boldsymbol{\alpha}\boldsymbol{P}^k(-\boldsymbol{D}_0)^{-1}\mathbb{1} - 1}{2\lambda\,\boldsymbol{\alpha}(-\boldsymbol{D}_0)^{-1}\mathbb{1} - 1}.$$

Starting from the joint density function of consecutive interarrivals we compute any joint moment for arbitrary series of interarrivals in a similar way as the lag-k correlation. For the interarrival series $a_0 = 0 < a_1 < a_2 < \ldots < a_k$ we have

$$f_{T_{a_0}, T_{a_1}, \ldots, T_{a_k}}(x_0, x_1, \ldots, x_k)$$
$$= \boldsymbol{\pi}\mathrm{e}^{\boldsymbol{D}_0 x_0}\boldsymbol{D}_1\boldsymbol{P}^{a_1-a_0-1}\mathrm{e}^{\boldsymbol{D}_0 x_1}\boldsymbol{D}_1\boldsymbol{P}^{a_2-a_1-1}\ldots\mathrm{e}^{\boldsymbol{D}_0 x_k}\boldsymbol{D}_1\mathbb{1},$$

and from that the joint moment $\mathbf{E}\left(T_{a_0}^{i_0}, T_{a_1}^{i_0}, \ldots, T_{a_k}^{i_0}\right)$ is

$$\mathbf{E}\left(T_{a_0}^{i_0}, T_{a_1}^{i_0}, \ldots, T_{a_k}^{i_0}\right)$$
$$= \boldsymbol{\pi} i_0!(-\boldsymbol{D}_0)^{-i_0}\boldsymbol{P}^{a_1-a_0} i_1!(-\boldsymbol{D}_0)^{-i_1}\boldsymbol{P}^{a_2-a_1}\ldots i_k!(-\boldsymbol{D}_0)^{-i_k}\mathbb{1}.$$

5.2.2 *Examples of Simple Markov Arrival Processes*

In this section we describe some basic arrival processes with MAP notations.

- PH renewal process: Consider an arrival process whose interarrival times are independent PH distributed with representation $(\boldsymbol{\alpha}, \boldsymbol{A})$. This is a special MAP characterized by $\boldsymbol{D}_0 = \boldsymbol{A}$, $\boldsymbol{D}_1 = \mathbf{a}\boldsymbol{\alpha}$.

- Interrupted Poisson process (IPP): Consider an arrival process determined by a background CTMC with two states, ON and OFF. The transition rate from ON to OFF is α and from OFF to ON it is β. There is no arrival in state OFF, and customers arrive according to a Poisson process at rate λ in state ON. The MAP description of the process is

$$\boldsymbol{D}_0 = \begin{array}{|c|c|}\hline -\alpha-\lambda & \alpha \\ \hline 0 & -\beta \\ \hline \end{array}, \quad \boldsymbol{D}_1 = \begin{array}{|c|c|}\hline \lambda & 0 \\ \hline 0 & 0 \\ \hline \end{array}.$$

- Markov modulated Poisson process (MMPP): Consider the arrival process determined by a background CTMC with generator $\boldsymbol{Q}$. While the CTMC is in state i, arrivals occur according to a Poisson process at rate λ_i. Let $\boldsymbol{\lambda}$ be the vector of arrival rates. This is a special MAP with representation $\boldsymbol{D}_0 = \boldsymbol{Q} - \text{diag}\langle\boldsymbol{\lambda}\rangle$, $\boldsymbol{D}_1 = \text{diag}\langle\boldsymbol{\lambda}\rangle$.
- Filtered MAP: Consider a MAP with representation $\hat{\boldsymbol{D}}_{\mathbf{0}}$, $\hat{\boldsymbol{D}}_{\mathbf{1}}$. The arrivals of this MAP are discarded with probability p. The obtained process is a MAP with representation $\boldsymbol{D}_0 = \hat{\boldsymbol{D}}_{\mathbf{0}} + p\hat{\boldsymbol{D}}_{\mathbf{1}}$, $\boldsymbol{D}_1 = (1-p)\hat{\boldsymbol{D}}_{\mathbf{1}}$.
- Cyclically filtered MAP: In the previous example, every MAP arrival is discarded with probability p. Now we consider the same MAP such that only every second arrival is discarded with probability p. It requires that we keep track of odd and even arrivals of the original MAP. It can be done by duplicating the phases such that the first half of them represents odd arrivals of the original MAP and the second half of them the even arrivals of the original MAP. The obtained process is a MAP with representation

$$\boldsymbol{D}_0 = \begin{array}{|c|c|}\hline \hat{\boldsymbol{D}}_{\mathbf{0}} & 0 \\ \hline p\hat{\boldsymbol{D}}_{\mathbf{1}} & \hat{\boldsymbol{D}}_{\mathbf{0}} \\ \hline \end{array}, \quad \boldsymbol{D}_1 = \begin{array}{|c|c|}\hline 0 & \hat{\boldsymbol{D}}_{\mathbf{1}} \\ \hline (1-p)\hat{\boldsymbol{D}}_{\mathbf{1}} & 0 \\ \hline \end{array}.$$

- Superposition of MAPs: Consider two MAPs with representation $\hat{\boldsymbol{D}}_{\mathbf{0}}$, $\hat{\boldsymbol{D}}_{\mathbf{1}}$ and $\tilde{\boldsymbol{D}}_{\mathbf{0}}$, $\tilde{\boldsymbol{D}}_{\mathbf{1}}$. The superposition of their arrival processes is a MAP with

$$\mathbf{D_0} = \hat{\boldsymbol{D}}_{\mathbf{0}} \bigoplus \tilde{\boldsymbol{D}}_{\mathbf{0}}, \text{ and } \mathbf{D_1} = \hat{\boldsymbol{D}}_{\mathbf{1}} \bigoplus \tilde{\boldsymbol{D}}_{\mathbf{1}},$$

where the Kronecker product is defined as $\boldsymbol{A} \bigotimes \boldsymbol{B} = \begin{array}{|c|c|}\hline A_{11}\boldsymbol{B} & \ldots\ A_{1n}\boldsymbol{B} \\ \hline \vdots & \vdots \\ \hline A_{n1}\boldsymbol{B} & \ldots\ A_{nn}\boldsymbol{B} \\ \hline \end{array}$ and the Kronecker sum as $\boldsymbol{A} \bigoplus \boldsymbol{B} = \boldsymbol{A} \bigotimes \mathbf{I_B} + \mathbf{I_A} \bigotimes \boldsymbol{B}$. This example indicates one advantage of the $\mathbf{D_0}$, $\mathbf{D_1}$ description of MAPs. Using these matrices the description of the superposed process inherits the related property of the Cartesian product of independent Markov chains.
- Consider an arrival process where the interarrival time is either exponentially distributed with parameter λ_1 or with parameter λ_2 ($\lambda_1 \neq \lambda_2$). The arrivals are correlated such that an interarrival period with parameter λ_1 is followed by one

with parameter λ_1 with probability p or one with parameter λ_2 with probability $1-p$. The interarrival periods with parameter λ_2 follow the same behavior. The obtained process is a MAP with

$$\boldsymbol{D}_0 = \begin{array}{|c|c|}\hline -\lambda_1 & 0 \\ \hline 0 & -\lambda 2 \\ \hline \end{array}, \quad \boldsymbol{D}_1 = \begin{array}{|c|c|}\hline p\lambda_1 & (1-p)\lambda_1 \\ \hline (1-p)\lambda_2 & p\lambda_2 \\ \hline \end{array}.$$

Probability p has a very intuitive meaning in this model. If $p \to 1$, then the correlation of the consecutive interarrivals is increasing and vice versa.

5.2.3 Batch Markov Arrival Process

A batch Markov arrival process (BMAP) is an extension of MAP with batch arrivals. It has an interpretation similar to that of a MAP.

A CTMC with generator $\boldsymbol{D}$ determines arrivals in the following way. While the Markov chain stays in state i, arrivals of batch size k occur according to a Poisson process at rate $\lambda_i^{(k)}$. When the Markov chain experiences a state transition from state i to j, arrivals of batch size k occur with probability $p_{ij}^{(k)}$ and no arrival occurs with probability $1-\sum_k p_{ij}^{(k)}$. Generator $\boldsymbol{D}$, rates $\lambda_i^{(k)}$ $(i = 1, \ldots, N)$, and probabilities $p_{ij}^{(k)}$ $(i, j = 1, \ldots, N, i \neq j)$ determine the stationary behavior of BMAPs. Additionally, the initial distribution of the CTMC is needed for the analysis of the transient behavior. A BMAP is commonly described by matrices $\boldsymbol{D_k}$, which are obtained from the previously introduced parameters in the following way:

$$\boldsymbol{D_0}_{ij} = \begin{cases} \boldsymbol{D}_{ij}\left(1-\sum_k p_{ij}^{(k)}\right) & \text{if } i \neq j, \\ \boldsymbol{D}_{ii} - \sum_k \lambda_i^{(k)} & \text{if } i = j, \end{cases} \qquad \boldsymbol{D_k}_{ij} = \begin{cases} \boldsymbol{D}_{ij}\, p_{ij}^{(k)} & \text{if } i \neq j, \\ \lambda_i^{(k)} & \text{if } i = j. \end{cases}$$

Based on this description the $(N(t), J(t))$ $(N(t) \in \mathbb{N}, J(t) \in \{1, \ldots, N\})$ process is a CTMC with transition structure depicted in Fig. 5.4. If the states are numbered in lexicographical order $((0,1), \ldots, (0,N), (1,1), \ldots, (1,N), \ldots)$, then the generator matrix has the form

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|}\hline \boldsymbol{D}_0 & \boldsymbol{D}_1 & D_2 & D_3 & D_4 \\ \hline & \boldsymbol{D}_0 & \boldsymbol{D}_1 & D_2 & D_3 \\ \hline & & \boldsymbol{D}_0 & \boldsymbol{D}_1 & D_2 \\ \hline & & & \boldsymbol{D}_0 & \boldsymbol{D}_1 \\ \hline & & & & \ddots \\ \hline \end{array},$$

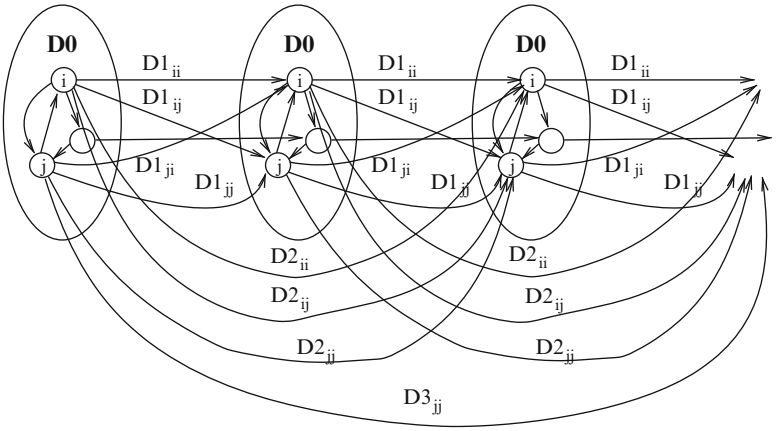


Fig. 5.4 Structure of Markov chain describing arrivals of a BMAP

To avoid complex cases it is commonly assumed that the considered BMAPs are regular:

- The phase process (**D**) is irreducible.
- The mean interarrival time is positive and finite ($\boldsymbol{D}_0$ nonsingular).
- The mean arrival rate, $\mathbf{d} = \sum_{k=0}^{\infty} k\mathbf{D_k}\mathbb{1}$, is finite.

BMAP properties are similar to MAP properties. We refer the reader to [62] for further details.

5.3 Quasi-Birth-Death Process

There are very few Markov chain structures that ensure solutions with convenient analytical properties. One of these few Markov chain structures is the quasi-birth-death (QBD) process.

Definition 5.8. A CTMC $\{N(t), J(t)\}$ with state space $\{n, j\}$ ($n \in \mathbb{N}$, $j \in \{1, \ldots, J\}$) is a *QBD process* if transitions are restricted to modify n by at most one and the transitions are homogeneous for different n values for $n \geq 1$, i.e., the transition rate from $\{n, j\}$ to $\{n', j'\}$ is zero if $|n - n'| \geq 2$ and the transition rate from $\{n, j\}$ to $\{n', j'\}$ equals the transition rate from $\{1, j\}$ to $\{n' - n + 1, j'\}$ (cf. Fig. 5.5).

These structural descriptions are relaxed subsequently by considering various versions of this basic regular QBD model. Similar to the case of MAPs, $N(t)$ is commonly referred to as a *level* process (it represents, e.g., the number of customers in a queue), and $J(t)$ is commonly referred to as a *phase* process (it represents,

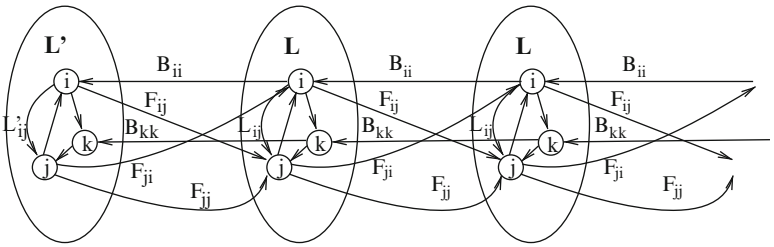


Fig. 5.5 Transition structure of QBD processes

e.g., the state of a randomly changing environment). Henceforth we assume that the considered QBD processes are irreducible with irreducible phase processes at the $n \geq 1$ levels (as detailed below).

Due to the structural properties of QBD processes, their state transitions can be classified as forward ($n \to n+1$), local ($n \to n$), and backward ($n \to n-1$). We apply the following notations:

- Matrix $\boldsymbol{F}$ of size $J \times J$ contains the rates of the forward transitions. The i, j element of $\boldsymbol{F}$ is the transition rate from $\{n, i\}$ to $\{n+1, j\}$ ($n \geq 0$).
- Matrix $\boldsymbol{L}$ of size $J \times J$ contains the rates of the local transitions for $n \geq 1$.
- Matrix $\boldsymbol{L}'$ of size $J \times J$ contains the rates of the local transitions for $n = 0$. Level 0 is irregular because there is no backward transition from level 0.
- Matrix $\boldsymbol{B}$ of size $J \times J$ contains the rates of the backward transitions. The i, j element of $\boldsymbol{F}$ is the transition rate from $\{n+1, i\}$ to $\{n, j\}$ ($n \geq 0$).

With these notations the structure of the generator matrix of a QBD process is

$$Q = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{L}' & \boldsymbol{F} & & & \\ \hline \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} & & \\ \hline & \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} & \\ \hline & & \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} \\ \hline & & & \ddots & \ddots \\ \hline \end{array} .$$

The name QBD process comes from the fact that on the matrix block level the generator matrix has a birth–death structure.

Condition of Stability

The phase process of a QBD process in the regular part ($n > 1$) is a CTMC with generator matrix $\boldsymbol{A} = \boldsymbol{F} + \boldsymbol{L} + \boldsymbol{B}$. Let $\boldsymbol{A}$ be irreducible with stationary distribution

$\boldsymbol{\alpha}$ (that is, $\boldsymbol{\alpha A} = \mathbf{0}, \boldsymbol{\alpha}\mathbb{1} = 1$). The drift associated with the stationary distribution of the regular phase process is $d = \boldsymbol{\alpha F}\mathbb{1} - \boldsymbol{\alpha B}\mathbb{1}$. The sign of this drift indicates whether the average tendency of the level process is increasing in the regular part. If $d < 0$, then the QBD process is positive recurrent [74]. That is, the condition of stability of QBD processes is $d = \boldsymbol{\alpha F}\mathbb{1} - \boldsymbol{\alpha B}\mathbb{1} < 0$, where $\boldsymbol{\alpha}$ is the solution of $\boldsymbol{\alpha}(\boldsymbol{F} + \boldsymbol{L} + \boldsymbol{B}) = \mathbf{0}, \boldsymbol{\alpha}\mathbb{1} = 1$.

5.3.1 Matrix-Geometric Distribution

The stationary solution of a QBD process with generator $\boldsymbol{Q}$ is the solution of the linear system of equations $\boldsymbol{\pi Q} = \mathbf{0}$, $\boldsymbol{\pi}\mathbb{1} = 1$, where $\boldsymbol{\pi}$ is the row vector of stationary probabilities. To utilize the regular structure of matrix $\boldsymbol{Q}$, we partition vector $\boldsymbol{\pi}$ according to the levels of the QBD process: $\boldsymbol{\pi} = \{\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots\}$. Using this partitioning the linear system of equations takes the following form:

$$\boldsymbol{\pi}_0 \boldsymbol{L}' + \boldsymbol{\pi}_1 \boldsymbol{B} = \mathbf{0}, \tag{5.5}$$

$$\boldsymbol{\pi}_{n-1} \boldsymbol{F} + \boldsymbol{\pi}_n \boldsymbol{L} + \boldsymbol{\pi}_{n+1} \boldsymbol{B} = \mathbf{0} \quad \forall n \geq 1, \tag{5.6}$$

$$\sum_{n=0}^{\infty} \boldsymbol{\pi}_n \mathbb{1} = 1. \tag{5.7}$$

Theorem 5.9. *The solution of Eqs. (5.5)–(5.7) in the case of a stable QBD process is* $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$*, where matrix* $\boldsymbol{R}$ *is the only solution of the quadratic matrix equation*

$$\boldsymbol{F} + \boldsymbol{RL} + \boldsymbol{R}^2 \boldsymbol{B} = \mathbf{0},$$

whose eigenvalues are inside the unit disk, and vector π_0 *is the solution of a linear system of size* J

$$\boldsymbol{\pi}_0 (\boldsymbol{L}' + \boldsymbol{RB}) = \mathbf{0}$$

with normalizing condition

$$\boldsymbol{\pi}_0 (\boldsymbol{I} - \boldsymbol{R})^{-1} \mathbb{1} = 1.$$

Proof. In the case of stable irreducible CTMCs, the solution of the linear system $\boldsymbol{\pi Q} = \mathbf{0}$, $\boldsymbol{\pi}\mathbb{1} = 1$ is unique and identical with the stationary distribution of the CTMC. In this proof we only show that $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$ satisfies the linear system and do not discuss the properties of the solutions of the quadratic matrix equations. The details of the spectral properties of the solutions are discussed, for example, in [62]. Substituting the $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$ solution into Eq. (5.6) gives

$$\boldsymbol{\pi}_0 \boldsymbol{R}^{n-1} \boldsymbol{F} + \boldsymbol{\pi}_0 \boldsymbol{R}^n \boldsymbol{L} + \boldsymbol{\pi}_0 \boldsymbol{R}^{n+1} \boldsymbol{B} = \boldsymbol{\pi}_0 \boldsymbol{R}^{n-1} (\boldsymbol{F} + \boldsymbol{RL} + \boldsymbol{R}^2 \boldsymbol{B}) = \mathbf{0} \quad \forall n \geq 1,$$

which holds according to the definition of $\boldsymbol{R}$. Due to the fact that the eigenvalues of $\boldsymbol{R}$ are inside the unit disk, the infinite sum $\sum_{n=0}^{\infty} \boldsymbol{R}^n$ is finite, and we have $\sum_{n=0}^{\infty} \boldsymbol{R}^n = (\boldsymbol{I} - \boldsymbol{R})^{-1}$. Using this and substituting the $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$ solution into Eqs. (5.5) and (5.7) gives

$$\boldsymbol{\pi}_0 \boldsymbol{L}' + \boldsymbol{\pi}_0 \boldsymbol{R} \boldsymbol{B} = \boldsymbol{0},$$

$$\sum_{n=0}^{\infty} \boldsymbol{\pi}_0 \boldsymbol{R}^n \mathbb{1} = \boldsymbol{\pi}_0 (\boldsymbol{I} - \boldsymbol{R})^{-1} \mathbb{1} = 1,$$

which is the linear system defining $\boldsymbol{\pi}_0$. □

The stationary distribution of the form $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \boldsymbol{R}^n$ are commonly referred to as matrix geometric distributions. This terminology refers also to the relation of homogeneous birth and death processes and QBD processes since the stationary distribution of homogeneous birth and death processes is geometric. Similar to the relation of Poisson processes and MAPs, QBD processes can be interpreted as an extension of birth and death processes such that their generator matrices have the same structure on the level of matrix blocks.

An extensive literature exists that deals with the properties of QBD processes and the efficient computation of matrix $\boldsymbol{R}$; therefore, we present here only two computational methods for matrix $\boldsymbol{R}$ and refer interested readers to [12] and references therein.

Linear algorithm

$\boldsymbol{R} := \boldsymbol{0}$;
REPEAT
  $\boldsymbol{R}_{old} := \boldsymbol{R}$;
  $\boldsymbol{R} := \boldsymbol{F} \left(-\boldsymbol{L} - \boldsymbol{R}\boldsymbol{B}\right)^{-1}$;
UNTIL $||\boldsymbol{R} - \boldsymbol{R}_{old}|| \leq \epsilon$

Logarithmic algorithm

$\mathbf{H} := \boldsymbol{F} \left(-\boldsymbol{L}\right)^{-1}$;
$\mathbf{K} := \boldsymbol{B} \left(-\boldsymbol{L}\right)^{-1}$;
$\boldsymbol{R} := \mathbf{H}$;
$\mathbf{T} := \mathbf{K}$;
REPEAT
  $\boldsymbol{R}_{old} := \boldsymbol{R}$;
  $\boldsymbol{U} := \mathbf{H}\mathbf{K} + \mathbf{K}\mathbf{H}$;
  $\mathbf{H} := \mathbf{H}^2 \left(\boldsymbol{I} - \boldsymbol{U}\right)^{-1}$;
  $\mathbf{K} := \mathbf{K}^2 \left(\boldsymbol{I} - \boldsymbol{U}\right)^{-1}$;
  $\boldsymbol{R} := \boldsymbol{R} + \mathbf{H}\mathbf{T}$;
  $\mathbf{T} := \mathbf{K}\mathbf{T}$;
UNTIL $||\boldsymbol{R} - \boldsymbol{R}_{old}|| \leq \epsilon$

The input data of these algorithms are matrices $\boldsymbol{F}$, $\boldsymbol{L}$, $\boldsymbol{B}$, and a predefined accuracy parameter ϵ. The main differences between the algorithms are that the linear algorithm has a simpler iteration step and is more sensitive to drift d. When the drift is close to 0, the linear algorithm performs a huge number of iterations. The properties of the logarithmic algorithm are different. It has a more complex iteration step, but the number of iterations is tolerable also for drift values close to 0.

The following sections present different QBD variants whose stationary distributions are different variants of the matrix geometric distribution.

5.3.2 Quasi-Birth-and-Death Process with Irregular Level 0

Many practical examples exist where the system has a regular behavior when it is in normal operation mode in some sense, but it has a different behavior (e.g., a different state transition structure or rates or even a different number of phases) when it is idle in some sense. Additionally, any CTMC that exhibits a regular QBD structure from a given point on can be considered a QBD process with irregular level 0, where level 0 is defined such that it contains the whole irregular part of the state space.

In general, a QBD process with irregular level 0 has the following block structure

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{L'} & \boldsymbol{F'} & & & \\ \hline \boldsymbol{B'} & \boldsymbol{L} & \boldsymbol{F} & & \\ \hline & \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} & \\ \hline & & \boldsymbol{B} & \boldsymbol{L} & \boldsymbol{F} \\ \hline & & & \ddots & \ddots \\ \hline \end{array},$$

where the sizes of the blocks are identical for levels $1, 2, \ldots$, but the sizes of the blocks at level 0 can be different from the regular block size. If J is the regular block size and J_0 the block size at level 0, then matrices $\boldsymbol{F}$, $\boldsymbol{L}$, and $\boldsymbol{B}$ are of size $J \times J$, matrix $\boldsymbol{F'}$ is of size $J_0 \times J$, matrix $\boldsymbol{L'}$ is of size $J_0 \times J_0$, and matrix $\boldsymbol{B'}$ is of size $J \times J_0$.

In this case, the partitioned form of the linear system $\boldsymbol{\pi Q} = \boldsymbol{0}$, $\boldsymbol{\pi}\mathbb{1} = 1$ is

$$\boldsymbol{\pi}_0\boldsymbol{L'} + \boldsymbol{\pi}_1\boldsymbol{B'} = \boldsymbol{0}, \tag{5.8}$$

$$\boldsymbol{\pi}_0\boldsymbol{F'} + \boldsymbol{\pi}_1\boldsymbol{L} + \boldsymbol{\pi}_2\boldsymbol{B} = \boldsymbol{0}, \tag{5.9}$$

$$\boldsymbol{\pi}_{n-1}\boldsymbol{F} + \boldsymbol{\pi}_n\boldsymbol{L} + \boldsymbol{\pi}_{n+1}\boldsymbol{B} = \boldsymbol{0} \quad \forall n \geq 2, \tag{5.10}$$

$$\sum_{n=0}^{\infty} \boldsymbol{\pi}_n\mathbb{1} = 1. \tag{5.11}$$

Theorem 5.10. *The solution of Eqs.* (5.8)–(5.11) *in the case of a stable QBD process is* $\boldsymbol{\pi}_0$ *and* $\boldsymbol{\pi}_n = \boldsymbol{\pi}_1 \boldsymbol{R}^{n-1}$ $(n \geq 1)$*, where matrix* $\boldsymbol{R}$ *is the only solution of the quadratic matrix equation*

$$\boldsymbol{F} + \boldsymbol{R}\boldsymbol{L} + \boldsymbol{R}^2\boldsymbol{B} = \boldsymbol{0}$$

whose eigenvalues are inside the unit disk and vectors $\boldsymbol{\pi}_0$, $\boldsymbol{\pi}_1$ *come from the solution of the linear system of size* $J_0 + J$

$$\boldsymbol{\pi}_0 \boldsymbol{L}' + \boldsymbol{\pi}_1 \boldsymbol{B}' = \boldsymbol{0},$$

$$\boldsymbol{\pi}_0 \boldsymbol{F}' + \boldsymbol{\pi}_1 (\boldsymbol{L}' + \boldsymbol{R}\boldsymbol{B}) = \boldsymbol{0},$$

with normalizing condition

$$\boldsymbol{\pi}_0 \mathbb{1} + \boldsymbol{\pi}_1 (\boldsymbol{I} - \boldsymbol{R})^{-1} \mathbb{1} = 1.$$

Proof. The proof follows the same pattern as that of Theorem 5.9. Substituting the matrix-geometric solution into the partitioned form of the stationary equations indicates that the solution satisfies the stationary equations. □

The linear system for $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ can be rewritten into the matrix form

$$[\boldsymbol{\pi}_0 | \boldsymbol{\pi}_1] \begin{array}{|c|c|} \hline \boldsymbol{L}' & \mathbf{F}' \\ \hline \mathbf{B}' & \boldsymbol{L} + \boldsymbol{R}\boldsymbol{B} \\ \hline \end{array} = [\,\boldsymbol{0} \mid \boldsymbol{0}\,].$$

5.3.3 Finite Quasi-Birth-and-Death Process

Another frequently applied variant of QBD processes is the case where the level process has an upper limit. When the upper limit is at level m, the generator matrix takes the form

$$Q = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{L}' & \boldsymbol{F} & & & \\ \hline \boldsymbol{B} & \boldsymbol{L} & \ddots & & \\ \hline & \boldsymbol{B} & \ddots & \boldsymbol{F} & \\ \hline & & \ddots & \boldsymbol{L} & \boldsymbol{F} \\ \hline & & & \boldsymbol{B} & \boldsymbol{L}'' \\ \hline \end{array},$$

and the partitioned form of the stationary equation is

$$\boldsymbol{\pi}_0 \boldsymbol{L}' + \boldsymbol{\pi}_1 \boldsymbol{B} = \mathbf{0}, \tag{5.12}$$

$$\boldsymbol{\pi}_{n-1} \boldsymbol{F} + \boldsymbol{\pi}_n \boldsymbol{L} + \boldsymbol{\pi}_{n+1} \boldsymbol{B} = \mathbf{0} \quad 1 \leq n \leq m-1, \tag{5.13}$$

$$\boldsymbol{\pi}_{m-1} \boldsymbol{F} + \boldsymbol{\pi}_m \boldsymbol{L}'' = \mathbf{0}, \tag{5.14}$$

$$\sum_{n=0}^{m} \boldsymbol{\pi}_n \mathbb{1} = 1. \tag{5.15}$$

Theorem 5.11. *The solution of Eqs. (5.12)–(5.15) in the case of a finite QBD process with $d < 0$ is $\boldsymbol{\pi}_n = \boldsymbol{\alpha} \boldsymbol{R}^n + \boldsymbol{\beta} \boldsymbol{S}^{m-n}$ $(0 \leq n \leq m)$, where matrix $\boldsymbol{R}$ is the only solution of the quadratic matrix equation*

$$\boldsymbol{F} + \boldsymbol{R}\boldsymbol{L} + \boldsymbol{R}^2 \boldsymbol{B} = \mathbf{0}$$

whose eigenvalues are inside the open unit disk, matrix $\boldsymbol{S}$ is the only solution of the quadratic matrix equation

$$\boldsymbol{B} + \boldsymbol{S}\boldsymbol{L} + \boldsymbol{S}^2 \boldsymbol{F} = \mathbf{0}$$

whose eigenvalues are on the closed unit disk, and vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the solution of the size $2J$ linear system

$$\boldsymbol{\alpha} \left(\boldsymbol{L}' + \boldsymbol{R}\boldsymbol{B} \right) + \boldsymbol{\beta} \boldsymbol{S}^{m-1} \left(\boldsymbol{S}\boldsymbol{L}' + \boldsymbol{B} \right) = \mathbf{0},$$

$$\boldsymbol{\alpha} \boldsymbol{R}^{m-1} \left(\boldsymbol{F} + \boldsymbol{R}\boldsymbol{L}'' \right) + \boldsymbol{\beta} \left(\boldsymbol{S}\boldsymbol{F} + \boldsymbol{L}'' \right) = \mathbf{0},$$

with normalizing condition

$$\boldsymbol{\alpha} \sum_{n=0}^{m} \boldsymbol{R}^n \mathbb{1} + \boldsymbol{\beta} \sum_{n=0}^{m} \boldsymbol{S}^n \mathbb{1} = 1.$$

Proof. The proof follows the same pattern as that of Theorem 5.9. Substituting the solution into the partitioned form of the stationary equations indicates that the solution satisfies the stationary equations. □

The matrix form of the linear system for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is

$$[\boldsymbol{\alpha} | \boldsymbol{\beta}] \begin{array}{|c|c|} \hline \boldsymbol{L}' + \boldsymbol{R}\boldsymbol{B} & \boldsymbol{R}^{m-1} \left(\boldsymbol{F} + \boldsymbol{R}\boldsymbol{L}'' \right) \\ \hline \boldsymbol{S}^{m-1} \left(\boldsymbol{S}\boldsymbol{L}' + \boldsymbol{B} \right) & \boldsymbol{S}\boldsymbol{F} + \boldsymbol{L}'' \\ \hline \end{array} = [\, \mathbf{0} \mid \mathbf{0} \,] .$$

Matrix $\boldsymbol{S}$ can be computed by the same linear or logarithmic procedures as matrix $\boldsymbol{R}$. If the drift is positive ($d > 0$) in a finite QBD process, then the numbering of the levels needs to be inverted ($0 \to m, 1 \to m-1, \ldots, m \to 0$), and we obtain a new finite QBD process whose drift is negative. It is worth mentioning that due to the fact that $d < 0$, matrix $\boldsymbol{S}$ has an eigenvalue on the unit circle, and consequently $\sum_{n=0}^{\infty} \boldsymbol{S}^n$ does not converge. Fortunately, this does not affect the applicability of Theorem 5.11 because we need to compute only the finite sum $\sum_{n=0}^{m} \boldsymbol{S}^n$.

5.4 Exercises

Exercise 5.1. X and Y are independent continuous-time PH distributed random variables with representations $(\boldsymbol{\alpha}, \boldsymbol{A})$ and $(\boldsymbol{\beta}, \boldsymbol{B})$, respectively. Define the distribution of the following random variables:

- $Z_1 = c_1 X$;
- Z_2 equals X with probability p and to Y with probability $1 - p$;
- $Z_3 = c_1 X + c_2 Y$;
- $Z_4 = \text{Min}(X, Y)$;
- $Z_5 = \text{Max}(X, Y)$.

Exercise 5.2. X and Y are independent discrete-time PH distributed random variables with representations $(\boldsymbol{\alpha}, \boldsymbol{A})$ and $(\boldsymbol{\beta}, \boldsymbol{B})$, respectively. Define the distribution of the following random variables:

- $Z_1 = c_1 X$;
- Z_2 equals to X with probability p and to Y with probability $1 - p$;
- $Z_3 = c_1 X + c_2 Y$;
- $Z_4 = \text{Min}(X, Y)$;
- $Z_5 = \text{Max}(X, Y)$.

Exercise 5.3. There are two machines, A and B, at a production site. Their failure times are exponentially distributed with parameters λ_A and λ_B, respectively. Their repair times are also exponentially distributed with parameters μ_A and μ_B, respectively. A lone repairman can work on only one machine at a time. At a given time, both machines work. Compute the distribution and the moments of the time to the first complete breakdown when both machines fail.

Part II
Queueing Systems

Chapter 6
Introduction to Queueing Systems

6.1 Queueing Systems

The theory of queueing systems dates back to the seminal work of A. K. Erlang (1878–1929), who worked for the telecom company in Copenhagen and studied telephone traffic in the early twentieth century. To this day the terminology of queueing theory is closely related to telecommunications (e.g., channel, call, idle/busy, queue length, utilization).

Due to the wide range of potential application fields (e.g., vehicular traffic, logistics, trade, banking, customer service, production lines, manufacturing systems, stock-in-trade) queueing theory has attracted attention and developed quickly. This attention is also apparent in the number of queueing-related publications. The queueing theory book of Saaty [82], published in 1961, contained 896 references, and its Russian translation, published in 1965, contained 1,115 references.

The early works of A.K. Erlang already contained the main elements of queueing theory: the (stochastic) arrival process of requests (calls), the (stochastic) service process of customers and, consequently, the departure process of customers, rejected/waiting customers, servers, etc. Later on real physical systems broke away from queueing and developed its own terminology and the theory was applied in a wide range of application fields using the aforementioned basic terminology.

The mathematical description of queueing systems requires a description of the following elements:

- Arrival process: the stochastic description of customer arrivals, where customers might have any abstract or physical meaning depending on the considered system.

  Customer arrivals might depend on the current system's properties, e.g., the number of customers in the system. In the case of basic queueing models (where the interarrival times are independent), the arrival process is characterized by the interarrival time distribution.
- Service process: the stochastic description of customer service.

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, 191
DOI 10.1007/978-1-4614-5317-8_6, © Springer Science+Business Media, LLC 2013

Like customer arrival, customer service might also depend on the current system's properties, and in basic queueing models the service times are i.i.d. random variables.

- System structure: the resources of the queueing system, typically the number of servers and the size of the waiting room.
- Service discipline: a set of rules that determines the service order and service mode of customers. The most common orders are FCFS (first come, first served), FIFO (first in, first out), and LIFO (last in, first out). Service resources can also be used to serve all customers in parallel. This discipline is referred to as processor sharing (PS). Service order plays an important role when different types of customers arrive at a system. In this case, priority (with and without preemptions) can be used to provide faster service to one customer type.
- Performance parameters: to build an appropriately detailed model of a system, one should consider those performance parameters that must be computed. The most common performance parameters are system utilization, mean and distribution of waiting time, loss probability (the probability that a customer will be rejected by the system), etc. These measures are precisely defined later.

6.2 Classification of Basic Queueing Systems

The same queueing models might appear in completely different application fields. To avoid the parallel development of the same models in different fields, in 1953, Kendall proposed a classification and a standard notation of basic queueing systems. The current version of this set of notations is composed by six elements – A/B/c/d/e-x – where

A is the type of arrival process;
B is the type of service process;
c is the number of servers;
d is the system capacity, the maximum number of customers in the system;
e is the population of the set of customers (if it is finite, the arrival intensity decreases with an increasing number of customers in the system); and
x defines the service discipline (the most common service disciplines – e.g., FCFS, LCFS, PS – are defined above).

In basic queueing systems A and B take one of the following options:

M – memoryless, refers to exponentially distributed interarrival or service time;
E_r – order r Erlang distributed interarrival or service time;
H_r – order r hyperexponentially distributed interarrival or service time;
D— deterministic inter-arrival or service time;
G or GI – i.i.d. random interarrival or service time with any general distribution.

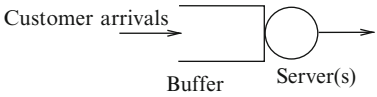


Fig. 6.1 Common representation of queueing systems

The symbols d, e, and x are not indicated if they take their default values: $d = \infty$ infinite system capacity, $e = \infty$ infinite customer population, and $x = FCFS$ service in arrival order.

Additionally, queueing systems have the following properties. If there is an idle server when a customer arrives at the system, then the service to the customer starts immediately. It is assumed that $c \leq d$. If $c = d$ (the system capacity is identical to the number of servers), then there is no buffer position available for customers that arrive at the system when all servers are busy. In this case, the arriving customer leaves the system without service. These systems are also referred to as loss systems. If $c < d$, then customers arrive at the system when all servers are busy and there is still an available buffer position; customers are not lost but wait until a server becomes available to start their service. The time period from the arrival of such customers to the beginning of the service is referred to as the waiting time. The elements of queueing systems are commonly depicted as in Fig. 6.1.

6.3 Queueing System Performance Parameters

The optimal operation of queueing systems can be analyzed through several performance parameters, the most important of which follow.

1. *Customer loss probability* (of queuing systems with finite capacity, $d < \infty$): Let $0 \leq t_1 \leq t_2 \leq \ldots$ be the arrival times of the first, second, ... customers, and let m_n be the number of the first n customers that are lost. If $\lim_{n\to\infty} m_n/n = q$ in a stochastic sense, then q is referred to as the **loss probability**. In finite-capacity systems, it is also important to check if the $\lim_{n\to\infty} m_n/n$ limit exists at all.
2. *Waiting time distribution*: Let W_n, $n \geq 1$, be the waiting time of the nth customer; the number of the first n customers whose waiting time is less than x is

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_{\{W_i \leq x\}}, \quad x > 0.$$

   $F_n(x)$ is the empirical distribution function of the waiting time based on the first n customers. If $\lim_{n\to\infty} F_n(x) = F(x)$ in a stochastic sense for $\forall x > 0$, then $F(x)$ is the CDF of the **waiting time distribution**.
3. *Mean waiting time*: If $\lim_{n\to\infty} \frac{1}{n}(W_1 + \ldots + W_n) = W^{(1)}$ in a stochastic sense, then $W^{(1)}$ is the mean waiting time. The higher moments of the waiting time are defined by stochastic convergence in a similar way:

$$W^{(k)} = \lim_{n\to\infty} \frac{1}{n}(W_1^k + \ldots + W_n^k), \quad k \geq 1.$$

In a wide range of practical cases,

$$W^{(k)} = \int_0^\infty x^k dF(x), \quad k \geq 1,$$

holds.

4. *Distribution of a server's busy period*: Consider a server of a queueing system. Let $[a_n, b_n),\ n \geq 1$, denote the consecutive intervals during which the server is busy (serving a customer). a_n and b_n are such that $a_n < b_n < a_{n+1},\ n \geq 1$, and the server is idle (serving no customers) during the intervals $[b_n, a_{n+1}),\ n \geq 1$. In this case $[a_n, b_n)$ denotes the n**th busy period** of the given server. If

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}_{\{b_i - a_i \leq x\}} = G(x), \quad \forall x > 0,$$

in a stochastic sense, then $G(x)$ is the **distribution of the busy period** of the given server. The distribution of the idle period can be defined in an analogous way.

5. *Queue length distribution*: Let $L(t),\ t \geq 0$, be the number of customers in the system (including those in the servers and those waiting in the buffer) at time t and $\bar{L}_k(t)$ be the portion of time in $(0, t)$ during which there were k customers in the system:

$$\bar{L}_k(t) = \frac{1}{t} \int_0^t \mathcal{I}_{\{L(s)=k\}} \mathrm{d}s.$$

If

$$p_k = \lim_{t\to\infty} \bar{L}_k(t),\ k \geq 0,$$

exists in a stochastic sense, then $(p_k,\ k \geq 0)$ defines the **queue length distribution**.

As was done previously, one can define the moments of the busy and idle periods and the queue length.

Comment 6.1. *If the state (e.g., number of customers in the system) of the queueing system can be described by the discrete-state $\mathcal{X} \subseteq \mathbb{N}^+$ homogeneous ergodic Markov chain $X(t),\ t \geq 0$, and $f(i),\ i \in \mathcal{X}$, is an arbitrary bounded function, then*

$$\bar{f} = \lim_{t\to\infty} \frac{1}{t} \int_0^t f(\nu(s))\mathrm{d}s$$

and

$$\bar{f} = \sum_{i \in \mathcal{X}} f(i)\pi_i,$$

where $\{\pi_i,\ i \in \mathcal{X}\}$*, denotes the stationary distribution of the Markov chain.*

6.4 Little's Law

Little's law describes the relation of various performance parameters of queueing systems. We need the following notations to present it.

$N(t),\quad t \geq 0$: **number of arrivals** (number of customers arriving at *system* in $[0, t)$);

$M(t),\quad t \geq 0$: **number of departures** (number of customers leaving *system* in $[0, t)$);

$L(t) = N(t) - M(t),\quad t \geq 0$: **queue length** (number of customers in *system* at time t).

Note that *system* can be replaced by any part of the queueing system in the foregoing definitions, e.g., by the buffer or a subset of servers. Based on these quantities we can compute the following ones.

$\bar{L}(t) = \dfrac{1}{t}\displaystyle\int_0^t L(s)\mathrm{d}s,\quad t > 0$: **mean number of customers** in system in $[0, t)$;

$\lambda(t) = \dfrac{N(t)}{t},\quad t > 0$: **arrival intensity** in $[0, t)$;

$\tau_n(t),\quad t > 0$: time nth customer spends in system in $[0, t)$;

$\tau(t) = \displaystyle\sum_{n=1}^{N(t)} \tau_n(t),\quad t > 0$: aggregate time customers that arrived before time t spend in system in $[0, t)$.

$T(t) = \dfrac{\tau(t)}{N(t)},\quad t > 0$: the average time a customer spends in the system in $[0, t)$ considering the customers that arrived before time t.

Comment 6.2. *These quantities have the following relations:*

$$\tau(t) = \int_0^t [N(s) - M(s)]\mathrm{d}s = \int_0^t L(s)\mathrm{d}s,\quad t \geq 0,$$

$$\bar{L}(t) = \frac{\tau(t)}{t},\quad t > 0.$$

In the analysis of queueing systems we are usually interested in the long-term or stationary behavior. Little's law defines the relation of the limiting values of these quantities.

Theorem 6.3 (*Little*). *If*

$$\lim_{t\to\infty} \lambda(t) = \lim_{t\to\infty} \frac{N(t)}{t} = \lambda, \quad \lambda > 0,$$

$$\lim_{t\to\infty} \bar{L}(t) = \lim_{t\to\infty} \frac{1}{t} \int_0^t L(s)\mathrm{d}s = L$$

in a stochastic sense, then $T(t)$ *converges to* T *in a stochastic sense as* $t \to \infty$ *and*

$$L = \lambda T.$$

Proof. The stochastic convergence of $T(t)$ is obtained from

$$\lim_{t\to\infty} T(t) = \lim_{t\to\infty} \frac{\tau(t)}{N(t)} = \lim_{t\to\infty} \frac{\tau(t)/t}{N(t)/t} = \lim_{t\to\infty} \frac{\bar{L}(t)}{\lambda(t)} = T,$$

and it also results in the main statement of the theorem. □

Comment 6.4. *The* $L = \lambda T$ *relation was known as an experimental law for a long time. It was first proved by J. Little [66] in 1961 and was later commonly referred to as Little's law [95]. In words, Little's law can be expressed as*

***mean number of customers in a* system**

***= arrival intensity * mean time a customer spends in the* system**

and is independent of the definition of the system (as discussed previously), the arrival and service time distributions, number of servers, and buffer size.

Depending on the definition of *system*, we obtain the following versions of Little's law:

(a) If the *system* is the buffer only, then

$$L_\mathrm{w} = \lambda W,$$

where L_w is the mean number of waiting customers and W is the mean waiting time.

(b) If the *system* is the set of all servers, then

$$L_\mathrm{s} = \lambda \bar{Y},$$

where L_s is the mean number of busy servers and $\bar{Y}$ is the mean service time.

6.5 Exercises

Exercise 6.1. Interpret the following Kendall's notations:

- $M/M/1/\infty/\infty - FIFO,\ M/M/1$
- $M/M/2//4$
- $M/M/1//m - PS$
- $M/M/m - LIFO$

Exercise 6.2. In a single-server, infinite-buffer queueing model, the arrival rate is λ and the service time is exponentially distributed with the parameter μ.

- Define Little's law for the entire queueing system, for the buffer, and for the server.
- Which one of these expressions defines the server utilization?
- What is the utilization?

Exercise 6.3. Which of the following queueing systems are lossless?

- $M/M/1$
- $M/M/2/5/4$
- $M/M/1/2 - PS$
- $M/M/m/m$
- $M/M/m$

Exercise 6.4. Which of the following queueing systems provide immediate service for customers?

- $M/M/1$
- $M/M/4/5/3$
- $M/M/1/2 - PS$
- $M/M/m/m$
- $M/M/m$

Chapter 7
Markovian Queueing Systems

Queueing systems whose underlying stochastic process is a continuous-time Markov chain (CTMCs) are the simplest and most often used class of queueing systems. The analysis of these systems is based on the essential results available for the analysis of CTMCs. As a consequence, several interesting properties of these queueing systems can be described by simple closed-form analytical expressions both in transient (as a function of time and initial state) and in steady state.

The most often studied property of basic queueing systems is the queue length process, $N(t), t \geq 0$, which represents the number of customers in the system at time t. If customers belong to K different customer classes, then the vector-valued function $N(t) = (N_1(t), \ldots, N_K(t))$, $t \geq 0$, describes the queue length process. In this case the ith component of $N(t)$, $N_i(t)$, is the number of class i customers in the system.

The queue length process of basic queueing systems with a single class of customers is the birth-death process, which is a special CTMC. We will utilize the previously introduced results of CTMCs and birth-death processes for the analysis of queueing systems.

7.1 $M/M/1$ Queue

The most basic queueing system is the $M/M/1$ queue, which is composed of a single server and an infinite buffer (see Fig. 7.1). Identical customers arrive according to a time-homogeneous Poisson process at rate λ, and the service time of a customer is exponentially distributed at rate μ. The customers are served in the order of their arrival (FCFS). The server is always busy as there is at least one customer in the system. This last property is referred to as a *work-conserving* property, which we commonly assume in the sequel unless otherwise stated.

Let $N(t)$ be the number of customers in a system (either being served or waiting in the buffer) at time t. Due to the memoryless property of the arrival and the service process, $N(t)$ is a CTMC (Fig. 7.2). Its (nonvanishing) state-transition probabilities

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, 199
DOI 10.1007/978-1-4614-5317-8_7, © Springer Science+Business Media, LLC 2013

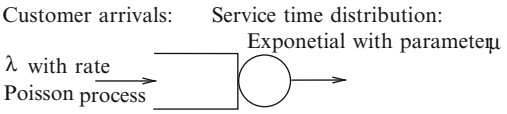


Fig. 7.1 $M/M/1(/\infty)$ queueing system

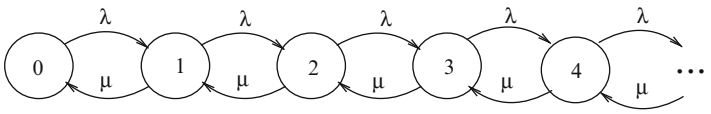


Fig. 7.2 Markov chain of number of customers in $M/M/1$ queue

are

$$\begin{aligned} p_{i,i+1}(\Delta) &= \lambda\Delta + o(\Delta), \quad i = 0, 1, \ldots, \\ p_{i,i-1}(\Delta) &= \mu\Delta + o(\Delta), \quad i = 1, 2, \ldots, \\ p_{ii}(\Delta) &= 1 - (\lambda + \mu)\Delta + o(\Delta), \quad i = 1, 2, \ldots, \\ p_{0,0}(\Delta) &= 1 - \lambda\Delta + o(\Delta), \end{aligned}$$

where $p_{i,j}(t) = \mathbf{P}(N(t) = j \mid N(0) = i)$. That is, $N(t)$ is an infinite state birth-death process with $\lambda_i = \lambda \ \ (i = 0, 1, \ldots)$ and $\mu_i = \mu \ \ (i = 1, 2, \ldots)$. The Markov chain is irreducible, and from its stationary equations we have

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i, \quad i = 0, 1, \ldots.$$

According to Eq. (3.19) this Markov chain is stable if

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

that is,

$$\lambda < \mu.$$

The intuitive explanation of this relation is straightforward. It means that the queue is stable if the mean service time $(1/\mu)$ is less than the mean interarrival time $(1/\lambda)$. Introducing $\rho = \frac{\lambda}{\mu}$ from Eqs. (3.21) and (3.20) we have

$$p_i = \lim_{t\to\infty} \mathbf{P}(N(t) = i) = p_0\rho^i \quad (i = 0, 1, \ldots),$$

where

$$p_0 = \frac{1}{\sum_{j=0}^{\infty} \rho^j} = \frac{1}{\frac{1}{1-\rho}} = 1 - \rho$$

and

$$p_i = (1-\rho)\rho^i, \quad i = 0, 1, \ldots$$

Consequently, the stationary number of customers in the system is geometrically distributed on $\{0, 1, \ldots\}$ with parameter $(1-\rho)$. The mean of this distribution is

$$\bar{N} = \sum_{i=0}^{\infty} i\ p_i = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}. \tag{7.1}$$

Now we compute the mean time a customer spends in the system in stationary state. Let us consider that the customer arrives at time t. The number of customers in the system at this time instant is $N(t)$. According to the FCFS service order, the new customer must wait while all of the customers present in the system at time t are served. Due to the memoryless property of the exponentially distributed service time, the remaining service for the customer being served at time t (if any) is also exponentially distributed with the same parameter. Furthermore, the service times of the $N(t) - 1$ customers waiting in the buffer at time t [if $N(t) \geq 1$] and the service time of the newly arrived customer are also exponentially distributed with parameter μ. Summing up all these, the total time the customer spends in the system is

$$D(t) = \sum_{i=1}^{N(t)} Y_i + Y, \tag{7.2}$$

where $Y_1, \ldots, Y_{N(t)}$ and Y are i.i.d. exponentially distributed random variables with parameter μ. In stationary state, neither the distribution of $N(t)$ nor the distribution of $D(t)$ depends on t, that is, $\mathbf{E}\left(D(t)\right) = \bar{D}$. $\bar{D}$ can be computed using the following lemma (see also Exercise 2.1).

Lemma 7.1 (*Wald's lemma). If N is a nonnegative-integer-valued random variable and $\{Y_i\}$ are nonnegative, i.i.d. random variables, independent of N, then*

$$\boldsymbol{E}\left(\sum_{i=1}^{N} Y_i\right) = \boldsymbol{E}\left(Y_1\right)\ \boldsymbol{E}\left(N\right).$$

Proof.

$$\begin{aligned}
\mathbf{E}\left(\sum_{i=1}^{N} Y_i\right) &= \mathbf{E}\left(\sum_{i=1}^{\infty} Y_i\, \mathcal{I}_{\{i \leq N\}}\right) = \sum_{i=1}^{\infty} \mathbf{E}\left(Y_i\, \mathcal{I}_{\{i \leq N\}}\right) \\
&= \sum_{i=1}^{\infty} \mathbf{E}\left(Y_i\right) \mathbf{E}\left(\mathcal{I}_{\{i \leq N\}}\right) = \mathbf{E}\left(Y_1\right) \sum_{i=1}^{\infty} \mathbf{P}\left(N \geq i\right) \\
&= \mathbf{E}\left(Y_1\right) \sum_{k=1}^{\infty} k \cdot \mathbf{P}\left(N = k\right) = \mathbf{E}\left(Y_1\right) \mathbf{E}\left(N\right).
\end{aligned}$$

□

Based on the Wald lemma and Eqs. (7.1) and (7.2) we have

$$\bar{D} = \mathbf{E}(Y_1) \cdot \bar{N} + \mathbf{E}(Y) = \frac{1}{\mu} \cdot \frac{\lambda}{\mu - \lambda} + \frac{1}{\mu} = \frac{1}{\mu - \lambda}.$$

We can also evaluate the stationary system time (time spent in the system by a customer) distribution $D(t) \equiv D$ because

$$\mathbf{P}(D \le s) = \sum_{i=0}^{\infty} \mathbf{P}(D \le s | N(t) = i) \cdot \mathbf{P}(N(t) = i),$$

where $\mathbf{P}(D \le s | N(t) = i)$ is the distribution of the sum of $(i + 1)$ i.i.d. exponentially distributed random variables with parameter μ according to Eq. (7.2). Thus,

$$\begin{aligned}
\mathbf{P}(D < s) &= \sum_{i=0}^{\infty} \left(\int_0^s \mu \frac{(\mu u)^i}{i!} \mathrm{e}^{-\mu u} \mathrm{d}u \right) (1 - \rho) \rho^i \\
&= \int_0^s \left(\sum_{i=0}^{\infty} \mu \frac{\mu^i u^i}{i!} \mathrm{e}^{-\mu u} \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^i \right) \mathrm{d}u \\
&= \mu \left(1 - \frac{\lambda}{\mu}\right) \int_0^s \left(\mathrm{e}^{-\mu u} \cdot \sum_{i=0}^{\infty} \frac{(\lambda u)^i}{i!} \right) \mathrm{d}u \\
&= \mu \left(1 - \frac{\lambda}{\mu}\right) \int_0^s \mathrm{e}^{-\mu u} \cdot \mathrm{e}^{\lambda u} \mathrm{d}u \\
&= \mu \left(1 - \frac{\lambda}{\mu}\right) \int_0^s \mathrm{e}^{-(\mu - \lambda) u} \mathrm{d}u \\
&= \mu \cdot \frac{\mu - \lambda}{\mu} \cdot \frac{1}{\mu - \lambda} \left(1 - \mathrm{e}^{-(\mu - \lambda) s}\right) \\
&= 1 - \mathrm{e}^{-(\mu - \lambda) s}, \quad s \ge 0,
\end{aligned}$$

where we used that the sum of $(i + 1)$ i.i.d. exponentially distributed random variables with parameter μ is Gamma (or Erlang) distributed with parameters μ and $i + 1$. We obtained that the system time is exponentially distributed with parameter $(\mu - \lambda)$ and its mean is

$$\bar{D} = \frac{1}{\mu - \lambda},$$

as we saw previously.

The departure process of a stationary $M/M/1$ queue (the point process of the consecutive departure instants) is a Poisson process with parameter λ. This property

is referred to as Burke's theorem and plays an important role in the analysis of queueing networks, as discussed in Sect. 10.1.

In single-server systems, the probability of finding the server busy is referred to as utilization.

$$\mathbf{P}\,(\text{the server is busy}) = \sum_{k=1}^{\infty} p_k \;= 1 - p_0 \;= 1 - (1-\rho) \;=\; \rho.$$

Let X_s be the number of customers being served in stationary state. In this case, $\mathbf{E}\,(X_\mathrm{s}) = 0 \cdot p_0 + 1 \cdot (1 - p_0)$, whence

$$\mathbf{E}\,(X_\mathrm{s}) = \rho. \tag{7.3}$$

According to Little's law, if $\rho < 1$ (i.e., the system is stable), then

$$\mathbf{E}\,(X_\mathrm{s}) = \bar{\lambda}\bar{x} = \frac{\lambda}{\mu} = \rho \tag{7.4}$$

because $\bar{\lambda} = \lambda$ and $\bar{x} = \dfrac{1}{\mu}$.

The mean number of customers in the system is

$$\begin{aligned}\mathbf{E}\,(X) &= \sum_{k=0}^{\infty} k\; p_k \;= \sum_{k=0}^{\infty} k\;(1-\rho)\rho^k \;= \rho(1-\rho)\sum_{k=0}^{\infty} k\;\rho^{k-1} \\ &= \rho(1-\rho)\sum_{k=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}\rho}\rho^k \;= \rho(1-\rho)\,\frac{\mathrm{d}}{\mathrm{d}\rho}\sum_{k=0}^{\infty}\rho^k \;= \rho(1-\rho)\,\frac{1}{(1-\rho)^2},\end{aligned}$$

whence

$$\mathbf{E}\,(X) = \frac{\rho}{1-\rho}. \tag{7.5}$$

The mean number of waiting customers in the system is

$$\begin{aligned}\mathbf{E}\,(X_\mathrm{w}) &= \sum_{k=1}^{\infty} (k-1)\; p_k = \sum_{k=1}^{\infty} k\; p_k - \sum_{k=1}^{\infty} p_k \\ &= \sum_{k=1}^{\infty} k\;(1-\rho)\rho^k - (1-p_0) = \frac{\rho}{1-\rho} - \rho,\end{aligned}$$

whence

$$\mathbf{E}\,(X_\mathrm{w}) = \frac{\rho^2}{1-\rho}. \tag{7.6}$$

By definition,

$$X = X_{\mathrm{w}} + X_{\mathrm{s}}, \tag{7.7}$$

and thus

$$\mathbf{E}(X) = \mathbf{E}(X_{\mathrm{w}}) + \mathbf{E}(X_{\mathrm{s}}), \tag{7.8}$$

$$\mathbf{E}(X_{\mathrm{w}}) = \mathbf{E}(X) - \mathbf{E}(X_{\mathrm{s}}) = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}. \tag{7.9}$$

A customer's system time, waiting time, and service time (T, W, and x, respectively) fulfill

$$T = W + x \tag{7.10}$$

and

$$\bar{T} = \bar{W} + \bar{x}. \tag{7.11}$$

According to Little's law,

$$\bar{T} = \frac{\mathbf{E}(X)}{\lambda} = \frac{1}{\mu(1-\rho)}, \tag{7.12}$$

$$\bar{W} = \frac{\mathbf{E}(X_{\mathrm{w}})}{\lambda} = \frac{\rho}{\mu(1-\rho)}, \tag{7.13}$$

$$\bar{x} = \frac{1}{\mu}, \tag{7.14}$$

which confirms Eq. (7.11).

$\bar{T}$ can also be computed as

$$\bar{T} = \frac{1}{\mu(1-\rho)} = \frac{1-\rho+\rho}{\mu(1-\rho)} = \frac{1}{\mu} + \frac{\rho}{\mu(1-\rho)} = \frac{1}{\mu} + \frac{1}{\mu}\frac{\rho}{1-\rho} = \frac{1}{\mu} + \sum_{k=0}^{\infty} k\,\frac{1}{\mu}\,p_k.$$

The probability that there are at least k customers in the system is

$$\mathbf{P}(X \geq k) = \sum_{i=k}^{\infty} (1-\rho)\rho^i = (1-\rho)\rho^k \sum_{j=0}^{\infty} \rho^j = \rho^k. \tag{7.15}$$

Example 7.2. Let us consider a data packet transmission unit that receives data packets from a set of terminals and transmits them to a destination unit through a transmission line. The packets arrive according to a Poisson arrival process. On average, one packet arrives every 4 ms. The packet transmission time is exponentially distributed. The mean packet transmission time is 3 ms.

What is the mean number of packets in the transmission unit if it has an infinite buffer?

$$\rho = \frac{1}{4} \cdot 3 = \frac{3}{4},$$

$$\mathbf{E}(X) = \frac{\rho}{1-\rho} = 3.$$

What is the mean system time of a customer?

$$\mathbf{E}(T) = \frac{\mathbf{E}(X)}{\lambda} = \frac{3}{1/4\ 1/\text{ms}} = 12\,\text{ms}.$$

By how much must the arrival rate increase for the mean system time to double?

$$\mathbf{E}(T') = 24\,\text{ms} = \frac{1/\mu}{1-\rho'} = \frac{3\,\text{ms}}{1-\rho'},$$

$$\rho' = 1 - \frac{1}{8} = 7/8,$$

whence

$$\lambda' = \rho'\mu = \frac{7}{8 \cdot 3} = \frac{7}{24}.$$

This means that a small (17%) increase in the arrival rate doubles the mean system time.

Example 7.3. Customers arrive at an infinite buffer queueing system according to a Poisson process at rate $K\lambda$, and the service time is exponentially distributed. The mean service time of a high-capacity server is $1/(K\mu)$, and the mean service time of a low-capacity server is $1/\mu$. Compare the performance of the single queue using one high-capacity server with the performance of K parallel queues having low-capacity servers. In the latter case, customers arrive at each queue according to a Poisson process at rate λ (cf. the decomposition of a Poisson process in Sect. 2.7.3).

Performance of a single high-capacity queue:

$$\rho = \frac{K\lambda}{K\mu} = \frac{\lambda}{\mu},$$

$$\mathbf{E}(T) = \frac{\mathbf{E}(X)}{1-\rho} = \frac{1}{K\mu(1-\rho)};$$

Performance of the K low-capacity queues:

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

per server, and

$$\mathbf{E}(T') = \frac{\mathbf{E}(X)}{1-\rho} = \frac{1}{\mu(1-\rho)} = K \cdot \mathbf{E}(T).$$

Consequently, the system time is K times longer in the latter case.

The result demonstrates that aggregating the resources and demands in service systems increases system performance.

7.2 Transient Behavior of an $M/M/1$ Queueing System

Let $A(x) = 1 - e^{-\lambda x}$ and $B(x) = 1 - e^{-\mu x}$ be the CDF of the interarrival and the service time distribution, and $L(t)$, $t \geq 0$ be the number of customers in the system at time t. From the fact that $L(t)$ is a birth-death process (Sect. 3.4) we will derive the following characteristics (Fig. 7.3):

(A) The parameters of $\{L(t),\ t \geq 0\}$.
(B) The distribution of $L(t)$ at an arbitrary $t \geq 0$ instant (using point 2 of Theorem 3.68).
(C) The distribution of the length of the busy period of the server (based on Theorem 3.70).
(D) The distribution of the stationary virtual waiting time (the time required to serve customers in the system at an arbitrary time instant).

(A) Parameters of birth-death process:

Let t be an instant when the system becomes idle. For this t we have

$$L(t) = 0.$$

The next state change of the system happens at the arrival of the next customer after t. The time till this state change is the time spent in state 0. This idle time is exponentially distributed with parameter $\lambda > 0$. Applying the notations of Sect. 3.4 we have

$$a_0 = \lambda, \quad p_0 = 1.$$

Due to the memoryless property of the exponential distribution, starting from t' such that the system was idle for a period of time does not modify the distribution of the remaining time till the next customer arrival. The same holds for all other $k > 0$ states of the system.

Assuming that there are k, $k \geq 1$, customers in the system at time t; the next state change occurs when a new customer arrives or when a customer's service is completed. The probability that these two events will occur at the same time is 0.

We need the following notations:

ξ_t: the time to arrival of the next customer after time t,
η_t: the remaining service time of the customer being served at time t.

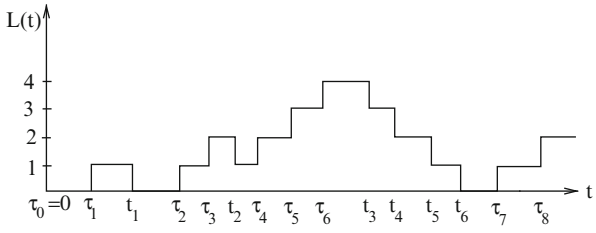


Fig. 7.3 Number of customers in the $M/M/1(/\infty)$ queueing system

ξ_t and η_t are independent, and (due to the memoryless property of the exponential distribution) their distributions are $A(x)$ and $B(x)$, respectively.

Let $\zeta_t = \min(\xi_t, \eta_t)$ be the time till the next state change at time t. The distribution of ζ_t is

$$\begin{aligned}\mathbf{P}(\zeta_t \le x) &= 1 - \mathbf{P}(\zeta_t > x) = 1 - \mathbf{P}(\xi_t > x, \eta_t > x)\\ &= 1 - \mathbf{P}(\xi_t > x)\,\mathbf{P}(\eta_t > x) = 1 - \mathrm{e}^{-(\lambda+\mu)x},\end{aligned}$$

from which α_k (the parameter of the sojourn time in state k) is $\alpha_k = \lambda + \mu, \quad k \ge 1$.

After ζ_t the system moves to state $(k+1)$ or $(k-1)$, depending on the relation of ξ_t and η_t. If $\xi_t < \eta_t$, then it moves to state $(k+1)$; if $\xi_t > \eta_t$, then it moves to state $(k-1)$ $[\mathbf{P}(\xi_t = \eta_t) = 0]$.

$$\begin{aligned}\mathbf{P}(\zeta_t \le x,\, \xi_t \le \eta_t) &= \mathbf{P}(\xi_t \le x,\, \xi_t \le \eta_t)\\ &= \int_0^x \mathbf{P}(\xi_t \le x, \xi_t \le \eta_t | \xi_t = u)\,\mathrm{d}\mathbf{P}(\xi_t \le u)\\ &= \int_0^x \mathbf{P}(u \le \eta_t)\,\mathrm{d}(1 - \mathrm{e}^{-\lambda u}) = \int_0^x \mathrm{e}^{-\mu u}\mathrm{d}(1 - \mathrm{e}^{-\lambda u})\\ &= \frac{\lambda}{\lambda+\mu}\left(1 - \mathrm{e}^{-(\lambda+\mu)x}\right),\end{aligned}$$

and similarly

$$\mathbf{P}(\zeta_t \le x,\, \xi_t > \eta_t) = \mathbf{P}(\zeta_t \le x) - \mathbf{P}(\zeta_t \le x,\, \xi_t \le \eta_t) = \frac{\mu}{\lambda+\mu}\left(1 - \mathrm{e}^{-(\lambda+\mu)x}\right).$$

This means that, independently of state k, the parameter of the exponentially distributed time spent in state k is $\alpha_k = (\lambda + \mu)$, and the probabilities of moving to $(k+1)$ and $(k-1)$ are $p_k = \frac{\lambda}{\lambda+\mu}$ and $q_k = \frac{\mu}{\lambda+\mu}$ $(k \ge 1)$, respectively.

Consequently, $L(t)$ is a birth-death process with parameters

$$\begin{aligned}&a_k = \lambda, \quad k \ge 0, \quad b_k = \mu, \quad k \ge 1;\\ &\alpha_k = \lambda + \mu, \quad p_0 = 1, \quad p_k = \frac{\lambda}{\lambda+\mu}, \quad q_k = \frac{\mu}{\lambda+\mu}, \quad k \ge 1.\end{aligned}$$

(B) Distribution of $L(t)$:

We assume that the system is idle at time 0 $[L(0) = 0]$ with probability 1 $(\varphi_0 = 1, \varphi_k = 0, \quad k \ge 1)$ and compute the distribution of $L(t), \ t \ge 0$. More precisely, we evaluate

$$P_k(t) = \mathbf{P}\left(L(t) = k\right), \quad p_k^*(s) = \int_0^\infty \mathrm{e}^{-st} P_k(t)\mathrm{d}t, \quad \mathrm{Re}\, s > 0.$$

The $p_k^*(s)$ functions are given by point 2 of Theorem 3.68:

$$sp_0^*(s) - 1 = -\lambda p_0^*(s) + \mu p_1^*(s), \tag{7.16}$$

$$sp_k^*(s) = \lambda p_{k-1}^*(s) - (\lambda + \mu) p_k^*(s) + \mu p_{k+1}^*(s), \quad k \geq 1. \tag{7.17}$$

Following the approach proposed in [69] we define the probability generating function

$$p^*(z, s) = \sum_{k=0}^{\infty} p_k^*(s)\, z^k.$$

Multiplying both sides of Eq. (7.17) by z^k $(0 < |z| \leq 1, \quad k \geq 1)$ and summing up the terms $k = 1, 2, \ldots$ we obtain

$$\begin{aligned} sp^*(z, s) - sp_0^*(s) = \lambda z p^*(z, s) - (\lambda + \mu)[p^*(z, s) - p_0^*(s)] \\ + \frac{1}{z}\mu[p^*(z, s) - p_0^*(s) - z p_1^*(s)]. \end{aligned}$$

Further adding Eq. (7.16) and rearranging the terms we obtain

$$p^*(z, s)\left[s - \lambda z + (\lambda + \mu) - \frac{\mu}{z}\right] = 1 + \mu p_0^*(s) - \frac{\mu}{z} p_0^*(s),$$

whence

$$p^*(z, s) = \frac{z - \mu(1 - z) p_0^*(s)}{s\, z - (1 - z)(\mu - \lambda z)}. \tag{7.18}$$

The function $p^*(z, s)$ is the Laplace transform of a generator function. It is analytic and bounded for $|z| \leq 1$ and fixed $\mathrm{Re}\, s > 0$. Consequently, on the right-hand side of Eq. (7.18) the numerator must have a root at the root of the denominator. The denominator has a root at

$$z = \gamma_1(s) = \frac{\lambda + \mu + s - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}}{2\lambda}.$$

Using that the numerator also has a root at $z = \gamma_1(s)$ we obtain that $|\gamma_1(s)| < 1$ if $\mathrm{Re}\, s > 0$. It is easy to see for real $s > 0$ because

$$\lambda + \mu + s - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu} < 2\lambda$$

holds for $\mu - \lambda + s < 0$ since in this case $\mu - \lambda + s < \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}$. If $\mu - \lambda + s \geq 0$, then the squares of the two sides of the equation remain equal and we obtain a simple identity.

The numerator of $p^*(z, s)$ in Eq. (7.18) must be zero at $z = \gamma_1(s)$, that is,

$$\gamma_1(s) - \mu(1 - \gamma_1(s))p_0^*(s) = 0, \quad \text{that is,} \quad p_0^*(s) = \frac{\gamma_1(s)}{\mu(1 - \gamma_1(s))}.$$

Thus

$$p^*(z, s) = \frac{z - (1 - z)\frac{\gamma_1(s)}{1 - \gamma_1(s)}}{s\,z - (1 - z)(\mu - \lambda z)} = \frac{z - \gamma_1(s)}{(1 - \gamma_1(s))(s\,z - (1 - z)(\mu - \lambda z))}.$$

Introducing

$$\gamma_2(s) = \frac{\lambda + \mu + s + \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}}{2\lambda}$$

and using

$$s\,z - (1 - z)(\mu - \lambda z) = -\lambda(z - \gamma_1(s))(z - \gamma_2(s))$$

we modify the expression in the following way:

$$p^*(z, s) = \frac{1}{\lambda(1 - \gamma_1(s))(\gamma_2(s) - z)}.$$

It can be seen that $|\gamma_2(s)| > 1$ [$\gamma_2(s) \neq z$, $|z| \leq 1$ since $p^*(z, s)$ is bounded there] and

$$\gamma_1(s)\gamma_2(s) = \frac{\mu}{\lambda}.$$

The series expansion of the fraction $\frac{1}{\gamma_2(s) - z}$, according to z, gives

$$p^*(z, s) = [\lambda(1 - \gamma_1(s))\gamma_2(s)]^{-1} \sum_{k=0}^{\infty} \left[\frac{z}{\gamma_2(s)}\right]^k.$$

Comparing the coefficients of the z^k terms and using the series expansion of the fraction $\frac{1}{1 - \gamma_1(s)}$ we have

$$p_k^*(s) = [\lambda(1 - \gamma_1(s))(\gamma_2(s))^{k+1}]^{-1} = \frac{1}{\lambda(\gamma_2(s))^{k+1}} \sum_{j=0}^{\infty} (\gamma_1(s))^j$$

$$= \frac{1}{\lambda\,[\gamma_1(s)\gamma_2(s)]^{k+1}} \sum_{j=k+1}^{\infty} (\gamma_1(s))^j = \frac{1}{\lambda}\left(\frac{\lambda}{\mu}\right)^{k+1} \sum_{j=k+1}^{\infty} (\gamma_1(s))^j$$

$$= \frac{1}{\lambda}\left(\frac{\lambda}{\mu}\right)^{k+1} \sum_{j=k+1}^{\infty} \left(\frac{\mu}{\lambda}\right)^j (\gamma_2(s))^{-j}\,.$$

The last expression allows the explicit description of $P_k(t)$. Let $I_m(z)$ be the modified first-order Bessel function, i.e.,

$$I_m(z) = \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(m+k+1)} \left(\frac{z}{2}\right)^{m+2k}.$$

Since the Laplace transform of

$$\int_0^{\infty} \mathrm{e}^{-sx} x^{-1} I_m(cx)\mathrm{d}x$$

is (see [76])

$$\frac{c^m}{m}(s + \sqrt{s^2 - c^2})^{-m},$$

the inverse Laplace transform of

$$\left(\frac{s + \sqrt{s^2 - 4\lambda\mu}}{2\lambda}\right)^{-m}$$

is

$$m\left(\frac{\lambda}{\mu}\right)^{m/2} t^{-1} I_m(2\sqrt{\lambda\mu}t).$$

Additionally, using $\mathrm{e}^{-st}\mathrm{e}^{-(\lambda+\mu)t} = \mathrm{e}^{-(s+\lambda+\mu)t}$ the inverse Laplace transform of

$$\gamma_2(s)^{-m} = \left(\frac{s + \lambda + \mu + \sqrt{(s+\lambda+\mu)^2 - 4\lambda\mu}}{2\lambda}\right)^{-m}$$

is

$$\mathrm{e}^{-(\lambda+\mu)t} m\left(\frac{\lambda}{\mu}\right)^{m/2} t^{-1} I_m(2\sqrt{\lambda\mu}t),$$

and finally we obtain

$$P_k(t) = \frac{1}{\lambda}\left(\frac{\lambda}{\mu}\right)^{k+1} \mathrm{e}^{-(\lambda+\mu)t} \sum_{j=k+1}^{\infty} jt^{-1} \left(\frac{\mu}{\lambda}\right)^{j/2} I_j(2\sqrt{\lambda\mu}t).$$

(C) Distribution of busy intervals:

Let Π_k random variables be the length of the busy period starting from state k, and let

$$\Pi_k(t) = \mathbf{P}(\Pi_k \le t), \quad \text{and } \pi_k^*(s) = \mathbf{E}\left(\mathrm{e}^{-s\Pi_k}\right)$$

be its CDF and Laplace transform, respectively. To compute $\Pi_k(t)$ and $\pi_k^*(s)$, we assume that state 0 is an absorbing state and put

$$\varphi_k = 1, \quad a_0 = 0, \quad p_0 = 0, \quad a_n = \lambda, \quad \alpha_n = \lambda + \mu, \quad p_n = \frac{\lambda}{\lambda + \mu}, \quad n \ge 1.$$

We note (Remark 3.71) that

$$P_0(t) = \mathbf{P}(L(t) = 0) = \mathbf{P}(\Pi_k \le t) = \Pi_k(t), \quad P_0'(t) = \Pi_k'(t).$$

According to Theorem 3.70 for $p_n^*(s), \quad n \ge 1$, we have

$$sp_0^*(s) = \mu p_1^*(s), \tag{7.19}$$

$$sp_1^*(s) - \delta_{1,k} = -(\lambda + \mu)p_1^*(s) + \mu p_2^*(s), \tag{7.20}$$

$$sp_n^*(s) - \delta_{n,k} = \lambda p_{n-1}^*(s) - (\lambda + \mu)p_n^*(s) + \mu p_{n+1}^*(s), \quad n \ge 2. \tag{7.21}$$

Furthermore, according to point 1 of Theorem 3.70 $\Pi_k'(t) = P_0'(t) = \mu P_1(t)$, and consequently

$$\pi_k^*(s) = \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}\Pi_k(x) = \int_0^\infty \mathrm{e}^{-sx}\Pi_k'(x)\mathrm{d}x = \int_0^\infty \mathrm{e}^{-sx}bP_1(x)\mathrm{d}x = bp_1^*(s).$$

Multiplying Eq. (7.21) by z^n, summing it up for $n \ge 2$, and adding Eq. (7.20) z times, we obtain

$$s\,p^*(z,s) - z^k = \lambda\, z\, p^*(z,s) - (\lambda + \mu)\, p^*(z,s) + \frac{\mu}{z}\, p^*(z,s) - \mu\, p_1^*(s),$$

where $p^*(z,s) = \sum_{n=1}^{\infty} p_n^*(s)\, z^n$ and $z^n\delta_{n,k} = z^k I(n = k)$. Further rearranging the expression gives

$$[s\,z - (1-z)(\mu - \lambda z)]\,\frac{p^*(z,s)}{z} = z^k - \mu p_1^*.(s). \tag{7.22}$$

As was shown previously, the roots of $s\,z - (1-z)(\mu - \lambda z) = 0$ are $z = \gamma_1(s)$ and $z = \gamma_2(s)$. Since $|\gamma_1(s)| < 1$, if $\mathrm{Re}\, s > 0$, then $p^*(z,s)/z$ is bounded for $|z| \le 1$ and $\ne 0$; if $z \ne 0$, then from Eq. (7.22) we get for $\mu p_1^*(s)$ that [because in the case of $|z| \le 1$ the only root of $z^k - \mu p_1^*(s) = 0$ is $z = \gamma_1(s)$]

$$\mu p_1^*(s) = \gamma_1(s)^k = \left(\frac{\mu}{\lambda}\right)^k \gamma_2(s)^{-k}.$$

Using this and $\int_0^\infty \mathrm{e}^{-sx}\,\Pi_k'(x)\mathrm{d}x = \mu p_1^*(s)$ we have that

$$\Pi_k'(t) = \left(\frac{\mu}{\lambda}\right)^k \mathrm{e}^{-(\lambda+\mu)t} \left(\sqrt{\frac{\lambda}{\mu}}\right)^k k\, t^{-1} I_k(2\sqrt{\lambda\mu}t)$$

$$= \left(\sqrt{\frac{\mu}{\lambda}}\right)^k \frac{k}{t} \mathrm{e}^{-(\lambda+\mu)t} I_k(2\sqrt{\lambda\mu}t).$$

In the special case where $k = 1$, we have

$$\Pi_1'(t) = \sqrt{\frac{\mu}{\lambda}} \frac{1}{t} \mathrm{e}^{-(\lambda+\mu)t} I_1(2\sqrt{\lambda\mu}t).$$

(D) Distribution of virtual waiting time:

At time t the virtual waiting time, $W(t)$, satisfies

$$W(t) \stackrel{d}{=} \begin{cases} 0, & \text{if } L(t) = 0, \\ \sum_{i=1}^{k} \xi_i, & \text{if } L(t) = k, \end{cases}$$

where $\stackrel{d}{=}$ denotes the equality in distribution and $\xi_1, \ldots, \xi_k$ are the i.i.d. service times of the waiting customers ($\xi_i, i = 2, \ldots, k$) and the remaining service time of the customer being served (ξ_1). Due to the memoryless property of the exponential service time distribution, all of these random variables are exponentially distributed with parameter μ, and their CDF is $1 - \mathrm{e}^{-\mu x}$. According to the law of total probability, this gives

$$W(x,t) = \mathbf{P}\left(W(t) \le x\right) = \sum_{k=0}^{\infty} \mathbf{P}\left(W(t) \le x | L(t) = k\right) \mathbf{P}\left(L(t) = k\right) \quad (7.23)$$

$$= P_0(t) + \sum_{k=1}^{\infty} P_k(t)\, \mathbf{P}\left(\sum_{i=1}^{k} \xi_i \le x\right), \quad (7.24)$$

where $P_k(t) = \mathbf{P}\left(L(t) = k\right)$.

Introducing the Laplace transforms $W^*(x,s) = \int_0^\infty \mathrm{e}^{-st} W(x,t)\mathrm{d}t$ and $p_k^*(s) = \int_0^\infty \mathrm{e}^{-st} P_k(t)\mathrm{d}t$ from Eq. (7.24) we obtain

$$W^*(x,s) = p_0^*(s) + \sum_{k=1}^{\infty} p_k^*(s)\mathbf{P}\left(\sum_{i=1}^{k}\xi_i < x\right).$$

Since ξ_i, $i = 1,2,\ldots,k$, are independent exponentially distributed random variables with parameter μ we have

$$\mathbf{P}\left(\sum_{i=1}^{k}\xi_i \le x\right) = \mathbf{P}\left(\sum_{i=1}^{k}(\mu\xi_i) \le \mu x\right) = \int_0^{\mu x}\frac{u^{k-1}}{(k-1)!},\mathrm{e}^{-u}\mathrm{d}u$$

and according to point (B) we also have

$$p_k^*(s) = [\lambda(1-\gamma_1(s))(\gamma_2(s))^{k+1}]^{-1}, \quad k \ge 0.$$

Using all these expressions we obtain the Laplace transform of $W(x,t)$:

$$\begin{aligned} W^*(x,s) &= [\lambda(1-\gamma_1(s))\gamma_2(s)]^{-1}\left(1+\sum_{k=1}^{\infty}[\gamma_2(s)]^{-k}\int_0^{\mu x}\frac{u^{k-1}}{(k-1)!}\mathrm{e}^{-u}\mathrm{d}u\right) \\ &= [\lambda(1-\gamma_1(s))\gamma_2(s)]^{-1}\left(1+\sum_{k=1}^{\infty}(\gamma_2(s))^{-1}\int_0^{\mu x}\frac{[u/\gamma_2(s)]^{k-1}}{(k-1)!}\mathrm{e}^{-u}\mathrm{d}u\right) \\ &= [\lambda(1-\gamma_1(s))\gamma_2(s)]^{-1}\left(1+\frac{1}{\gamma_2(s)-1}(1-\mathrm{e}^{-\mu(1-\gamma_2^{-1}(s))x})\right). \end{aligned}$$

According to Eq. (7.24), in the case of a stable system ($\lambda/\mu < 1$) there exists the limit

$$\bar{W}(x) = \lim_{t\to\infty} W(x,t) = \pi_0 + \sum_{k=1}^{\infty}\pi_k\mathbf{P}\left(\sum_{i=1}^{k}\xi_i < x\right),$$

where according to Theorem 3.68

$$\pi_k = \lim_{t\to\infty} P_k(t) = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^k, \quad k \ge 0.$$

Thus

$$\bar{W}(x) = \left(1-\frac{\lambda}{\mu}\right)\left(1+\sum_{k=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^k\int_0^{\mu x}\frac{u^{k-1}}{(k-1)!}\mathrm{e}^{-u}\mathrm{d}u\right)$$

$$= \left(1 - \frac{\lambda}{\mu}\right)\left(1 + \frac{\lambda}{\mu}\int_0^{\mu x} \mathrm{e}^{-(1-\lambda/\mu)u}\mathrm{d}u\right) = 1 - \frac{\lambda}{\mu}\mathrm{e}^{-(\mu-\lambda)x}, \quad x > 0,$$

and

$$\lim_{t\to\infty} \mathbf{P}\,(W(t) = 0) = \lim_{t\to\infty} \mathbf{P}\,(L(t) = 0) = \lim_{t\to\infty} P_0(t) = \pi_0 = 1 - \frac{\lambda}{\mu}.$$

7.3 $M/M/m$ Queueing System

The arrival process (Poisson process at rate λ) and the service time distribution (exponential with parameter μ) are the same as before, but there are m servers in the service unit of this queueing system. While there is at least one idle server, an arriving customer is assigned to one of the idle servers upon arrival, and service of this customer starts immediately. If all the servers are busy at an arrival, then the arriving customer waits in the buffer. When i $(1 \le i \le m)$ servers are busy, the i service processes go on in parallel. Due to the memoryless property of the service time distribution, the remaining service times are also independent exponentially distributed random variables. The minimum of i independent exponentially distributed random variables with parameter μ is exponentially distributed with parameter $i\mu$. Another intuitive interpretation of the service process is through the service rate. A single server serves a customer at rate μ, i.e., the probability that an ongoing service will be completed in the next interval of length δ is $\mu\Delta + o(\Delta)$. If i servers are working in parallel, then they serve customers at a rate $i\mu$, i.e., the probability that one of the i ongoing service will be completed in the next Δ long interval is $i\mu\Delta + \sigma(\Delta)$. The transitions of the birth-death process describing the number of customers in the system are as follows:

$$\begin{aligned}
p_{i,i+1}(\Delta) &= \lambda\Delta + o\,(\Delta)\,, \quad (i = 0, 1, \ldots),\\
p_{i,i-1}(\Delta) &= i\mu\Delta + o\,(\Delta)\,, \quad (0 < i \le m),\\
p_{ii}(\Delta) &= 1 - (\lambda + i\mu)\Delta + o\,(\Delta)\,, \quad (0 \le i \le m),\\
p_{i,i}(\Delta) &= 1 - (\lambda + m\mu)\Delta + o\,(\Delta)\,, \quad (i \ge m).
\end{aligned}$$

The Markov chain is stable if $0 < \lambda < m\mu < \infty$. In this case the stationary equations are

$$\begin{aligned}
p_{k-1}\lambda + p_{k+1}(k+1)\mu &= p_k(\lambda + k\mu), \ \ 0 < k < m,\\
p_{k-1}\lambda + p_{k+1}m\mu &= p_k(\lambda + m\mu), \ \ k \ge m,\\
p_1\mu &= p_0\lambda\,.
\end{aligned}$$

The solution of this set of equations is

$$p_k = \frac{\lambda}{k\mu} p_{k-1} = \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} p_0 \quad \text{if } k = 1, 2, \cdots, m, \tag{7.25}$$

$$p_k = \frac{\lambda}{m\mu} p_{k-1}, \quad \text{if } k = m+1, m+2, \cdots, \tag{7.26}$$

whence

$$p_{m+i} = \left(\frac{\lambda}{m\mu}\right)^i p_m, \quad i \geq 1. \tag{7.27}$$

Combining the two cases we have

$$p_j = \begin{cases} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} p_0, & j = 1, 2, \cdots, m, \\ \left(\frac{\lambda}{\mu}\right)^j \frac{1}{m!} \frac{1}{m^{j-m}} p_0, & j > m, \end{cases} \tag{7.28}$$

from which the normalized solution of p_0 is

$$p_0 = \frac{1}{1 + \sum_{j=1}^{m} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} + \sum_{j=m+1}^{\infty} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{m!} \frac{1}{m^{j-m}}}. \tag{7.29}$$

The second term of the denominator can be rewritten as

$$\frac{m^m}{m!} \sum_{j=m}^{\infty} \left(\frac{\lambda}{m\mu}\right)^j = \frac{\left(\frac{\lambda}{m\mu}\right)^m}{\left(1 - \frac{\lambda}{m\mu}\right)} \frac{m^m}{m!}. \tag{7.30}$$

The mean system time can be computed as

$$\bar{T} = \bar{x} + \bar{W} = \frac{1}{\mu} + \sum_{k=m}^{\infty} \mathbf{E}(W \mid k) \; p_k^{(a)}, \tag{7.31}$$

where $p_k^{(a)}$ denotes the queue length distribution at arrival instants. In the case of an $M/M/m$ queue, $p_k^{(a)} = p_k$. The mean system time can be expressed as

$$\mathbf{E}(T) = \frac{1}{\mu} + \sum_{k=m}^{\infty} \frac{k - m + 1}{m\mu} p_k, \tag{7.32}$$

whence

$$\bar{T} = \frac{1}{\mu} + \sum_{k=m}^{\infty} \frac{k-m+1}{m\mu} p_m \left(\frac{\lambda}{m\mu}\right)^{k-m} = \frac{1}{\mu} + \frac{p_m}{m\mu} \sum_{k=m}^{\infty} (k-m+1) \left(\frac{\lambda}{m\mu}\right)^{k-m}. \tag{7.33}$$

Using that the arrival process is a Poisson process we further have

$$\bar{T} = \frac{1}{\mu} + \frac{p_m}{m\mu} \sum_{i=1}^{\infty} i \left(\frac{\lambda}{m\mu}\right)^{i-1} = \frac{1}{\mu} + \frac{m\mu p_m}{(m\mu-\lambda)^2}. \tag{7.34}$$

With the help of Little's law we can also compute the mean number of customers in the system:

$$\mathbf{E}(X) = \lambda \bar{T} = \frac{\lambda}{\mu} + \frac{m\lambda\mu p_m}{(m\mu-\lambda)^2}. \tag{7.35}$$

The probability that all servers will be busy and an arriving customer will have to wait is

$$\begin{aligned}
\mathbf{P}(\text{waiting}) &= \sum_{k=m}^{\infty} p_k = \sum_{k=m}^{\infty} p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{m!} \frac{m^m}{m^k} = \frac{m^m}{m!} p_0 \sum_{k=m}^{\infty} \left(\frac{\lambda}{m\mu}\right)^k \\
&= p_0 \frac{m^m}{m!} \frac{\left(\frac{\lambda}{m\mu}\right)^m}{1-\frac{\lambda}{m\mu}} = p_0 \frac{1}{m!} \left(\frac{\lambda}{\mu}\right)^m \frac{1}{1-\rho} \\
&= \frac{\frac{1}{m!} \left(\frac{\lambda}{\mu}\right)^m \frac{1}{1-\rho}}{\sum_{k=0}^{m-1} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} + \frac{1}{m!} \left(\frac{\lambda}{\mu}\right)^m \frac{1}{1-\rho}},
\end{aligned}$$

where $\rho = \dfrac{\lambda}{m\mu}$.

This expression is known as the C (or waiting probability) formula of Erlang [55]. The parameters of this formula are m the number of servers and the λ/μ ratio, which is also referred to as *traffic*. The shorthand notation of the C formula is $C(m, \lambda/\mu)$.

Example 7.4. There are four leased telephone lines between two sites of a company. Phone call requests arrive according to a Poisson process at a rate $1/2$ (calls/min). The lengths of the calls are exponentially distributed. The mean call holding time is 4 (min). If all lines are busy when a call arrives, then the caller waits until a telephone line becomes available. What is the probability that a caller will have to wait?

We have

$$\lambda = 1/2, \quad 1/\mu = 4, \quad a = \lambda/\mu = 2, \quad \rho = a/m = 2/4 = 0.5,$$

from which

$$p_0 = \frac{1}{1 + 2 + 2^2/2 + 2^3/6 + 16/24(1/(1-0.5)} = 3/23$$

and

$$C(4,2) = \frac{2^4/4!}{1-0.5}\frac{3}{23} = 4/23 = 0.17.$$

Example 7.5. Compare the performance of the $M/M/1$ and the $M/M/2$ queueing systems if $\lambda = 1/2$ in both systems and the service rates of the $M/M/1$ and the $M/M/2$ systems are $\mu_1 = 1$ and $\mu_2 = 1/2$, respectively.

The parameters of the $M/M/1$ queueing system are

$$\rho = \frac{\lambda}{\mu_1} = \frac{1/2}{1} = 0.5,$$

$$\mathbf{E}(W) = \frac{\rho/\mu}{1-\rho} = 1\,\text{s},$$

$$\mathbf{E}(T) = \frac{1/\mu}{1-\rho} = 2\,\text{s}.$$

The parameters of the $M/M/2$ queueing system are

$$a = \frac{\lambda}{\mu_2} = \frac{1/2}{1/2} = 1, \ \rho = a/m = 1/2 = 0.5,$$

$$p_0 = \frac{1}{1 + 2 + \frac{a^2/2}{1-0.5}} = 1/3,$$

from which

$$C(2,1) = \frac{a^2/2}{1-0.5}p_0 = 1/3,$$

$$\mathbf{E}\left(W'\right) = \frac{1/\mu_2}{1-\rho}C(2,1) = 2/3,$$

$$\mathbf{E}(T) = 2/3 + 1/\mu_2 = 8/3\,\text{s}.$$

Consequently, the system time of the $M/M/1$ system is lower, though its waiting time is higher.

7.4 $M/M/\infty$ Queueing System

An $M/M/\infty$ queueing system is obtained as the number of servers in the $M/M/m$ queueing system tends to infinity. Obviously, no waiting is possible at the limiting case because there is always an idle server in the system. The $M/M/\infty$ queue does not occur in practice, but this model can be used efficiently to approximate the behavior of high-capacity service units.

The analysis of an $M/M/\infty$ queueing system can be carried out in an analogous way to the analysis of the $M/M/1$ system. In an $M/M/\infty$ system, the number of customers, $L(t)$, is also a birth-death process with the following parameters:

$$p_0 = 1, \quad a_k = \lambda, \quad k \geq 0, \qquad b_k = k\mu, \quad k \geq 1,$$
$$\alpha_k = \lambda + k\mu, \quad p_k = \frac{\lambda}{\lambda + k\mu}, \quad q_k = \frac{k\mu}{\lambda + k\mu}, \quad k \geq 1.$$

In the case of an $M/M/\infty$ system,

$$P_k(t) = \exp\left\{-\frac{\lambda}{\mu}(1 - \mathrm{e}^{-\mu t})\right\} \frac{1}{k!}\left(\frac{\lambda}{\mu}(1 - \mathrm{e}^{-\mu t})\right)^k, \quad k \geq 0,$$

and according to Theorem 3.68,

$$\pi_k = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \mathrm{e}^{-\lambda/\mu}, \quad k \geq 0.$$

The condition of stability is $0 < \lambda, \mu < \infty$. The infinitesimal generator of the Markov chain describing the number of customers in the system contains the following nonzero elements:

$$q_{ij} = \begin{cases} \lambda_i = \lambda & \text{if} \quad i \geq 0, j = i + 1, \\ \mu_i = i\mu & \text{if} \quad i \geq 1, j = i - 1, \\ -\lambda_i - \mu_i = -\lambda - i\mu & \text{if} \quad i \geq 0, j = i, \end{cases}$$

whence

$$p_k = p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}, \qquad k \geq 1, \tag{7.36}$$

and if the Markov chain is stable, then the normalized solution of p_0 is

$$p_0 = \frac{1}{1 + \sum\limits_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}} = \frac{1}{\sum\limits_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}} = \mathrm{e}^{-\lambda/\mu},$$

and the other state probabilities are

$$p_k = \frac{(\lambda/\mu)^k}{k!} \mathrm{e}^{-\lambda/\mu}, \quad k \geq 0\,. \tag{7.37}$$

This means that the number of customers in the stationary $M/M/\infty$ queue is Poisson distributed with parameter λ/μ and

$$\mathbf{E}\,(X) = \bar{\lambda}\ \bar{T} = \lambda \frac{1}{\mu} = \frac{\lambda}{\mu}. \tag{7.38}$$

7.5 $M/M/m/m$ Queueing System

An $M/M/m/m$ queueing system contains m servers but does not contain a buffer for waiting customers. Thus customers that arrive while the servers are busy are lost. It can be interpreted as a finite-state variant of the $M/M/m$ queueing system because the number of customers in the system cannot exceed m. The infinitesimal generator of the Markov chain describing the number of customers in the system contains the following nonzero elements:

$$q_{ij} = \begin{cases} \lambda_i = \lambda & \text{if} \quad 0 \leq i < m, j = i+1, \\ \mu_i = i\mu & \text{if} \quad 1 \leq i \leq m, j = i-1, \\ -\lambda_i - \mu_i = -\lambda - i\mu & \text{if} \quad 0 \leq i < m, j = i, \\ -\mu_m = -m\mu & \text{if} \quad i = m, j = i. \end{cases}$$

The stationary distribution is

$$p_k = p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}, \quad k = 1, 2, \cdots, m\,, \tag{7.39}$$

where

$$p_0 = \frac{1}{\displaystyle\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}\,. \tag{7.40}$$

An $M/M/m/m$ system is stable ($p_k > 0 \ \ \forall k$) if $0 < \lambda, \mu < \infty$. The mean service time, the mean customer arrival intensity, and the mean number of customers are

$$\bar{x} = \frac{1}{\mu},\ \bar{\lambda} = \sum_{k=0}^{m-1} \lambda_k p_k = \lambda(1 - p_m), \quad \text{and } \mathbf{E}\,(X) = \frac{\lambda}{\mu}(1 - p_m). \tag{7.41}$$

which fulfills Little's law. $M/M/m/m$ queueing systems are referred to as loss systems in telecommunications. The probability that an arriving customer will be lost is

$$\mathbf{P}\,(\text{loss}) = p_m^{(a)} = p_m = \frac{\left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!}}{\sum_{k=0}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}} = B(m, \lambda/\mu), \tag{7.42}$$

which is known as the B (loss) formula of Erlang [55]. The dimensioning of switched telephone networks was based on this formula for several decades in the twentieth century.

Example 7.6. Consider the same system as in Example 7.4 and assume that the calls that arrive when all lines are busy are lost. Compute the parameters of this loss system, and compare them with those of the waiting system from Example 7.4:

$$p_{\text{loss}} = B(4,2) = \frac{16/24}{1 + 2 + 2^2/2 + 2^3/6 + 16/24} = \frac{2/3}{5 + 4/3 + 2/3}$$
$$= 2/21 = 0.095$$

where $B(4,2) = 0.095 < C(4,2)0.17$. This relation can be explained by the load of the two systems. In the case of a waiting system, all arriving customers must be served, while in the case of loss systems, the load of the servers is reduced by the lost customers.

7.6 $M/M/1//N$ Queueing System

All previous queueing systems have infinite populations and a state-independent Poisson customer arrival process. In an $M/M/1//N$ queueing system, the population is finite and the customer arrival intensity depends on the state of the system because the customers in the system do not contribute to new arrivals. For example, if all customers of the population are in the system, then the new customer arrival intensity reduces to 0. The infinitesimal generator of a Markov chain describing the possible changes in the system contains:

$$q_{ij} = \begin{cases} \lambda_i = (N-i)\lambda & \text{if} \quad 0 \le i < N, j = i+1, \\ \mu_i = i\mu & \text{if} \quad 1 \le i \le N, j = i-1, \\ -\lambda_i - \mu_i = -(N-i)\lambda - \mu & \text{if} \quad 0 \le i \le N, j = i. \end{cases}$$

The Markov chain is stable if $0 < \lambda, \mu < \infty$. In this case the stationary distribution is

$$p_k = p_0 \left(\frac{\lambda}{\mu}\right)^k [N(N-1)\cdots(N-k+1)] = p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{N!}{(N-k)!}$$
$$k = 1, 2, \cdots, N, \tag{7.43}$$

where

$$p_0 = \frac{1}{1 + \sum_{j=1}^{N} \left(\frac{\lambda}{\mu}\right)^j \frac{N!}{(N-j)!}} \tag{7.44}$$

and the utilization is $\rho = 1 - p_0$. The mean arrival intensity of this system is

$$\bar{\lambda} = \sum_{i=0}^{N} \lambda_i p_i = \sum_{i=0}^{N-1} (N-i)\ \lambda\ p_i. \tag{7.45}$$

According to Little's law,

$$\rho = \bar{\lambda}\mathbf{E}(x) = \bar{\lambda}/\mu,$$

from which we obtain an expression for the mean arrival intensity:

$$\bar{\lambda} = \frac{\rho}{\mathbf{E}(x)} = \mu\rho = \mu(1 - p_0). \tag{7.46}$$

There is another way to express the mean arrival intensity. We can interpret the life cycle of a customer such that it stays outside the system for an exponentially distributed amount of time with parameter λ and after that it enters the system and spends a system time (waiting time + service time) there. Thus, the cycle time of a customer is $1/\lambda + \mathbf{E}(T)$, and a customer generates a new arrival at the system once every cycle. Consequently, a customer generates arrivals at an average rate of $(1/\lambda + \mathbf{E}(T))^{-1}$, and the N members of the population generate arrivals at a rate of

$$\bar{\lambda} = \frac{N}{1/\lambda + \mathbf{E}(T)}.$$

From this expression we have

$$\mathbf{E}(T) = \frac{N}{\bar{\lambda}} - \frac{1}{\lambda}, \tag{7.47}$$

and using Little's law again we have

$$\mathbf{E}(X) = \bar{\lambda}\mathbf{E}(T) = N - \frac{\bar{\lambda}}{\lambda} \tag{7.48}$$

and $\mathbf{E}(W) = \mathbf{E}(T) - \frac{1}{\mu}$. The probability that a member of the population will be in the system is

$$\mathbf{P}(\text{in system}) = \frac{\mathbf{E}(T)}{1/\lambda + \mathbf{E}(T)}.$$

Example 7.7. There are N terminals in a computer system. Each terminal infinitely repeats the following steps:

- Generates a task in an exponentially distributed amount of time with parameter λ.
- Submits the task to the central processing unit.
- Waits for the answer.

The central processing unit processes a task in an exponentially distributed amount of time with parameter μ. Approximate the task completion rate and the system time of this system assuming the two extreme cases where the system is heavily loaded (N, λ/μ are small) and when the system load is light (N, λ/μ are large).

- In the case of a light load:

$$\mathbf{E}(T) \approx \frac{1}{\mu},$$

$$\bar{\lambda} = \frac{K}{1/\lambda + \mathbf{E}(T)} \approx \frac{K}{1/\lambda + 1/\mu}.$$

- In the case of a heavy load:

$$\bar{\lambda} \approx \mu,$$

$$\mathbf{E}(T) \approx \frac{K}{\mu} - \frac{1}{\alpha}.$$

7.7 Exercises

Exercise 7.1. Compute the mean and variance of the waiting time in an $M/M/1$ queue based on Wald's identity.

Exercise 7.2. Two kinds of customers arrive at a queueing system with three servers. Type 1 customers arrive according to a Poisson process at a rate λ_1. A type 1 customer occupies one server for an exponentially distributed amount of time with the parameter μ_1. Type 2 customers arrive according to a Poisson process at a rate λ_2. A type 2 customer occupies two servers for an exponentially distributed amount of time with the parameter μ_2. Compute the loss probability of type 2 customers if there is no buffer in the system.

Exercise 7.3. One shop assistant serves customers in a shop with an exponentially distributed service time with the parameter μ. The shop assistant wants to smoke after an exponentially distributed time with the parameter α. If the shop is idle, he leaves to smoke immediately. If he is busy when he wants to smoke, then he serves

the customers while the shop is not idle and then leaves to smoke. The length of the smoking break is exponentially distributed with the parameter β. Customers arrive according to a Poisson process at a rate λ. Compute the mean shopping time of customers if at most three customers can enter the shop (compute the same measure if infinitely many customers can enter the shop.).

Exercise 7.4. A queueing system has two servers and two types of customers. Type i customers arrive according to a Poisson process at a rate λ_i, and their service time is exponentially distributed with the parameter μ_i, $i = 1, 2$. Server i is typically assigned to type i customers. If there is a type i customer in the system when server i is idle, then it serves a type i customer. If there is no type i customer in the system when server i is idle, then it can serve a customer of the other type. The arrival of a new customer does not interrupt the ongoing service. Compute the loss probability of type i customers if the buffer size is 3.

Exercise 7.5. Two kinds of customers arrive at a discrete-time queueing system. In every time slot a type i customer arrives with probability p_i, $i = 1, 2$, and no customer arrives with probability $1 - p_1 - p_2$. There is one server. The service time of a type 1 customer is geometrically distributed with the parameter q_1. The service time of a type 2 customer is time slot k, and the buffer size is b. Compute the mean system time of type i customers for $i = 1, 2$ if $k = 1, 2$ and $b = 0, 3, \infty$.

Exercise 7.6. To improve the energy efficiency of a discrete-time queueing system, the server is switched off (goes on vacation) for a geometrically distributed amount of time with the parameter r if the system is idle at the end of a time slot. At the end of the vacation period the server starts serving customers (if any) or goes for another vacation (if none). In every time slot one customer arrives with probability p and no customer arrives with probability $1 - p$. The service time is geometrically distributed with the parameter q, and the buffer size is b. Compute the mean service time, the mean vacation time, and the mean idle time of the server for $b = 3, \infty$.

Exercise 7.7. Compute the stationary number of customers in an $M/M/2/3/4$ queue if $\lambda = 1$ and $\mu = 2$.

Exercise 7.8. Compute the loss probability of an $M/M/m/m/K$ system for $K > m$.

Exercise 7.9. Compare the probability of waiting in an $M/M/m$ queue with a loss probability in an $M/M/m/m$ queue for $m = 1, 2, 3$, where the arrival and service intensities are identical. Interpret the relation of the results.

Exercise 7.10. A complex system is composed of two main units. The failure and the repair time of unit i, $i = 1, 2$, are exponentially distributed with the parameters λ_i and μ_i, respectively. The units are maintained by a single repairman. Define the Markov chain of the system behavior if the service discipline of the repairman is FIFO, preemptive LIFO, or processor sharing if the repair of unit 1 has a preemptive priority over that of unit 2 and if the repair of unit 1 has a nonpreemptive priority over that of unit 2.

Exercise 7.11. Customers of a discrete-time queueing system (being served and waiting) can be lost. Each customer is lost with probability r in each time slot. One customer arrives with probability p (and with probability $1 - p$ no customer arrives) in each time slot, and the service time is geometrically distributed with the parameter q. Compute the probability of successful service completion if the buffer size is 3.

Chapter 8
Non-Markovian Queueing Systems

8.1 $M/G/1$ Queueing System

The $M/G/1$ queueing system (Fig. 8.1) is similar to the $M/M/1$ queueing system and the only difference is that the service time distribution is no exponential. First we mention some idea, most of which were described in the previous chapter in connection with an $M/M/1$ system.

8.1.1 Description of $M/G/1$ System

Conditions of functioning:

1. At the starting moment τ_0 the system is empty. For the sake of simplicity we generally assume $\tau_0 = 0$ (Fig. 8.2).
2. $\{N(t),\ t \geq \tau_0\}$ describes the number of entering customers; this is a Poisson process with intensity $\lambda > 0$.
3. There is one server functioning without breakdowns; after having served a customer it immediately starts serving the next one. If a customer enters the system and the server is busy, the customer joins the waiting queue. There is no limitation on the queue's size.
4. The service discipline is FCFS (FIFO).
5. The service times are independent identically distributed (i.i.d.) random variables with distribution function $\mathbf{P}(Y < x) = B(x)$ and mean $\mathbf{E}(Y) = \mu_B < \infty$, and they do not depend on the arrival process.

The main characteristic of the system is the queue length $\{L(t),\ t \geq 0\}$, i.e., how many customers are in the system at moment t. Let $L(t)$ be continuous from the right, i.e., $L(t) = L(t+0),\ t \geq 0$.

In the case of queueing systems, the basic issues concern the distribution of queue length, whether there exists a limit distribution and how it can be found, the

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, 225
DOI 10.1007/978-1-4614-5317-8_8, © Springer Science+Business Media, LLC 2013

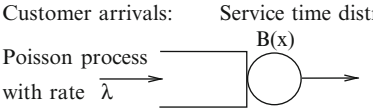


Fig. 8.1 $M/G/1$ system

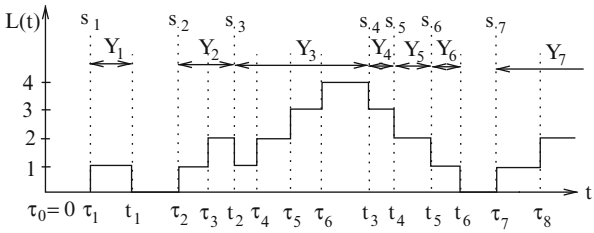


Fig. 8.2 Number of customers in $M/G/1$ system

average number of customers in the system, etc. In this chapter we will deal with the asymptotic behavior of queue length $L(t)$ as $t \to \infty$.
We introduce the following notations:

- $\tau_0 + X_1$: moment of entry of first customer; X_n: interarrival time between $(n-1)$st and nth customers;
- $\tau_n = \tau_0 + X_1 + \ldots + X_n$ $(n \geq 1)$: moment of entry of nth customer;
- Y_n: service time of nth customer;
- s_n, $n \geq 1$: starting moment of service of nth customer;
- t_n, $n \geq 1$: moment when nth customer leaves system (service in system is completed at this moment).

According to these assumptions, $\{(X_n, Y_n),\ n \geq 1\}$ is a sequence of i.i.d. random variables, where the components of vectors are independent, too. Furthermore, the intervals between consecutive arrivals X_n, $n \geq 1$, have exponential distribution with parameter λ, the service times Y_n, $n \geq 1$, have distribution function $B(x)$. It is also clear that $\{\tau_n,\ n \geq 1\}$, are moments of jumps of the Poisson process $N(t)$ (Fig. 8.2).

8.1.2 Main Differences Between $M/M/1$ and $M/G/1$ Systems

For the $M/M/1$ system both the interarrival and service times are independent exponentially distributed random variables. These distributions have the memoryless property, so one can derive that $\{L(t),\ t \geq 0\}$ is a Markov (birth-death) process. This fact simplifies the investigation of the system.

In the $M/G/1$ queueing system the examined processes (queue length, waiting time, etc.) are not necessarily of the Markov type since the service time distribution may not have the memoryless property, so their investigation requires different methods.

The foregoing conditions do not guarantee that $L(t)$ is a Markov process, but by means of an auxiliary variable one can make it a Markov process with an extended state vector.

If $U(t)$ denotes the service time passed till t [$U(t)$ is right continuous], then the vector process $\{(L(t), U(t)),\ t \geq 0\}$ is already Markov and can be considered as the state of the queueing system.

Generally, the vector $W(t) = \{L(t), U(t)\}$ $(t \geq \tau_0)$ describing (from a certain viewpoint) the functioning of a system is called the *state vector* of the system if at arbitrary $t_1 > t$ one can determine the vector $W(t_1)$ in a stochastic sense based on the value of $W(t)$ and the arrivals for $(t, t_1]$.

Compared with the $M/M/1$ system the difference is not only that the system state is characterized by a vector process, but – and this is an important feature – the state space will not be discrete since the possible values of $U(t)$ are not discrete and take on values from the set $R_+ = [0, \infty)$ (or its subset).

8.1.3 Main Methods for Investigating $M/G/1$ System

1. **Method of embedded Markov chains**, also called Kendall's method because its wide use is connected with Kendall [52]. This method appeared in the 1950s, but the possibility of such an approach was noted by Khinchin [53] (see also Palm [75]). We will consider this method in detail.
2. **Lindley's integral equation** [64]: can be derived for the more general $G/G/1$ systems and, hence, is applicable in our case, too. This approach leads to a special Wiener–Hopf type of integral equation for the limit distribution of customer waiting time.
3. **Method of auxiliary variables** [24, 52]: based on the fact that the system may be investigated via the state vector $\{L(t), U(t),\ t \geq 0\}$ with the help of an auxiliary variable $U(t)$ (Fig. 8.3). Instead of the time interval from the beginning of service, one can use the time interval till the end of service (Henderson [42]).
4. **Method of random walk** and **combinatorial approach** [90].
5. **Method of recurrent processes** [14, 15].

In the following sections we will investigate the $M/G/1$ queue using these approaches.

8.2 Embedded Markov Chains

This method includes the following steps:

(A) Choose random $(\tau_0 = 0 <) t_1 < t_2 < \ldots$ moments when the process describing the evolution of system is of a Markov type.
(B) Prove the ergodicity of the Markov chain $L_n = L(t_n)$, $n \geq 1$.
(C) Determine the ergodic distribution

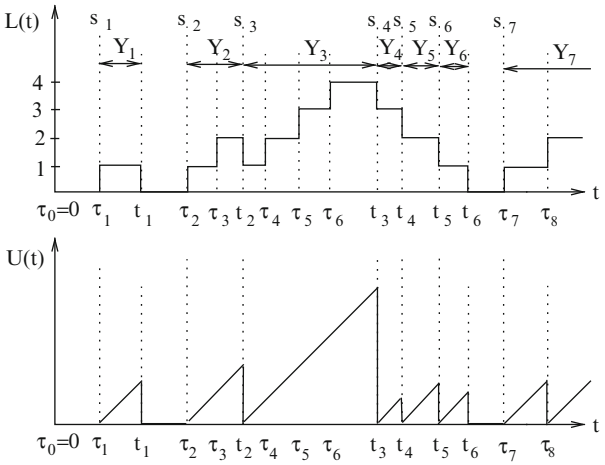


Fig. 8.3 $L(t)$ and $U(t)$ process of $M/G/1$ queue

$$\pi_k = \lim_{n\to\infty} \mathbf{P}\,(L_n = k)\,, \quad k \geq 0,$$

of the Markov chain.

(D) Prove the coincidence of limiting values $\lim_{t\to\infty} \mathbf{P}\,(L(t) = k) = \pi_k, k \geq 0$.

8.2.1 Step (A): Determining Queue Length

As earlier, t_n $(n = 1, 2, \ldots)$ denotes the moment when the service of the nth customer is completed. Let $L_n = L(t_n)$, $n \geq 1$ $(L_0 = 0)$. Since $Y_{t_n} = 0$, $n \geq 0$, at moments t_n, the behavior of the state-vector process $\{L(t), Y_t),\ t \geq 0\}$ is described by the sequence $\{L_n, n \geq 1\}$. In our case the main idea of application of embedded Markov chains is to consider the process $\{(L(t), Y_t),\ t \geq 0\}$ at moments t_n, $n = 1, 2, \ldots$. Using this method we come to a Markov chain $\{L_n, n \geq 1\}$ with countable state space $\mathcal{X} = \{0, 1, 2, \ldots\}$, and so we obtain the final result [see (D) for the method]. In practice this means that the states of the system (the number of customers in the system) are considered at moments just after having served a customer. In this restricted view of the process, every state transition between consecutive service completion moments (e.g., customer arrival) is considered at the service completion moments.

The process $L(t)$ is Markov regenerative; this fact will be used at the proof of step (D).

We prove the following theorem, fulfilling the tasks formulated in steps (A) and (B).

Theorem 8.1. *The stochastic process* $\{L_n, n \geq 1\}$ *is a homogeneous, irreducible, aperiodic Markov chain with state space* $\mathcal{X} = \{0, 1, 2, \ldots\}$. *If the condition* $\rho = \lambda\mu_B < 1$ *is fulfilled, then the Markov chain* $\{L_n, n \geq 1\}$ *is ergodic.*

Proof. First we prove that the stochastic process $\{L_n, n \geq 1\}$ is a Markov chain. Let Δ_n, $n = 1, 2, \ldots$, denote the number of customers entering the system for the service time Y_n of the nth customer, i.e.,

$$\Delta_n = N(t_n) - N(t_n - Y_n) = N(s_n + Y_n) - N(s_n), \quad n \geq 1.$$

Service to the nth customer may start at $s_n = t_n - Y_n > t_{n-1}$ if the system is empty at t_{n-1}, i.e., $L_{n-1} = 0$. In this case, $s_n = \tau_n$, and consequently $\Delta_n = N(t_n) - N(t_{n-1})$. Then

$$L_n = \begin{cases} L_{n-1} - 1 + \Delta_n, & \text{if } L_{n-1} > 0, \\ \Delta_n, & \text{if } L_{n-1} = 0, \end{cases}$$

or

$$L_n = \mathcal{I}_{\{L_{n-1}>0\}}(L_{n-1} - 1) + \Delta_n = (L_{n-1} - 1)^+ + \Delta_n. \tag{8.1}$$

Since the arrival process $\{N(t), t \geq 0\}$ is Poisson, independent of service times $\{Y_n, n \geq 1\}$, the number of customers Δ_n entering at service time Y_n is independent of $L_1, \ldots, L_{n-1}$, and consequently the sequence L_n constitutes a Markov chain.

Now we determine the distribution and mean value of the random variable Δ_n $(n \geq 1)$.

Using the fact that the behavior of $N(t)$ is independent of past events, the distribution of Δ_n can be written by the total mean value formula

$$a_k = \mathbf{P}(\Delta_n = k) = \int_0^\infty \mathbf{P}(\Delta_1 = k | Y_1 = x)\, \mathrm{d}B(x) = \int_0^\infty \frac{(\lambda x)^k}{k!} \mathrm{e}^{-\lambda x} \mathrm{d}B(x), \ k \geq 0.$$

Excluding the degenerate case $\mathbf{P}(Y = 0) = 1$, the inequality $a_k > 0$, $k \geq 0$, is always valid.

For the mean value of Δ_n we obtain

$$\begin{aligned} \mathbf{E}(\Delta_n) &= \sum_{k=1}^\infty k a_k = \sum_{k=1}^\infty \int_0^\infty \frac{(\lambda x)^k}{(k-1)!} \mathrm{e}^{-\lambda x} \mathrm{d}B(x) \\ &= \int_0^\infty \lambda x \sum_{k=0}^\infty \frac{(\lambda x)^k}{k!} \mathrm{e}^{-\lambda x} \mathrm{d}B(x) = \lambda \int_0^\infty x \mathrm{d}B(x) = \lambda\mu_B = \rho. \end{aligned} \tag{8.2}$$

Since $\lambda\mu_B < 1$, by Eq. (8.1), the Foster criterion is fulfilled (Theorem 3.42). □

The possibility of changing the order of summation and integration in the previous formula follows from the Fubini theorem but can also be proved in an elementary way. Since for the function

$$Q(A,n) = \int_0^A \sum_{k=1}^{n} \frac{(\lambda x)^k}{(k-1)!} \mathrm{e}^{-\lambda x} \mathrm{d}B(x), \quad A \in R_+, \quad n \in N,$$

there exists a limit as $A \to \infty$ and $n \to \infty$, and moreover $\mu_B = \int_0^\infty x \mathrm{d}B(x) < \infty$, then as $A \to \infty$ uniformly in n

$$\begin{aligned} |Q(\infty,n) - Q(A,n)| &= \int_A^\infty \sum_{k=1}^{n} \frac{(\lambda x)^k}{(k-1)!} \mathrm{e}^{-\lambda x} \mathrm{d}B(x) \\ &\le \int_A^\infty \lambda x \sum_{k=0}^{\infty} \frac{(\lambda x)^k}{k!} \mathrm{e}^{-\lambda x} \mathrm{d}B(x) = \int_A^\infty \lambda x \mathrm{d}B(x) \to 0, \end{aligned}$$

from which the interchangeability follows.

Proof of Homogeneity Let

$$p_{ij}(n) = \mathbf{P}\left(L_{n+1} = j \,|\, L_n = i\right), \quad i, j \ge 0, \; n \ge 0,$$

be one-step transition probabilities. Then, using Eq. (8.1),

$$p_{ij} = \mathbf{P}\left(\mathcal{I}_{\{i>0\}}(i-1) + \Delta_{n+1} = j\right),$$

so

$$p_{ij}(n) = p_{ij} = \begin{cases} a_j & \text{if } i = 0,1, \; j = 0,1,2,\ldots, \\ 0 & \text{if } i \ge 2, \; j \le i-2, \\ a_{j+1-i} & \text{if } i \ge 2, \; j \ge i-1, \end{cases}$$

i.e., the sequence $\{L_n, n \ge 0\}$ is a homogeneous Markov chain. This behavior is depicted in Fig. 8.4, and the associated matrix of one-step transition probabilities may be written in the form

$$P = (p_{ij})_{i,j=1}^{\infty} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & \ldots \\ a_0 & a_1 & a_2 & a_3 & \ldots \\ 0 & a_0 & a_1 & a_2 & \ldots \\ 0 & 0 & a_0 & a_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{8.3}$$

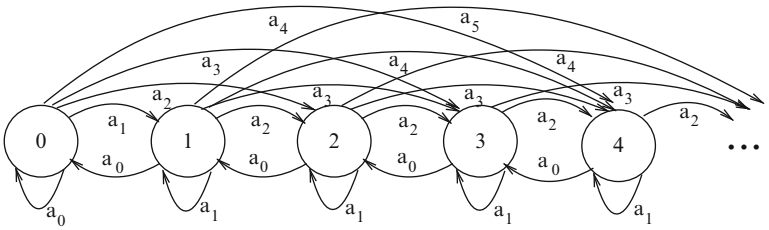


Fig. 8.4 Embedded Markov chain of $M/G/1$ queue

In this matrix, a_i gives the probability that i customers arrive at the system while a customer is being served. Fixing the initial state, adding the arriving customers, and subtracting the customer being served, we get the next state. We can descend one level if no new customers appear, remain at the same level if one new customer arrives, and go up if at least two new customers arrive. This explains the structure of the matrix. In this matrix the first two rows coincide. In the case of one already present customer, the foregoing reasoning is valid, but the zero state is a special situation. We arrive at the zero state when the last customer in a busy period is served. After a free period the first customer of the next busy period arrives, and we will consider the system state after this customer has been served. This new state will be determined by the number of customers arriving while this customer is being served. The coincidence of two rows is explained by the fact that in both cases we must consider the number of new customers arriving for the service of one customer. In the first case it is within a busy period, while in another case it is at the beginning of a busy period.

8.2.2 Proof of Irreducibility and Aperiodicity

Both properties may be derived from the matrix of one-step transition probabilities, but they may also be obtained from the following considerations.

Since the interarrival times have an exponential distribution with the parameter λ, it is clear that

- From arbitrary state $i \in \mathcal{X}$ for i services (steps) with positive probability we arrive at the state 0; this is enough so that no new customers enter.
- From state 0 with positive probability we can get to any state $j \in \mathcal{X}$ in one step.

The i and one-step transition probabilities (in the case of arbitrary $i, j \in \mathcal{X}$) are $p_{i0}^{(i)} > 0$, $p_{0j}^{(1)} > 0$, and consequently $p_{ij}^{(i+1)} > 0$, from which it follows that the Markov chain $\{L_n, n \geq 0\}$ is irreducible [for all $i, j \in \mathcal{X}$ there exists such n that $p_{ij}^{(n)} > 0$].

Obviously, for arbitrary $i \in \mathcal{X}$ $p_{ii}^{(1)} > 0$ (since for all $i \geq 1$ for the service of a customer with positive probability a new customer enters and there is no entry at $i = 0$). So the Markov chain $\{L_n, n \geq 0\}$ is aperiodic [if $d(i)$ is the period of state i, i.e., $d(i) = \{$greatest common divisor (g.c.d.) of n for which $p_{ii}^{(n)} > 0\}$, in our case $d(i) = 1$].

8.2.3 Step (B): Proof of Ergodicity

One way to prove ergodicity is to show that all states of the Markov chain are recurrent nonzero ones (with probability 1 it returns to all states and the mean value of the return time is finite, i.e.,

$$F_{ii} = \sum_{n=1}^{\infty} f_{ii}(n) = 1, \quad m_i = \sum_{n=0}^{\infty} n f_{ii}(n) < \infty,$$

where $f_{ij}(n)$ is the probability that the Markov chain which starts from state i goes to state j for the first time in the nth step. This approach requires a lot of computation, so we use the sufficient condition for the ergodicity of Markov chains obtained by Klimov (Theorem 3.41).

We check the conditions of Theorem 3.41 in the case $\rho < 1$. It is enough to find a function $g(i), i \in \mathcal{X}$, for which its conditions are fulfilled.

Let $\varepsilon = 1 - \rho\ (> 0)$ and $g(i) = i,\ i \geq 0$ (this case is known in the literature as Foster's criterion). From Eq. (8.1) it follows that

$$\mathbf{E}\,(g(L_{n+1})|L_n = i) = \mathbf{E}\,(i - 1 + \Delta_{n+1}) = i - 1 + \lambda\mu_B = i - \varepsilon, \quad i \geq 1,$$

and

$$\mathbf{E}\,(g(L_{n+1})|L_n = 0) = \mathbf{E}\,(\Delta_{n+1}) = \lambda\mu_B = 1 - \varepsilon, \quad i = 0,$$

i.e., the conditions of Klimov's theorem are fulfilled, and we have proved the ergodicity of the Markov chain $\{L_n, n \geq 0\}$.

8.2.4 Pollaczek–Khinchin Mean Value Formula

Equation (8.1) makes it possible to find the moments of ergodic distribution. We present it for the case of mean value; the computations are similar for other moments. The derivation requires less computation than the later Pollaczek–Khinchin transform equation, but in that case we automatically obtain the necessary conditions ($\rho < 1$ and the service time has finite second moment).

Assume that the following finite limits exist:

$$\lim_{n\to\infty} \mathbf{E}\,(L_n) = m_1 \quad \text{and} \quad \lim_{n\to\infty} \mathbf{E}\left(L_n^2\right) = m_2 \tag{8.4}$$

(we do not deal with the conditions of existence).

Equation (8.4) follows from (8.1) if $\mathbf{E}\left(\Delta_n^2\right) < \infty$, and the service time also has finite second moment. Taking on both sides of Eq. (8.1) limit as $n \to \infty$

$$\begin{aligned} m_1 &= \lim_{n\to\infty} \mathbf{E}\,(L_n) = \lim_{n\to\infty} [\mathbf{E}\left(\mathcal{I}_{\{L_{n-1}>0\}}(L_{n-1}-1)\right) + \mathbf{E}\,(\Delta_n)] \\ &= \lim_{n\to\infty} [\mathbf{E}\,(L_{n-1}) - \mathbf{E}\left(\mathcal{I}_{\{L_{n-1}>0\}}\right) + \rho] = m_1 - \lim_{n\to\infty} \mathbf{E}\left(\mathcal{I}_{\{L_{n-1}>0\}}\right) + \rho, \end{aligned}$$

whence

$$\lim_{n\to\infty} \mathbf{P}\,(L_n > 0) = \lim_{n\to\infty} \mathbf{E}\left(\mathcal{I}_{\{L_{n-1}>0\}}\right) = \rho$$

and

$$\pi_0 = \lim_{n\to\infty} \mathbf{P}\,(L_n = 0) = 1 - \lim_{n\to\infty} \mathbf{E}\left(\mathcal{I}_{\{L_{n-1}>0\}}\right) = 1 - \rho.$$

Though this procedure leads to important results, it does not produce the desired mean value. Repeating it for the second moments we meet our objective. Using the independence of L_{n-1} and Δ_n, we obtain

$$\begin{aligned} m_2 &= \lim_{n\to\infty} \mathbf{E}\left(L_n^2\right) \\ &= \lim_{n\to\infty} \mathbf{E}\left((L_{n-1}^2 - 2L_{n-1} + 1)\mathcal{I}_{\{L_{n-1}>0\}} + 2(L_{n-1}-1)\mathcal{I}_{\{L_{n-1}>0\}}\,\Delta_n + \Delta_n^2\right) \\ &= m_2 - 2m_1 + \rho + 2\,m_1\,\rho - 2\rho^2 + \mathbf{E}\left(\Delta_1^2\right), \end{aligned}$$

whence

$$m_1 = \frac{\rho - 2\rho^2 + \mathbf{E}\left(\Delta_1^2\right)}{2(1-\rho)} = \rho + \frac{\mathbf{E}\left(\Delta_1^2\right) - \rho}{2(1-\rho)}.$$

Later, by means of the generating function, we obtain the equality $\mathbf{E}\left(\Delta_1^2\right) = \lambda^2 \mathbf{E}\left(Y_1^2\right) + \rho$; using it from the last equation we come to the Pollaczek–Khinchin mean value formula:

$$m_1 = \rho + \frac{\lambda^2 \mathbf{E}\left(Y_1^2\right)}{2(1-\rho)}. \tag{8.5}$$

8.2.5 *Proof of Equality* $\mathbf{E}\left(\Delta_1^2\right) = \lambda^2 \mathbf{E}\left(Y_1^2\right) + \rho$

Let $B^{\sim}(s) = \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}B(x), s \geq 0$, be the Laplace–Stieljes transform of the distribution function $B(x)$. The generating function of entering customers for one service will be

$$\mathbf{E}\left(z^{\Delta_1}\right) = A(z) = \sum_{i=0}^{\infty} a_i z^i = \sum_{i=0}^{\infty} \int_0^\infty \frac{(\lambda x z)^i}{i!} \mathrm{e}^{-\lambda x}\mathrm{d}B(x) = B^{\sim}(\lambda(1-z)). \tag{8.6}$$

Similarly to the derivation of mean value $\mathbf{E}(\Delta_1)$ we get

$$A'(1) = \mathbf{E}(\Delta_1) = \sum_{k=1}^{\infty} k a_k = \sum_{k=1}^{\infty} k \int_0^{\infty} \frac{(\lambda x)^k}{k!} \mathrm{e}^{-\lambda x} \mathrm{d}B(x)$$

$$= \int_0^{\infty} \lambda x \sum_{k=0}^{\infty} \frac{(\lambda x)^k}{k!} \mathrm{e}^{-\lambda x} \mathrm{d}B(x) = \lambda \int_0^{\infty} x \mathrm{d}B(x) = \rho. \tag{8.7}$$

There exists a second moment of service time, so the Laplace–Stieltjes transform is twice continuously differentiable from the right, and for the right derivatives

$$B^{\sim\prime}(0) = -\int_0^{\infty} x \mathrm{d}B(x) = -\mathbf{E}(Y_1) = -\mu_B,$$

$$B^{\sim\prime\prime}(0) = \int_0^{\infty} x^2 \mathrm{d}B(x) = \mathbf{E}\left(Y_1^2\right).$$

From here (and taking the left-side derivatives at point 1)

$$\mathbf{E}(\Delta_1) = A'(1) = -\lambda B^{\sim\prime}(0) = \lambda \mu_B = \rho,$$

$$\mathbf{E}\left(\Delta_1^2\right) = (zA'(z))'_{z=1} = -\lambda B^{\sim\prime}(0) + \lambda^2 B^{\sim\prime\prime}(0) = \lambda \mu_B + \lambda^2 \mathbf{E}\left(Y_1^2\right)$$

$$= \rho + \lambda^2 \mathbf{E}\left(Y_1^2\right).$$

8.2.6 Step (C): Ergodic Distribution of Queue Length

From the ergodicity of the Markov chain $\{L_n, n \geq 0\}$ follows the existence of the ergodic distribution

$$\pi_k = \lim_{n \to \infty} \mathbf{P}(L_n = k), \quad k = 0, 1, 2, \ldots,$$

which can be obtained as the solution of the system of equations

$$\pi_k = \sum_{j=0}^{\infty} \pi_j p_{jk}, \quad k = 0, 1, 2, \ldots,$$

$$\sum_{k=0}^{\infty} \pi_k = 1.$$

The matrix P has a special structure [see Eq. (8.3)], and the stationary equations take the form

$$\begin{aligned}\pi_k &= \pi_0 a_k + \pi_1 a_k + \pi_2 a_{k-1} + \ldots + \pi_{k+1} a_0 \\ &= \sum_{i=0}^{k} \pi_{k-i+1} a_i + \pi_0 a_k, \quad k = 0, 1, 2, \ldots . \end{aligned} \tag{8.8}$$

We solve this system of equations by the method of generating functions. Let us introduce the notation

$$\pi(z) = \sum_{k=0}^{\infty} \pi_k z^k, \quad A(z) = \sum_{k=0}^{\infty} a_k z^k, \quad |z| \le 1.$$

First, $\pi(1) = A(1) = 1$, and, according to our previous computations, $A'(1) = \lim_{z \to 1-0} A'(z) = \sum_{k=1}^{\infty} k a_k \ (= \mathbf{E}(\Delta_1)) = \rho$. Multiplying both sides of Eq. (8.8) by z^k and summing up by k for $k \ge 0$, we obtain

$$\begin{aligned}\pi(z) &= \sum_{k=0}^{\infty} z^k \sum_{m=0}^{k} a_m \pi_{k-m+1} + \pi_0 A(z) \\ &= \sum_{m=0}^{\infty} a_m z^m \sum_{k=m}^{\infty} z^{k-m} \pi_{k-m+1} + \pi_0 A(z) \\ &= A(z) \frac{\pi(z) - \pi_0}{z} + \pi_0 A(z), \end{aligned}$$

whence

$$\pi(z)[1 - A(z)/z] = -\pi_0 A(z)(1/z - 1),$$

and so

$$\pi(z) = \pi_0 \frac{(1-z)A(z)}{A(z) - z}, \quad |z| < 1. \tag{8.9}$$

This includes the unknown probability π_0, which will be found from the condition $\pi(1) = \sum_{k=0}^{\infty} \pi_k = 1$. In the derivation of the Pollaczek–Khinchin mean value formula under special conditions we already found the value of π_0, and here it will come from Eq. (8.9) when $\int_0^{\infty} x^2 \mathrm{d}B(x) < \infty$.

$\pi(z)$ is continuous from left at point 1, so at $z = 1$ the numerator and denominator of Eq. (8.9) disappear. By l'Hospital's rule

$$\pi(1) = \lim_{z \to 1-0} \pi(z) = \lim_{z \to 1-0} \pi_0 \frac{-A(z) + A(z)(1-z)}{A(z) - 1} = \frac{-\pi_0}{A'(1) - 1} = \frac{\pi_0}{1 - \rho} = 1,$$

and we obtain $\pi_0 = 1 - \rho$.

Earlier we proved (8.6), i.e.,

$$A(z) = B^{\sim}(\lambda(1-z)), \quad |z| \le 1.$$

From it and Eq. (8.9) we get the Pollaczek–Khinchin transform equation (or, more precisely, one of its forms):

$$\pi(z) = \frac{(1-\rho)(1-z)B^{\sim}(\lambda\,(1-z))}{B^{\sim}(\lambda\,(1-z)) - z}. \tag{8.10}$$

Recall that this gives the generating function of ergodic distribution for the embedded Markov chain $\{L_n, n \ge 0\}$.

Corollary 8.2. *The inversion of the Pollaczek–Khinchin transform equation generally is not simple, but the moments may be obtained from it without inversion.*

Taking into account $B^{\sim\prime}(0) = -\mathbf{E}(Y) = -\mu_B$, $B^{\sim\prime\prime}(0) = \mathbf{E}\left(Y^2\right) = \int_0^\infty x^2 \mathrm{d}B(x)$, and using l'Hospital's rule twice we obtain the mean value of the number of customers in the system (Pollaczek–Khinchin mean value formula):

$$\begin{aligned}
&\sum_{k=0}^{\infty} k\pi_k = \pi'(1) = \lim_{z\to 1-} \pi(z) \\
&= \lim_{z\to 1-} (1-\rho)\frac{-B^{\sim 2}(\lambda(1-z)) + \lambda z B^{\sim\prime}(\lambda(1-z)) - \lambda z^2 B^{\sim\prime}(\lambda(1-z)) + B^{\sim}(\lambda(1-z))}{[B^{\sim}(\lambda(1-z)) - z]^2} \\
&= \lim_{z\to 1-} (1-\rho)\frac{\lambda^2 B^{\sim\prime\prime}(\lambda(1-z)) - 2\lambda B^{\sim\prime}(\lambda(1-z)) - 2\lambda^2[B^{\sim\prime}(\lambda(1-z))]^2}{2\lambda^2[B^{\sim\prime}(\lambda(1-z))]^2 + 4\lambda B^{\sim\prime}(\lambda(1-z)) + 2} \\
&= (1-\rho)\frac{\lambda^2 B^{\sim\prime\prime}(0) + 2\rho - 2\rho^2}{2(1-\rho)^2} \\
&= \rho + \frac{\lambda^2 \mathbf{E}\left(Y^2\right)}{2(1-\rho)}.
\end{aligned}$$

The variance of stationary queue length can be computed in a similar way:

$$\sigma^2 = \frac{\lambda^3 \mathbf{E}\left(Y^3\right)}{3(1-\rho)} + \frac{\lambda^4 \mathbf{E}\left(Y^2\right)}{4(1-\rho)^2} + \frac{\lambda^2 \mathbf{E}\left(Y^2\right)(3-2\rho)}{2(1-\rho)} + \rho(1-\rho),$$

where $\mathbf{E}\left(Y^i\right), i = 2, 3$, denotes the ith moment of service time [19].

Example 8.3 (Inversion in the case of $M/M/1$ system). In this case the intensity of arrivals is $\lambda > 0$, the intensity of service $\mu > 0$ (the interarrival and service times are independent exponentially distributed random variables with parameters λ and μ, respectively). Then

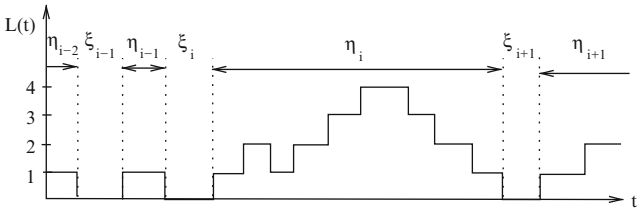


Fig. 8.5 Busy periods

$$B^{\sim}(s) = \frac{\mu}{s+\mu}, \quad \text{Re}\, s > -\mu,$$

$$\pi(z) = \frac{\mu}{\lambda - \lambda z + \mu} \cdot \frac{(1-\rho)(1-z)}{[\mu/(\lambda - \lambda z + \mu)] - z}.$$

Since $\rho = \lambda\tau = \lambda/\mu$,

$$\pi(z) = \frac{1-\rho}{1-\rho z},$$

and for the stationary distribution we obtain

$$\pi_k = (1-\rho)\rho^k, \quad k \geq 0.$$

8.2.7 *Investigation on Busy/Free Intervals in M/G/1 Queueing System*

Observing a queueing system we see that there are periods during which it is empty or occupied. The time interval when the server is occupied is called the busy period. It begins with the arrival of a customer at the empty system and is finished when the last customer leaves the system (Fig. 8.5).

If $(\xi_i, \eta_i), i = 1, 2, \ldots$, denote consecutive free and busy periods, then (ξ_i, η_i) is a sequence of i.i.d. random variables, where the components ξ_i and η_i are also independent of each other. The sequence $(\xi_i + \eta_i), i = 1, 2, \ldots$, is a renewal process, ξ_i has an exponential distribution with the parameter λ. Finding the distribution of busy periods η_i is more complicated and will be considered later.

Let $\Psi(x) = \mathbf{P}(\eta_i \leq x)$. Assume that at moment $t = 0$ a customer enters the system and a busy period begins. The customer's service time is $Y = y$. There are two cases:

1. During service no new customer enters the system and the busy period ends, i.e., its duration is $Y = y$.
2. For y, $n \geq 1$, customers enter the system and the busy period continues (Fig. 8.6).

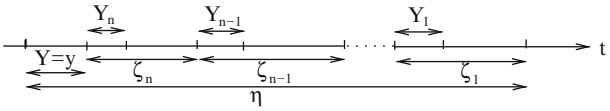


Fig. 8.6 Length of busy period

In the last case n successive service times are denoted by $Y_1, Y_2, \ldots, Y_n$. Assume that the service is realized in inverse order, i.e., according to the LCFS discipline [88], then according to our assumptions the interarrival and service times are independent. Their distributions are exponential with parameter λ and $B(x)$, and the distribution of busy periods remains the same (Ψ). The whole busy period η can be divided into intervals $Y, \zeta_n, \zeta_{n-1}, \ldots, \zeta_1$ (if $n = 0$, then $\eta = Y$), where $\zeta_n, \zeta_{n-1}, \ldots, \zeta_1$ mean busy periods generated by the different customers, they are

1. Independent,
2. Identically distributed, and
3. Their distribution coincides with that of η.

By the formula of total probability ($\zeta_n + \cdots + \zeta_1 = 0$ if $n = 0$),

$$\begin{aligned}
\Psi(x) &= \mathbf{P}\,(\eta \le x) \\
&= \mathbf{P}\,(Y + \zeta_n + \cdots + \zeta_1 \le x) \\
&= \sum_{j=0}^{\infty} \mathbf{P}\,(Y + \zeta_n + \cdots + \zeta_1 \le x,\ n = j) \\
&= \int_0^{\infty} \sum_{j=0}^{\infty} \mathbf{P}\left(Y + \zeta_j + \cdots + \zeta_1 \le x,\ n = j \mid Y = y\right) \mathrm{d}B(y) \\
&= \int_0^{\infty} \sum_{j=0}^{\infty} \mathbf{P}\left(y + \zeta_j + \cdots + \zeta_1 \le x\right) \frac{(\lambda y)^j}{j!} \mathrm{e}^{-\lambda y}\, \mathrm{d}B(y) \\
&= \int_0^{\infty} \sum_{j=0}^{\infty} \Psi_j(x - y) \frac{(\lambda y)^j}{j!} \mathrm{e}^{-\lambda y}\, \mathrm{d}B(y)
\end{aligned}$$

(the order of summation and integration may be changed), where

$$\Psi_j(x) = \mathbf{P}\left(\zeta_1 + \cdots + \zeta_j \le x\right).$$

This functional equation will be simpler if we use the Laplace–Stieltjes transforms. Let

$$B^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx}\,\mathrm{d}B(x), \qquad \Psi^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx}\,\mathrm{d}\Psi(x).$$

The ζ_j are independent and have the same distribution Ψ, so

$$\begin{aligned}
\Psi^{\sim}(s) &= \int_0^{\infty} \left\{ \sum_{j=0}^{\infty} \frac{(\lambda y)^j}{j!} \int_0^{\infty} \mathrm{e}^{-sx}\,\mathrm{d}x\Psi_j(x-y) \right\} \mathrm{e}^{-\lambda y}\,\mathrm{d}B(y) \\
&= \int_0^{\infty} \left\{ \sum_{j=0}^{\infty} \frac{(\lambda y)^j}{j!} \left[\mathrm{e}^{-sy}(\Psi^{\sim}(s))^j\right] \mathrm{e}^{-\lambda y} \right\} \mathrm{d}B(y) \\
&= \sum_{j=0}^{\infty} \frac{(\lambda\Psi^{\sim}(s))^j}{j!} \int_0^{\infty} y^j \mathrm{e}^{-(\lambda+s)y}\,\mathrm{d}B(y) \\
&= \sum_{j=0}^{\infty} (-1)^j \frac{(\lambda\Psi^{\sim}(s))^j}{j!} \frac{\mathrm{d}^j}{\mathrm{d}s^j}(B^{\sim}(\lambda+s)),
\end{aligned}$$

which corresponds to the Taylor expansion of the function $B^{\sim}(\lambda + s - \lambda\Psi^{\sim}(s))$ in the neighborhood of $\lambda\Psi^{\sim}(s)$; consequently,

$$\Psi^{\sim}(s) = B^{\sim}(\lambda + s - \lambda\Psi^{\sim}(s)). \tag{8.11}$$

The next theorem deals with the solution of the functional Eq. (8.11).

Theorem 8.4. *Equation (8.11) has a unique solution at Re $s > 0$, $|\Psi^{\sim}(s)| \le 1$, and $\Psi^{\sim}(s)$ is real for all $s > 0$. Let p^* $(0 \le p^* \le 1)$ denote the least positive number for which $B^{\sim}(\lambda(1-p^*)) = p^*$. Then*

$$\Psi(\infty) = p^*.$$

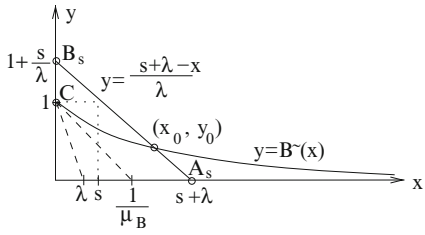
If $\rho = \lambda\tau \le 1$, then $p^* = 1$, and $\Psi(x)$ is a (nondegenerate) distribution function; if $\rho > 1$, then $p^* < 1$, and the busy period may be infinite with probability $1 - p^*$.

Comment 8.5. *Since $B^{\sim}(\lambda(1-p)), 0 \le p \le 1$, is a continuous and strictly monotonically function of p, and if $p = 1$, then $B^{\sim}(\lambda(1-p^*)) = p^*$, and p^* is well defined.*

Proof. First we show that Eq. (8.11) has a unique solution $\Psi^{\sim}(s)$ for which, at arbitrary $s > 0$, $|\Psi^{\sim}(s)| \le 1$. The proof uses the Rouché's theorem. □

Theorem 8.6 (Rouché). *Let $G(z)$ and $g(z)$ be regular functions on the domain D and continuous on its closure. If on the boundary of D $|g(z)| < |G(z)|$, then $G(z) + g(z)$ and $G(z)$ have the same number of roots in D (with multiplicities).*

Fig. 8.7 Solution of Eq. 8.12



Let s be an arbitrary complex number, $Re\ s > 0$, and consider the equation $z = B^{\sim}(s + \lambda - \lambda z)$. The right and left sides are analytical functions of z on a domain that contains the unit circle $|z| \leq 1$. If $|z| = 1$, then, because of $Re\ s > 0$ $Re(s + \lambda - \lambda z) > 0$,

$$\begin{aligned} |B^{\sim}(s + \lambda - \lambda z)| &\leq \int_0^{\infty} |e^{-(s+\lambda-\lambda z)x}| \, dB(x) \\ &= B^{\sim}(Re(s + \lambda - \lambda z)) \\ &< 1 = |z|. \end{aligned}$$

By Rouché's theorem, z and $(z - B^{\sim}(s + \lambda - \lambda z))$ have the same number of roots on the domain $|z| < 1$, i.e., one.

Now let us examine Eq. (8.11) on the positive real half-line. Let $s+\lambda-\lambda\Psi^{\sim}(s) = x$ and consider the solution of the equation

$$(s + \lambda - x)/\lambda = B^{\sim}(x) \tag{8.12}$$

at $s > 0$. We will see that in this case there exists a unique solution x_0 for which $s < x_0 < s + \lambda$. Figure 8.7 helps to understand the problem.

$B^{\sim}(x)$ is convex from below and continuous; consequently the root of Eq. (8.12) – and so $\Psi^{\sim}(s)$ also – for all $s > 0$ is uniquely determined on the whole $(0, \infty)$ half-line.

We remark that $B^{\sim}$ is a regular function, as is $\Psi^{\sim}$ [for all points $(0, \infty)$ there exists a neighborhood with radius $r > 0$, where it can be expanded]; consequently, it can be analytically continued for the right half-plane. (This means that Eq. (8.12) has an analytical inverse for x.)

If $s \to 0$, then $B_s \to C$, while the tangent of $B_s A_s$ remains $-\frac{1}{\lambda}$. At the same time

$$B^{\sim\prime}(0) = -\mu_B = -\int_0^{\infty} x \, dB(x),$$

so, by using the fact $B^{\sim}(x)$ is convex from below:

1. If $\rho > 1$ $(1/\mu_B < \lambda)$, then $B_s A_s$ (in the case $s \to 0+$) for a certain $x_* > 0$ intersects $B^{\sim}(x)$. Then $\lim_{s\to 0} x_0(s) = x_*$, $p^* = \Psi^{\sim}(0) = \frac{\lambda - x_*}{\lambda} < 1$ (in this case the busy period can be infinite with positive probability).
2. If $\rho \le 1$ $(1/\mu_B \ge \lambda)$, then the limit of $B_s A_s$ intersects $B^{\sim}(x)$ at the only point $x_0 = 0$ when $p^* = 1$. Consequently, $\Psi(\infty) = \Psi^{\sim}(0) = 1$.

Corollary 8.7. *Assume that $\rho < 1$. Differentiating Eq. (8.11) at $s = 0$ we obtain a linear equation for the mean value $\mathbf{E}\eta = -\Psi^{\sim}(0)$:*

$$\Psi^{\sim\prime}(0) = B^{\sim\prime}(0)(1 - \lambda\Psi^{\sim\prime}(0)).$$

From this we obtain the mean value of the busy period

$$\mathbf{E}\,(\eta) = -\Psi^{\sim\prime}(0) = \frac{\mu_B}{1-\rho}.$$

The other moments can be computed in a similar way, e.g.,

$$\mathbf{E}\left(\eta^2\right) = \frac{\mathbf{E}\left(Y^2\right)}{(1-\rho)^3}.$$

Results concerning the distribution function of a busy period's length may be derived from other considerations. With one customer have been served, n ones remain in the system; let H_n denote the time period till the moment when there will be $n-1$ customers. Furthermore, let Q_n denote the number of served customers for this period. The structure of this period (while we descend one level) coincides with the structure of the busy period and is independent of n.

Let the service time of a customer be $Y = y$; then (since we have a Poisson process with the parameter λ) the length of the busy period is

$$\{\eta | Y = y\} = \begin{cases} y & \text{with probability } \mathrm{e}^{-\lambda y}, \\ y + H_1 & \text{with probability } \lambda y \mathrm{e}^{-\lambda y}, \\ y + H_1 + H_2 & \text{with probability } \frac{(\lambda y)^2}{2!}\mathrm{e}^{-\lambda y}, \\ \dots \end{cases}$$

We have $\Psi^{\sim}(s) = \mathbf{E}\,(\mathrm{e}^{-s\eta}) = \Psi_1^{\sim}(s) = \Psi_2^{\sim}(s) = \dots$, so

$$\begin{aligned} \mathbf{E}\,(\mathrm{e}^{-s\eta} | Y = y) &= \sum_{i=0}^{\infty} \frac{(\lambda y)^i}{i!} \mathrm{e}^{-\lambda y} \mathrm{e}^{-sy} \left(\Psi^{\sim}(s)\right)^i \\ &= \mathrm{e}^{-\lambda y} \mathrm{e}^{-sy} \mathrm{e}^{\lambda y \Psi^{\sim}(s)} = \mathrm{e}^{-y(s+\lambda-\lambda\Psi^{\sim}(s))} \end{aligned}$$

and

$$\Psi^{\sim}(s) = \int_0^{\infty} \mathbf{E}\left(\mathrm{e}^{-s\eta}|y\right) \mathrm{d}B(y) = \int_0^{\infty} \mathrm{e}^{-y(s+\lambda-\lambda\Psi^{\sim}(s))} \mathrm{d}B(y)$$
$$= B^{\sim}(s + \lambda - \lambda\Psi^{\sim}(s)).$$

The number of customers served for a busy period is

$$\{Q|Y = y\} = \begin{cases} 1 & \text{with probability } \mathrm{e}^{-\lambda y}, \\ 1 + Q_1 & \text{with probability } \lambda y \mathrm{e}^{-\lambda y}, \\ 1 + Q_1 + Q_2 & \text{with probability } \frac{(\lambda y)^2}{2!} \mathrm{e}^{-\lambda y}, \\ \ldots & \end{cases}$$

Let $Q(z) = \mathbf{E}\left(z^Q\right) = Q_1(z) = Q_2(z) = \ldots,$

$$\mathbf{E}\left(z^Q|y\right) = z \sum_{i=0}^{\infty} \frac{(\lambda y)^i}{i!} \mathrm{e}^{-\lambda y} Q^i(z)$$
$$= z\mathrm{e}^{-\lambda y} \mathrm{e}^{\lambda y Q(z)} = z\mathrm{e}^{-y(\lambda-\lambda Q(z))},$$

and using this result we obtain

$$Q(z) = \int_0^{\infty} \mathbf{E}\left(z^Q|y\right) \mathrm{d}B(y) = \int_0^{\infty} z\mathrm{e}^{-y(\lambda-\lambda Q(z))} \mathrm{d}B(y)$$
$$= zB^{\sim}(\lambda(1 - Q(z))).$$

We have already computed the moments for the length of the busy period; the mean value of customers served for the busy period is

$$\mathbf{E}(Q) = \frac{1}{1-\rho}.$$

8.2.8 *Investigation on the Basis of the Regenerative Process*

The functioning of an $M/G/1$ system may be considered a regenerative process. Our aim now is to derive the Pollaczek–Khinchin transform equation on its basis.

We introduce the following notations:

$\mathbf{E}(\eta) = \frac{\mu_B}{1-\rho}$: mean value of busy period;

ω_i: mean value of time spent above ith level during a busy period;

η_i: mean value of time spent on ith level during a busy period.

Theorem 8.8. *Let us consider an M/G/1 queueing system with arrival rate λ and service time distribution $B(x)$. If the service time of a customer has a finite mean μ_B, $\lambda\mu_B < 1$, then there exists an equilibrium distribution in the system. These probabilities are determined by the fractions $p_i = \eta_i/\mathbf{E}(\eta)$ $(i = 0, 1, \ldots)$, where $\mathbf{E}(\eta)$ is the mean value of the busy period and η_i is the mean value of time spent on the i th level during a busy period.*

Proof. The proof of the theorem is a direct consequence of Theorem 4.40 (see also [94, Theorems 1.3.2 and 1.3.3]). The mean values appearing in the theorem are given by the following lemma. □

Lemma 8.9. *In the M/G/1 system*

$$\eta_0 = \mu_B, \qquad \eta_1 = \frac{1 - a_0}{a_0}\eta_0, \qquad \eta_2 = \frac{1 - a_0 - a_1}{a_0}(\eta_0 + \eta_1),$$

and η_k $(k \geq 3)$ satisfy the recurrence relation

$$\eta_k = \sum_{i=1}^{k-2} \frac{1 - a_0 - a_1 - \ldots - a_i}{a_0}\eta_{k-i} + \frac{1 - a_0 - a_1 - \ldots - a_{k-1}}{a_0}(\eta_0 + \eta_1).$$

Proof. Let j customers be present in the system, with one of them being served. An actual customer having been served, the number of present customers does not change with probability a_1. The number of present customers changes with probability $1 - a_1$, we come to another level, with probability $\frac{a_0}{1-a_1}$ to $j - 1$, and with probability $\frac{1-a_0-a_1}{1-a_1}$ to a level above j.

Let us consider a busy period and intervals in it where one or more customers stay in the system. When we used the embedded Markov chain technique the states of the system were identified by the number of customers remaining in the system after a customer had been served. Now it will be better to regard the number of customers at the beginning of service. The difference will be clear from the following reasoning. If one considers service periods of customers when at the starting moment there are no other customers, then each of the periods corresponds to state 1, excluding two cases. The first case is when we jump to a level above the first one, then the service of the last customer from the viewpoint of states corresponds to the new level (from the viewpoint of the number of present customers to the first level). But the whole duration does not change because coming from the second level to the first the inverse situation takes place. The situation will be similar for all levels above the first. The second case is the service of the last customer in the busy period; it corresponds to a zero state (after this customer is served there will be no customers in the system), so it must be excluded from the number of customers served on the first level.

We determine the mean value of a period during which there is only one customer in the system. For the service of a customer a new one enters with probability a_1, so this state is continued with probability a_1 and terminated with probability $1 - a_1$ (there is no entry or more than one customer appears). For such a period with

probabilities $1 - a_1$ is served one, $a_1(1 - a_1)$ are served two,..., with probability $a_1^{k-1}(1-a_1)$ are served k customers. The mean value of the number of customers served is

$$\sum_{k=1}^{\infty} k a_1^{k-1}(1 - a_1) = \frac{1}{1 - a_1}.$$

Now let us determine the mean value of a period above the first level (in this case we will have the aforementioned deviation concerning the states and number of customers, but finally we obtain the correct value). Assume that at the beginning of this period there are k customers in the system [while the last customer on the first level is being served with probability $1 - a_0 - a_1$, at least two customers arrived, with probabilities $\frac{a_k}{1-a_0-a_1}$ $(k = 2, 3, \ldots)$ we will have k ones]. To return to the first level, we have to complete $k - 1$ present and all further customers entered for their services. (The structure of a period during which one customer is served with the generated ones coincides with that of the busy period.) The mean value of a busy period is $\frac{\mu_B}{1-\rho}$; consequently, the length of such an interval is

$$\sum_{k=2}^{\infty} \frac{a_k}{1 - a_0 - a_1}(k - 1)\frac{\mu_B}{1 - \rho} = \frac{\mu_B}{(1 - \rho)(1 - a_0 - a_1)}\Big(\rho - a_1 - (1 - a_0 - a_1)\Big)$$

$$= \frac{\rho - 1 + a_0}{(1 - \rho)(1 - a_0 - a_1)}\mu_B,$$

where we used the equalities

$$\rho = \sum_{k=1}^{\infty} k a_k \qquad \text{és} \qquad \sum_{k=0}^{\infty} a_k = 1.$$

For the busy period we have a certain number of intervals with one present customer; such an interval is finished either without entry (meaning the end of the busy period) or with the entry of more than one customer. With probabilities

$$\frac{a_0}{1 - a_1}, \quad \frac{1 - a_0 - a_1}{1 - a_1}\frac{a_0}{1 - a_1}, \ldots, \frac{(1 - a_0 - a_1)^k}{(1 - a_1)^k}\frac{a_0}{1 - a_1}, \ldots$$

we will have 0,1,...,k,... intervals with the presence of more than one customer. Thus the mean values of intervals of two types are

$$\sum_{k=1}^{\infty} k \frac{(1 - a_0 - a_1)^{k-1}}{(1 - a_1)^{k-1}}\frac{a_0}{1 - a_1}\frac{\mu_B}{1 - a_1} = \frac{\mu_B}{a_0},$$

$$\sum_{k=1}^{\infty} k \frac{(1 - a_0 - a_1)^k}{(1 - a_1)^k}\frac{a_0}{1 - a_1}\frac{\rho - 1 + a_0}{(1 - \rho)(1 - a_0 - a_1)}\mu_B = \frac{\rho - 1 + a_0}{a_0(1 - \rho)}\mu_B.$$

The sum of these two values obviously gives the busy period's mean value:

$$\frac{\mu_B}{a_0} + \frac{\rho - 1 + a_0}{a_0(1-\rho)}\mu_B = \frac{\mu_B}{1-\rho}.$$

We derive the mean value of time spent above the kth level for a busy period. First let us consider the case of the second level. We have two possibilities:

1. From the first level we arrive at the second one.
2. From the first level we arrive at least at the third one.

If the period under consideration begins at the second level, then we are in the same situation as in the case of the first level. We serve a certain number of customers on the second level, then we go either to the first level or above the second one. In the first case, intervals on and above the second level will change, and spending on average ω_1 time above it we come to the first one. In the second case the period above the second level begins with a jump from the first level immediately to a level above the second, and the mean value of time to return to the second one is equal to

$$\sum_{k=3}^{\infty} \frac{a_k}{1-a_0-a_1-a_2}(k-2)\frac{\mu_B}{1-\rho} = \frac{\rho - 2 + 2a_0 + a_1}{(1-\rho)(1-a_0-a_1-a_2)}\mu_B = \varepsilon_2.$$

Now we are in the same situation as in the previous case, i.e., we spend above the second level ω_1 time. The probabilities of the two cases are

$$\frac{a_2}{1-a_0-a_1} \quad \text{and} \quad \frac{1-a_0-a_1-a_2}{1-a_0-a_1},$$

so for a period beginning and ending on the first level we spend above the second level on average

$$\frac{a_2}{1-a_0-a_1}\omega_1 + \frac{1-a_0-a_1-a_2}{1-a_0-a_1}(\omega_1 + \varepsilon_2) = \omega_1 + \varepsilon_2',$$

where

$$\varepsilon_2' = \frac{\rho - 2 + 2a_0 + a_1}{(1-\rho)(1-a_0-a_1)}\mu_B.$$

For a busy period we have i such intervals with probability $\frac{(1-a_0-a_1)^i}{(1-a_1)^i}\frac{a_0}{1-a_1}$; consequently,

$$\omega_2 = \sum_{i=1}^{\infty} i \frac{(1-a_0-a_1)^i}{(1-a_1)^i}\frac{a_0}{1-a_1}(\omega_1 + \varepsilon_2') = \frac{1-a_0-a_1}{a_0}\omega_1 + \frac{1-a_0-a_1-a_2}{a_0}\varepsilon_2.$$

Let us assume that our formula is valid for the $k-1$st level and compute ω_k. We consider again an interval starting and ending on the first level. ω_k may be written in the form

$$\begin{aligned}\omega_k : \quad & \omega_{k-1} \\ & \omega_{k-2} + \omega_{k-1} \\ & \ldots\ldots\ldots\ldots\ldots \\ & \omega_{k-i} + \omega_{k-i+1} + \ldots + \omega_{k-2} + \omega_{k-1} \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & \omega_1 + \omega_2 + \ldots + \omega_{k-2} + \omega_{k-1} + \varepsilon_k\end{aligned}$$

From the first level we can come to the second, third, ..., $k-1$st, kth, or one above the kth level. The first possibility is the second level. We are in the same situation as in the case with the time spent above the $k-1$st level from the viewpoint of the first one; the mean value is ω_{k-1}. In the case of the third level, first we have an interval starting with three and ending with two customers. This corresponds to the situation where one considers the time above the $k-2$nd level from the viewpoint of the first one; the mean value is ω_{k-2}. Now we are in the previous situation (two customers), and the mean value of the remaining part is ω_{k-1}. So under the condition that from the first level we come at once to the third level, the desired mean value is equal to $\omega_{k-2} + \omega_{k-1}$.

Let us consider the last case, which takes place when from the first level we jump to a level above k. The mean value of time to reach the kth level is

$$\begin{aligned}&\sum_{i=k+1}^{\infty} \frac{a_i}{1-a_0-a_1-\ldots-a_k}(i-k)\frac{\mu_B}{1-\rho} \\ &\quad = \frac{\rho - k + ka_0 + (k-1)a_1 + \ldots + 2a_{k-2} + a_{k-1}}{(1-\rho)(1-a_0-a_1-\ldots-a_k)}\mu_B = \varepsilon_k.\end{aligned}$$

After this period we will be at the kth level, and according to our previous reasoning, spending on average ω_1 time above the kth level we come to the $k-1$st level, spending ω_2 above the kth level we come to the $k-2$nd, ..., and finally starting from the second level and spending ω_{k-1} above the kth one we reach the first level. So, in the last case, the desired mean value is $\omega_1 + \omega_2 + \ldots + \omega_{k-1} + \varepsilon_k$. The probability of the first case is $\frac{a_2}{1-a_0-a_1}$, the probability of the second one is $\frac{a_3}{1-a_0-a_1}$, ..., and the probability of the last case is $\frac{1-a_0-a_1-\ldots-a_k}{1-a_0-a_1}$. Multiplying the conditional mean values by the corresponding probabilities we obtain

$$\omega_{k-1} + \frac{1-a_0-a_1-a_2}{1-a_0-a_1}\omega_{k-2} + \ldots +$$

$$+\frac{1-a_0-a_1-\ldots-a_{k-1}}{1-a_0-a_1}\omega_1+\frac{1-a_0-a_1-\ldots-a_k}{1-a_0-a_1}\varepsilon_k.$$

For the busy period we will stay above the first level i times with probability $\frac{(1-a_0-a_1)^i}{(1-a_1)^i}\frac{a_0}{1-a_1}$, so the mean value of time spent above the kth level for a busy period equals

$$\begin{aligned}\omega_k &= \sum_{i=1}^{\infty} i\frac{(1-a_0-a_1)^i}{(1-a_1)^i}\frac{a_0}{1-a_1}\left(\omega_{k-1}+\frac{1-a_0-a_1-a_2}{1-a_0-a_1}\omega_{k-2}+\right.\\ &\left.+\ldots+\frac{1-a_0-a_1-\ldots-a_{k-1}}{1-a_0-a_1}\omega_1+\frac{1-a_0-a_1-\ldots-a_k}{1-a_0-a_1}\varepsilon_k\right)\\ &= \sum_{i=1}^{k-1}\frac{1-a_0-a_1-\ldots-a_i}{a_0}\omega_{k-i}+\frac{1-a_0-a_1-\ldots-a_k}{a_0}\varepsilon_k.\end{aligned}$$

In a similar way

$$\omega_{k-1}=\sum_{i=1}^{k-2}\frac{1-a_0-a_1-\ldots-a_i}{a_0}\omega_{k-i-1}+\frac{1-a_0-a_1-\ldots-a_{k-1}}{a_0}\varepsilon_{k-1}.$$

The mean value of time spent on the kth level for the busy period is

$$\begin{aligned}\eta_k &= \omega_{k-1}-\omega_k=\frac{1-a_0-a_1}{a_0}(\omega_{k-2}-\omega_{k-1})+\frac{1-a_0-a_1-a_2}{a_0}(\omega_{k-3}-\omega_{k-2})\\ &+\ldots+\frac{1-a_0-\ldots-a_{k-2}}{a_0}(\omega_1-\omega_2)-\frac{1-a_0-\ldots-a_{k-1}}{a_0}\omega_1\\ &+\frac{\rho-(k-1)+(k-1)a_0+(k-2)a_1+\ldots+2a_{k-3}+a_{k-2}}{a_0(1-\rho)}\mu_B\\ &-\frac{\rho-k+ka_0+(k-1)a_1+(k-2)a_2+\ldots+2a_{k-2}+a_{k-1}}{a_0(1-\rho)}\mu_B\\ &=\frac{1-a_0-a_1}{a_0}\eta_{k-1}+\frac{1-a_0-a_1-a_2}{a_0}\eta_{k-2}+\ldots+\frac{1-a_0-\ldots-a_{k-2}}{a_0}\eta_2\\ &-\frac{1-a_0-a_1-\ldots-a_{k-1}}{a_0}\omega_1+\frac{1-a_0-a_1-\ldots-a_{k-1}}{a_0}\frac{\mu_B}{1-\rho}\\ &=\sum_{i=1}^{k-2}\frac{1-a_0-\ldots-a_i}{a_0}\eta_{k-i}+\frac{1-a_0-\ldots-a_{k-1}}{a_0}(\eta_0+\eta_1).\end{aligned}$$

The lemma is proved. □

We show that from these mean values one can derive the Pollaczek–Khintchine transform equation. Let us multiply the expression for η_i in the lemma by z^i and sum them up from the row η_2, excluding the last term (containing η_0). Then

$$\begin{aligned}
&\frac{1-a_0-a_1}{a_0}z(\eta_1 z+\eta_2 z^2+\ldots)+\frac{1-a_0-a_1-a_2}{a_0}z^2(\eta_1 z+\eta_2 z^2+\ldots)\\
&\quad+\frac{1-a_0-a_1-a_2-a_3}{a_0}z^3(\eta_1 z+\eta_2 z^2+\ldots)+\ldots\\
&=\left(\sum_{i=1}^{\infty}\eta_i z^i\right)\left(\frac{1-a_0-a_1}{a_0}z+\frac{1-a_0-a_1-a_2}{a_0}z^2\right.\\
&\quad\left.+\frac{1-a_0-a_1-a_2-a_3}{a_0}z^3+\ldots\right)\\
&=\left(\sum_{i=1}^{\infty}\eta_i z^i\right)\frac{1}{a_0}\left(\frac{z}{1-z}-\frac{a_0 z}{1-z}-\frac{a_1 z}{1-z}-\frac{a_2 z^2}{1-z}-\frac{a_3 z^3}{1-z}-\ldots\right)\\
&=\left(\sum_{i=1}^{\infty}\eta_i z^i\right)\frac{1}{a_0(1-z)}\Big(z(1-a_0)-(A(z)-a_0)\Big)\\
&=(\overline{P}(z)-\eta_0)\frac{1}{a_0(1-z)}\Big(z(1-a_0)-(A(z)-a_0)\Big),
\end{aligned}\tag{8.13}$$

where $\overline{P}(z)=\sum\limits_{i=0}^{\infty}\eta_i z^i$. For the terms containing η_0

$$\eta_0 z\sum_{i=1}^{\infty}\frac{1-a_0-\ldots-a_i}{a_0}z^i=\eta_0 z\frac{1}{a_0(1-z)}\Big(z(1-a_0)-(A(z)-a_0)\Big).\tag{8.14}$$

Summing up Eqs. (8.13) and (8.14), the formula for η_0 and η_1 multiplied by z, we obtain

$$\begin{aligned}
\overline{P}(z)&=(\overline{P}(z)-\eta_0)\frac{1}{a_0(1-z)}\Big(z(1-a_0)-(A(z)-a_0)\Big)\\
&\quad+\eta_0 z\frac{1}{a_0(1-z)}\Big(z(1-a_0)-(A(z)-a_0)\Big)+\eta_0+\frac{1-a_0}{a_0}\eta_0 z,
\end{aligned}$$

whence

$$\overline{P}(z)=\frac{(1-z)A(z)}{A(z)-z}\eta_0.$$

Dividing this by the mean value of the busy period $\frac{\mu_B}{1-\rho}$ and taking into account $\eta_0=\mu_B$, we finally get the well-known formula

$$P(z) = \frac{(1-\rho)(1-z)A(z)}{A(z)-z}.$$

For details see [60].

8.2.9 *Proof of Relation (D) (Khinchin (1932))*

Using the embedded Markov chain technique we found the ergodic distribution of the number of customers at moments just after having served individual customers. Actually, our objective is to show that this stationary distribution holds not only for the service completion moments but also for the continuous-time $L(t)$ process. We prove the equality

$$\lim_{t\to\infty} \mathbf{P}\,(L(t) = k) = \pi_k, \quad k \geq 0,$$

i.e., the same formula (8.10) is valid for the generating function of the limiting distribution of $L(t)$.

Let ξ_i, $\eta_i, i = 1, 2, \ldots$, denote the successive empty/busy periods, which are independent and separately identically distributed. The empty periods have an exponential distribution with the parameter λ, and the corresponding mean value is $\mathbf{E}\,(\xi_i) = 1/\lambda$.

The sequence $(\xi_i + \eta_i), i = 1, 2, \ldots$, is a renewal process; at the same time $(\xi_i + \eta)$ are the regenerative cycles of process $L(t)$.

Earlier we derived a functional equation for the Laplace–Stieltjes transform of a busy period's distribution function; from this for the mean value we obtained $\mathbf{E}\,(\eta_i) = \frac{\mu_B}{1-\rho}$, so the mean value of a regenerative cycle is

$$\kappa = \mathbf{E}\,(\xi_1 + \eta_1) = \frac{1}{\lambda} + \frac{\mu_B}{\lambda(1-\rho)} = \frac{1}{\lambda(1-\rho)}.$$

$L(t)$ is a regenerative process, and from the limit theorem for the regenerative processes

$$\lim_{t\to\infty} \mathbf{P}\,(L(t) > 0) = \kappa^{-1}\mathbf{E}\left(\int_0^{T_1} \mathcal{I}_{\{L(t)>0\}}\mathrm{d}t\right) = \kappa^{-1}\mathbf{E}\,(\eta_1)$$

$$= \frac{\mu_B}{1-\rho}\lambda(1-\rho) = \rho,$$

so, using the earlier proved relation $\pi_0 = 1 - \rho$, we get

$$p_0 = \lim_{t\to\infty} \mathbf{P}\,(L(t) = 0) = 1 - \rho = \pi_0.$$

By the repeated use of the theorem for regenerative processes one can show the existence of the limits

$$p_n = \lim_{t\to\infty} \mathbf{P}\left(L(t) = n\right), \; n \geq 1,$$

but finding them in explicit appears to be a difficult problem.

First we find the limit distributions of backward and forward service times, δ_t and γ_t, of a customer being served at moment t as $t \to \infty$

$$F(y) = \lim_{t\to\infty} \mathbf{P}\left(\delta_t < y\right), \quad \text{ill.} \quad G(y) = \lim_{t\to\infty} \mathbf{P}\left(\gamma_t < y\right).$$

Let $y > 0$. Using the aforementioned theorem for regenerative processes

$$\begin{aligned} F(y) &= \lambda(1-\rho)\mathbf{E}\left(\int_0^{T_1} \mathcal{I}_{\{0\leq\delta_s<y\}}\mathrm{d}s\right) \\ &= \lambda(1-\rho)\mathbf{E}\left(\xi_1 + \int_{\xi_1}^{\xi_1+\eta_1} \mathcal{I}_{\{0<\delta_s<y\}}\mathrm{d}s\right) \\ &= 1-\rho+\lambda(1-\rho)\mathbf{E}\left(\sum_{j=1}^{K} \int_{\xi_1+Y_1+\ldots+Y_{j-1}}^{\xi_1+Y_1+\ldots+Y_j} \mathcal{I}_{\{0<\delta_s<y\}}\mathrm{d}s\right) \\ &= 1-\rho+\lambda(1-\rho)\mathbf{E}\left(\sum_{j=1}^{K} \min(y, Y_j)\right), \end{aligned}$$

where K is a random variable, the number of customers served in the first regenerative cycle T_1, and it coincides with the number of customers served in the first busy period of the system. Integrating by parts, we get

$$\begin{aligned} \mathbf{E}\left(\min(y, Y_j)\right) &= \int_0^{\infty} \min(y, x)\mathrm{d}B(x) \\ &= \int_0^{y} x\mathrm{d}B(x) + \int_y^{\infty} y\mathrm{d}B(x) \\ &= -\int_0^{y} x\mathrm{d}(1-B(x)) + y(1-B(y)) \end{aligned}$$

$$= -y(1 - B(y)) + \int_0^y (1 - B(x))\mathrm{d}x + y(1 - B(y))$$

$$= \int_0^y (1 - B(x))\mathrm{d}x.$$

On the other hand, since K is a Markov moment for the sequence Y_j, $j = 1, 2, \ldots$, using the Wald identity we obtain

$$\mathbf{E}\left(\sum_{j=1}^{K} \min(y, Y_j)\right) = \mathbf{E}\,(K) \cdot \mathbf{E}\left(\min(y, Y_j\right) = \mathbf{E}\,(K) \int_0^y (1 - B(x))\mathrm{d}x.$$

Similarly,

$$\mathbf{E}\,(\eta_1) = \mathbf{E}\left(\sum_{j=1}^{K} Y_j\right) = \mathbf{E}\,(K) \cdot \mathbf{E}\left(Y_j\right) = \mathbf{E}\,(K) \cdot \mu_B = \frac{\mu_B}{1 - \rho},$$

whence $\mathbf{E}\,(K) = \frac{1}{1-\rho}$, and on the basis of these expressions we get the limiting distribution of δ_t:

$$F(y) = 1 - \rho + \lambda \int_0^y (1 - B(x))\mathrm{d}x.$$

We mention that $F(0+) = 1 - \rho$, $F(+\infty) = 1 - \rho + \lambda\mu_B = 1$.

The limiting distribution of γ_t may be obtained in a similar way.

$$1 - G(y) = \mu^{-1}\mathbf{E}\left(\int_0^{T_1} \mathcal{I}_{\{y<\gamma_s\}}\mathrm{d}s\right)$$

$$= \lambda(1 - \rho)\mathbf{E}\left(\int_{\xi_1}^{\xi_1+\eta_1} \mathcal{I}_{\{y<\gamma_s\}}\mathrm{d}s\right)$$

$$= \lambda(1 - \rho)\mathbf{E}\left(\sum_{j=1}^{K} \int_{\xi_1+Y_1+\ldots+Y_{j-1}}^{\xi_1+Y_1+\ldots+Y_j} \mathcal{I}_{\{y<\gamma_s\}}\mathrm{d}s\right)$$

$$= \lambda(1 - \rho)\mathbf{E}\left(\sum_{j=1}^{K} (Y_j - y)^+\right)$$

$$= \lambda(1-\rho)\mathbf{E}(K)\cdot\mathbf{E}\left((Y_j - y)^+\right)$$
$$= \lambda\int_y^\infty (x-y)\mathrm{d}B(x)$$
$$= \lambda\int_y^\infty (1-B(x))\mathrm{d}x,$$

and so

$$G(y) = 1 - \lambda\int_y^\infty (1-B(x))\mathrm{d}x.$$

From this it follows that $G(0+) = 1-\lambda\mu_B = 1-\rho = \pi_0$ and $G(+\infty) = 1$.

Now, let us prove that for the stationary distribution $p_n = \lim\limits_{t\to\infty}\mathbf{P}(L(t)=n)$ holds $p_n = \pi_n, n \geq 1$ (for the case $n = 0$ we have proved the equality). We will follow the reasoning by Khinchin [53].

In the stationary case the event that at the completion of a service n customers remain in the system has probability π_n, at an arbitrary moment p_n, and the remaining part of the service has distribution $G(x)$. For a small service time δ_t j customers enter the system with probability

$$a_j = \lambda\int_0^\infty \frac{(\lambda x)^j}{j!}\mathrm{e}^{-\lambda x}(1-B(x))\mathrm{d}x, \quad j \geq 0.$$

If the number of customers in the system at moment t is $L(t) = n > 0$, then $L(t-\delta_t)$ gives the possible number of customers there at the previous departure moment ($k = 1,\ldots,n$), or a new customer entered the empty system at $t-\delta_t$. Using the formula of total probability in the case $n > 0$ we obtain

$$p_n = \mathbf{P}(L(t)=n) = \sum_{k=0}^{n}\mathbf{P}(L(t)=n \mid L(t-\delta_t)=k)\,\mathbf{P}(L(t-\delta_t)=k)$$
$$= \pi_n a_0 + \pi_{n-1}a_1 + \ldots + \pi_1 a_{n-1} + \pi_0 a_{n-1}, \quad n > 0.$$

Let

$$h_k(x) = \frac{(\lambda x)^k}{k!}\mathrm{e}^{-\lambda x}, \quad k \geq 0,$$

and in the case $k \geq 1$,

$$g_{k-1}(x) = \pi_k h_0(x) + \pi_{k-1}h_1(x) + \ldots + \pi_1 h_{k-1}(x) + \pi_0 h_{k-1}(x). \quad (8.15)$$

Then

$$p_n = \lambda \int_0^\infty (1 - B(x)) g_{n-1}(x) \mathrm{d}x. \tag{8.16}$$

One can directly check that the functions $h_k(x)$ satisfy the difference-differential equation

$$h_k'(x) = \lambda[h_{k-1}(x) - h_k(x)],$$

so for the functions $g(x)$ we have

$$g_k'(x) = \lambda[g_{k-1}(x) - g_k(x)], \quad k \geq 1.$$

Since $h_0(0) = 1$ and $h_k(0) = 0$ if $k \geq 1$, then $g_n(0) = \pi_{n+1}$. On the other hand, the recurrence relation (8.8) is valid for π_k, and taking into account Eqs. (8.15) and (8.16) for $k \geq 0$ we have

$$\pi_k = \int_0^\infty h_k(x) \mathrm{d}B(x) = h_k(0) + \int_0^\infty h_k'(x)(1 - B(x)) \mathrm{d}x$$

$$= \pi_{k+1} + \lambda \int_0^\infty [h_{k-1}(x) - h_k(x)](1 - B(x)) \mathrm{d}x = \pi_{k+1} + p_k - p_{k+1}.$$

Using this result and the equality proved earlier, $p_0 = \pi_0$, we obtain

$$\pi_k - p_k = \pi_{k+1} - p_{k+1} = \text{const},$$

i.e.,

$$\pi_k = p_k, \quad k \geq 0.$$

8.3 Limit Distribution of Virtual Waiting Time

Let $\gamma(t), t \geq 0$, be the *virtual* (or *possible*) *waiting time* at moment t (customers that entered up to moment t leave the system up to $t + \gamma(t)$). If the system is empty at moment t, then $\gamma(t) = 0$. The notion of virtual waiting time was introduced and investigated by Takács [89].

Assume that $\gamma(0) = x_0$. Our aim is to determine the distribution function of $\gamma(t)$. Let $t_1, t_2, \ldots$ $(t_0 = 0)$ denote the arrival process. According to our previous assumptions $t_j - t_{j-1}$, $j \geq 1$, are independent exponentially distributed random variables with parameter λ. In this case if $t_n < t < t_{n+1}$, then

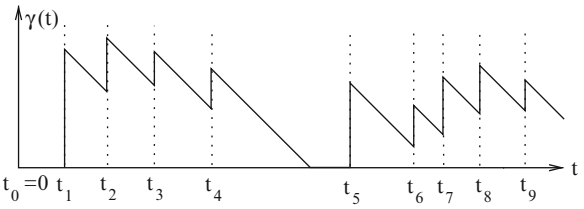


Fig. 8.8 Evolution of remaining service time

$$\gamma(t) = \begin{cases} 0 & \text{if } \gamma(t_n) < t - t_n, \\ \gamma(t_n) - (t - t_n) & \text{if } \gamma(t_n) \geq t - t_n. \end{cases}$$

If $t = t_n$, then $\gamma(t_n + 0) = \gamma(t_n - 0) + Y_n$, where Y_n is the service time of a customer entering at moment t_n (Fig. 8.8).

Theorem 8.10. *$\gamma(t)$, $t \geq 0$, is a Markov process.*

Proof. The arrival process $N(t)$ has independent increments (it is a Poisson process), so the number of customers and the associated service times of customers appearing by $[t, t + s)$ are independent of the number and service times of those appearing before t. $\gamma(t + s)$ is determined by the value of $\gamma(t)$ and the customers entering after t; they do not depend on those entering before t (the service times are independent of one another and the arrival process), so our statement is valid. □

Example 8.11. At moments $t_1, t_2, \ldots$ random amounts of water $Y_1, Y_2, \ldots$ flow to a reservoir. The outflow is uniform. In this case $\gamma(t)$ gives the actual amount of water in the reservoir.

8.3.1 Takács' Integrodifferential Equation

In previous sections we considered the number of customers in an $M/G/1$ system at special points. Here we intend to give a full description of its behavior. For the sake of simplicity let us denote the distribution function $F(t, x; x_0) = \mathbf{P}(\gamma(t) \leq x \mid \gamma(0) = x_0)$ by $F(t, x)$; assume that there exist the continuous partial derivatives by t and x on the set $t > 0,\ x \geq 0$, and

$$\lim_{t \to 0+} F(t, x) = F(0, x) = I(x \geq x_0).$$

Theorem 8.12 (Takács [90]). *Under these conditions the distribution function $F(t, x)$ satisfies the integrodifferential equation*

$$\frac{\partial F(t, x)}{\partial t} = \frac{\partial F(t, x)}{\partial x} - \lambda F(t, x) + \lambda \int_0^x B(x - y)\, \mathrm{d}_y F(t, y). \tag{8.17}$$

Proof. $\{\gamma(t+\Delta) < x\}$ is the union of three disjoint events (we take into account that the arrival process is Poisson and independent of service times):

- $\gamma(t+\Delta) < x$ and for $(t, t+\Delta)$ no customer enters the system. The probability of this event is $(1-\lambda\Delta)F(t, x+\Delta) + o(\Delta)$.
- At moment t, $0 \le \gamma(t) < x$, and for $(t, t+\Delta)$ one new customer enters [the corresponding probability is $\lambda\Delta$ independently of $\gamma(t)$ and the service time], and the customer's service time is $Y < x - \gamma(t)$. The probability of this event [Y and $\gamma(t)$ are independent, and the distribution of $Y + \gamma(t)$ is $B * F$: $\lambda\Delta \mathbf{P}(Y + \gamma(t) < x) + o(\Delta)$] is

$$\lambda\Delta \int_0^x B(x-y)\, \mathrm{d}_y F(t, y) + o(\Delta).$$

- $0 \le \gamma(t) < x$, and for $(t, t+\Delta)$ more than one customer enters the system, and its probability is $o(\Delta)$.

Then

$$F(t+\Delta, x) = (1-\lambda\Delta)F(t, x+\Delta) + \lambda\Delta \int_0^x B(x-y)\, \mathrm{d}_y F(t, y) + o(\Delta),$$

which can be rewritten as

$$\frac{1}{\Delta}\Big(F(t+\Delta, x) - F(t, x)\Big)$$

$$= \frac{1}{\Delta}\Big(F(t, x+\Delta) - F(t, x)\Big) - \lambda F(t, x+\Delta) + \lambda \int_0^x B(x-y)\, \mathrm{d}_y F(t, y) + o(1).$$

If $\Delta \to 0$, then we obtain Eq. (8.17). □

Takács derived this theorem in the case of an inhomogeneous Poisson arrival process with intensity $\lambda(t)$. He proved that this integrodifferential equation holds for all $t, x \ge 0$ for which $\frac{\partial}{\partial x} F(t, x)$ exists.

With the help of the previous theorem we prove the following one giving an integrodifferential equation for the stationary distribution.

Theorem 8.13. *If* $\mu_B = \int_0^\infty x\, \mathrm{d}B(x) < \infty$, $\rho = \lambda\tau < 1$, *then there exists*

$$\lim_{t\to\infty} F(t, x) = F(x),$$

and it is independent of the initial distribution $F(0, x)$. *It satisfies the equation*

$$F'(x) = \lambda F(x) - \lambda \int_0^x B(x-y)\,\mathrm{d}F(y),\ x > 0, \tag{8.18}$$

and $F(0+) = 1 - \rho$.

Proof. The proof is based on results for regenerative processes. We can use the fact that the distribution of cycles from a given index is absolute continuous, or the process $\gamma(t),\ t \geq 0$, is right continuous and has a limit from left. From both conditions it follows that the process has a limit distribution and can be written in the form given previously in Theorem 8.12.

Let $0 < \tau_1 < \tau_2 < \cdots$ be successive moments when free periods begin. Then $\{\gamma(t),\ t \geq 0\}$ is a regenerative process with regeneration points $\tau_k,\ k = 1, 2, \ldots$; the intervals $Z_k = \tau_k - \tau_{k-1},\ k = 1, 2, \ldots$ $(\tau_0 = 0)$, whose lengths are the sums of free and busy periods (perhaps excluding Z_1), are a (delayed) renewal process.

Let

$$G_1(x) = \mathbf{P}\,(Z_1 \leq x)\,, \quad G(x) = \mathbf{P}\,(Z_k \leq x)\,,\ k \geq 2,$$

and

$$G^{(n+1)}(x) = \mathbf{P}\,(\tau_k \leq x) = \int_0^x G^{(n)}(x-y)\,\mathrm{d}G(x),\ n \geq 1.$$

Since the free and busy periods are i.i.d. random variables (the free periods have an exponential distribution with parameter λ), the distribution function G, and thus $G^{(n)},\ n \geq 2$, is absolutely continuous (this is a sufficient condition for the existence of a limit distribution).

The mean value of a regenerative cycle is

$$\int_0^\infty x\,\mathrm{d}G(x) = \frac{1}{\lambda} + \frac{\mu_B}{1-\rho} < \infty,$$

and for arbitrary x there exists the limit distribution

$$\begin{aligned} F(x) &= \lim_{t\to\infty} F(t,x) \\ &= \lim_{t\to\infty} \mathbf{E}\left(\mathcal{I}_{\{\gamma(t)\leq x\}}\right) \\ &= \frac{1}{\kappa}\mathbf{E}\left(\int_0^T \mathcal{I}_{\{\gamma(s)\leq x\}}\,\mathrm{d}s\right). \end{aligned}$$

If in Eq. (8.17) $t \to \infty$, then we obtain Eq. (8.18). □

If we take the initial distribution $F(0,x) = F(x)$, then the distribution function $F(t,x) = F(x)$ satisfies Eq. (8.17). It is clear that

$$\begin{aligned} F(0+) &= \lim_{t\to\infty} F(t,0+) \\ &= \lim_{t\to\infty} \mathbf{P}\,(\gamma(t)=0) \\ &= \lim_{t\to\infty} \mathbf{P}\,(L(t)=0) \\ &= 1-\rho. \end{aligned}$$

One can see that [39], if $\rho \geq 1$, then

$$\lim_{t\to\infty} F(t,x) = 0, \quad x \in \mathbb{R}.$$

Equation (8.18) may be solved by means of the Laplace–Stieltjes transforms. Let

$$\begin{aligned} F^{\sim}(s) &= \int_0^\infty \mathrm{e}^{-sx}\,\mathrm{d}F(x) \\ &= 1-\rho+\int_0^\infty \mathrm{e}^{-sx}F'(x)\,\mathrm{d}x, \end{aligned}$$

where substituting F' from Eq. (8.18) yields

$$\begin{aligned} F^{\sim}(s) &= 1-\rho+\frac{\lambda}{s}F^{\sim}(s)-\frac{\lambda}{s}F^{\sim}(s)B^{\sim}(s) \\ &= 1-\rho+\frac{\lambda}{s}F^{\sim}(s)(1-B^{\sim}(s)), \end{aligned}$$

whence

$$F^{\sim}(s) = \frac{1-\rho}{1-\frac{\lambda}{s}(1-B^{\sim}(s))}. \tag{8.19}$$

This expression is called the Pollaczek–Khinchin formula for the waiting time. The inversion of the Laplace–Stieltjes transform gives the probability of an event in a stationary regime; the waiting time is less than x (see, e.g., [70]).

Example 8.14. Let us consider the case where the distribution function $B(x)$ is exponential with parameter μ, i.e.,

$$B(x) = 1-\mathrm{e}^{-\mu x}, \quad x \geq 0.$$

Then $B^{\sim}(s) = \frac{\mu}{s+\mu}$, according to the Pollaczek–Khinchin formula (8.19), for the Laplace–Stieltjes transform $F^{\sim}$ of the distribution function F we obtain

$$F^{\sim}(s) = 1-\rho+\lambda\frac{\mu-\lambda}{s+\mu-\lambda}.$$

The inversion of the Laplace–Stieltjes transform gives

$$F(x) = 1 - \rho + \rho(1 - \mathrm{e}^{-(\mu-\lambda)x}),$$

when

$$F(0+) = \lim_{x \to 0+} F(x) = 1 - \rho.$$

8.4 G/M/1 Queue

In the case of a $G/M/1$ queue, customers arrive according to a renewal process. The service times are independent exponentially distributed with parameter μ. There is one server, and the waiting room is infinite. The analysis methods available for a $G/M/1$ queue are very similar to those available for the $M/G/1$ queue. In this section we analyze the $G/M/1$ queue with the method of embedded Markov chain. The application of other analysis methods for the $G/M/1$ queue are left as exercises.

Let T_n be the time between the $(n-1)$st and nth arrivals (in the case of the $M/G/1$ queue it had another meaning). The arrivals constitute a renewal process, so $\{T_n\}$ is a sequence of i.i.d. random variables, and let T have the same distribution. For the sake of simplicity we assume that T is continuous with density function $a(x)$ and has a finite mean. Let $\lambda = 1/T$, i.e.,

$$\mathbf{P}(T \le x) = \int_0^x a(u)\,\mathrm{d}u \quad (x \ge 0)$$

and

$$\frac{1}{\lambda} = \int_0^\infty x\, a(x)\,\mathrm{d}x.$$

$L(t)$ denotes the number of customers in the system at moment t. Similarly to the $M/G/1$ queue, the process $\{L(t) : t \ge 0\}$ generally is not a Markov chain, and the future behavior at t depends not only on $L(t)$ but also on time elapsed from the moment of the last arrival. We will use the embedded Markov chain technique, and the embedded points will be the moments just before the arrivals.

8.4.1 Embedded Markov Chain

Let X_n be the number of customers in a $G/M/1$ system at the moment just before the entry of the nth one, formally

$$X_n = \lim_{\Delta \to 0+} N\left(\sum_{i=1}^n T_i - \Delta\right).$$

We show that $\{X_n\}$ is a homogeneous Markov chain. Let V'_n denote the number of customers served between the arrivals of the $(n-1)$st and nth customers (i.e., for T_n). $\{X_n\}$ satisfies the equation

$$X_{n+1} = X_n + 1 - V'_{n+1}. \tag{8.20}$$

It is not simple to work with the recursion Eq. (8.20) since V'_n depends on X_{n-1}, so $\{V'_n\}$ is not an identically distributed sequence (e.g., $V'_1 \equiv 0$). Let V_n be the number of customers that the system would have served had it not become empty. $\{V_n\}$ are i.i.d., which is a consequence of the fact that $\{T_n\}$ and the service times are independent. Equation (8.20) can be written in the form

$$X_{n+1} = (X_n + 1 - V_{n+1})^+, \tag{8.21}$$

from which it follows that $\{X_n\}$ is a homogeneous Markov chain.

We compute the transition probabilities of this chain:

$$\begin{aligned} p_{ij} &= \mathbf{P}\,(X_{n+1} = j \,|\, X_n = i) \\ &= \mathbf{P}\left((X_n + 1 - V_{n+1})^+ = j \,|\, X_n = i\right) \\ &= \mathbf{P}\left((i + 1 - V_{n+1})^+ = j \,|\, X_n = i\right) \\ &= \mathbf{P}\left((i + 1 - V_{n+1})^+ = j\right) \end{aligned} \tag{8.22}$$

because of the independence of X_n and V_{n+1}. Obviously, if $j > i+1$, then $p_{ij} = 0$. Let $0 < j \le i+1$, then in Eq. (8.22) we can cancel the sign of the positive part and

$$\begin{aligned} p_{ij} &= \mathbf{P}\,(V_{n+1} = i - j + 1) \\ &= \int_0^\infty \mathbf{P}\,(V_{n+1} = i - j + 1 | T_{n+1} = x)\; a(x)\,\mathrm{d}x \qquad (8.23) \\ &= \int_0^\infty \frac{(\mu x)^{i-j+1}}{(i-j+1)!} \mathrm{e}^{-\mu x} a(x)\,\mathrm{d}x \quad (0 < j \le i+1) \qquad (8.24) \end{aligned}$$

because, given that we always have customers, the moments of completion are a Poisson process with intensity μ, and its increment for x appears in Eq. (8.23). In this case, from Eq. (8.24) it follows that p_{ij} depends only on the differences in indices, i.e.,

$$p_{ij} = \mathbf{P}\,(V_{n+1} = i - j + 1) = \beta_{i-j+1} \quad (0 < j \le i+1).$$

The sum of elements in the rows of matrix $\mathbf{\Pi}$ is equal to 1; consequently,

$$p_{i0} = 1 - \sum_{j=1}^{\infty} p_{ij} = 1 - \sum_{j=1}^{i+1} \beta_{i-j+1} = 1 - \sum_{k=0}^{i} \beta_k.$$

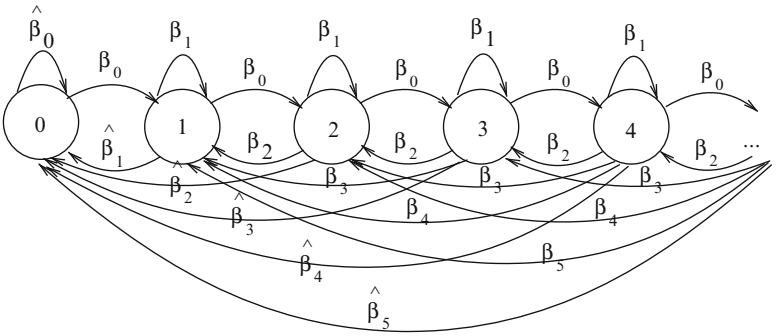


Fig. 8.9 Embedded CTMC of $G/M/1$ queue

The system behavior before customers arrive is depicted in Fig. 8.9, and the one-step state-transition probability matrix is

$$\mathbf{\Pi} = \begin{pmatrix} 1-\beta_0 & \beta_0 & 0 & 0 & \dots \\ 1-\beta_0-\beta_1 & \beta_1 & \beta_0 & 0 & \dots \\ 1-\beta_0-\beta_1-\beta_2 & \beta_2 & \beta_1 & \beta_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where

$$\beta_k = \int_0^\infty \frac{(\mu x)^k}{k!} \mathrm{e}^{-\mu x} a(x)\,\mathrm{d}x. \tag{8.25}$$

Consider the Markov chain defined by the preceding matrix. Let $P_{k\ell}(n)$ be the probability of the event that the system for n steps from state k arrives at state ℓ, and let us introduce the following notations:

$$\hat{\beta}_k = 1 - \sum_{i=0}^{k} \beta_i = \sum_{i=k+1}^{\infty} \beta_i, \qquad Q(z) = \sum_{i=0}^{\infty} \beta_i z^i, \qquad C(z) = \frac{Q(z)}{z},$$

$$P_\ell(n,z) = \sum_{k=0}^{\infty} P_{k\ell}(n) z^k, \qquad P_\ell(t,z) = \sum_{n=0}^{\infty} P_\ell(n,z) t^n, \qquad P_{0\ell}(t) = P_{0\ell}(n) t^n.$$

Let us fix the final state ℓ and write the inverse Kolmogorov equations for transition probabilities for $n+1$ steps:

$$P_{k\ell}(n+1) = \sum_{i=0}^{k} \beta_i P_{k+1-i,\ell}(n) + \hat{\beta}_k P_{0\ell}(n), \qquad k = 0, 1, 2, \dots .$$

Let us multiply these equations by z^k and sum up by k:

$$\sum_{k=0}^{\infty} P_{k\ell}(n+1)z^k = \sum_{k=0}^{\infty}\sum_{i=0}^{k} \beta_i P_{k+1-i,\ell}(n)z^k + \sum_{k=0}^{\infty} \hat{\beta}_k P_{0\ell}(n)z^k$$

$$= \frac{1}{z}Q(z)[P_\ell(n,z) - P_{0\ell}(n)] + P_{0\ell}(n)\sum_{k=0}^{\infty}\hat{\beta}_k z^k. \quad (8.26)$$

Since

$$\sum_{k=0}^{\infty}\hat{\beta}_k z^k = \sum_{k=0}^{\infty}(1-\beta_0-\ldots-\beta_k)z^k$$

$$= (1-\beta_0) + (1-\beta_0-\beta_1)z + (1-\beta_0-\beta_1-\beta_2)z^2 + \ldots$$

$$= (1-\beta_0)(1+z+z^2+\ldots) - \beta_1(z+z^2+\ldots) - \beta_2(z^2+z^3+\ldots) - \ldots$$

$$= (1-\beta_0)\frac{1}{1-z} - \beta_1\frac{z}{1-z} - \beta_2\frac{z^2}{1-z} - \ldots = \frac{1-Q(z)}{1-z},$$

from Eq. (8.26)

$$P_\ell(n+1,z) = \frac{1}{z}Q(z)\Big(P_\ell(n,z) - P_{0\ell}(n)\Big) + P_{0\ell}(n)\frac{1-Q(z)}{1-z}.$$

Multiplying this equation by t^n and summing up by n

$$\sum_{n=0}^{\infty} P_\ell(n+1)t^n = \frac{1}{z}Q(z)\sum_{n=0}^{\infty}\Big(P_\ell(n,t)t^n - P_{0\ell}(n)t^n\Big) + \frac{1-Q(z)}{1-z}\sum_{n=0}^{\infty}P_{0\ell}(n)t^n,$$

i.e.,

$$\frac{1}{t}[P_\ell(t,z) - P_\ell(0,z) = \frac{1}{z}Q(z)P_\ell(t,z) + P_{0\ell}(t)\frac{z-Q(z)}{z(1-z)},$$

or (using the initial value)

$$P_\ell(t,z) = \left(1 - t\frac{Q(z)}{z}\right)^{-1}\left(z^\ell + P_{0\ell}(t)t\frac{z-Q(z)}{z(1-z)}\right). \quad (8.27)$$

This expression contains the unknown generating function $P_{0\ell}(t)$, which will be determined [since $P_\ell(t,z)$ is analytical function] by means of the roots of the equation

$$1 - t\frac{Q(z)}{z} = 0$$

in $(0,1)$. We will need the following results.

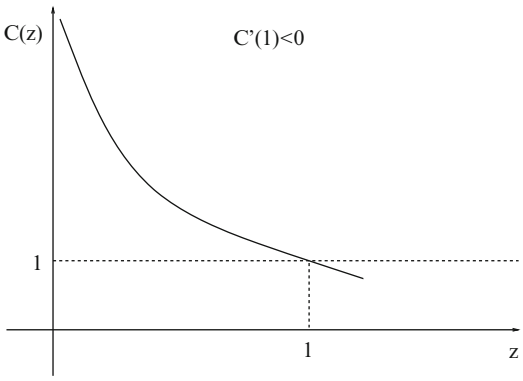


Fig. 8.10 The case where $C'(1) < 0$

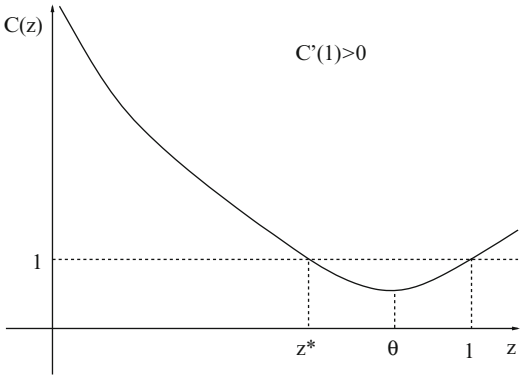


Fig. 8.11 The case where $C'(1) > 0$

Lemma 8.15. *If $C'(1) \leq 0$, then $C(z)$ is a continuous function on (0,1], and it is decreasing from $+\infty$ to 1.*

Proof. See Fig. 8.10.

$$C'(z) = \sum_{i=0}^{\infty}(i-1)\beta_i z^{i-1}, \qquad C''(z) = \sum_{i=0}^{\infty}(i-1)(i-2)\beta_i z^{i-3}.$$

$C''(z) > 0$ on the interval $(0, 1]$, so $C'(z)$ is monotonically increasing on $(0, 1]$ and on the open interval $C'(z) < C'(1) \leq 0$. Since $C(z)$ is continuous on $(0, 1]$, it decreases from $+\infty$ to $C(1) = 1$. □

Lemma 8.16. *If $0 < C'(1) < +\infty$, then there exists $\theta \in (0, 1)$ such that $C(z)$ monotonically decreases on $(0, \theta]$ from $+\infty$ to $C(\theta) < 1$ and on $[\theta, 1]$ is continuous and monotonically increases from $C(\theta)$ to $C(1) = 1$.*

Proof. See Fig. 8.11. $C''(z) > 0$ on $(0, 1)$, so $C'(z)$ monotonically increases from $-\infty$ to $C'(1) > 0$. Consequently, there is one and only one value $\theta \in (0, 1)$ for

which $C'(\theta) = 0$. On $(0, \theta)$, $C'(z) < 0$ and $C(z)$ is monotonically decreasing; on $(\theta, 1)$, $C'(z) > 0$ and $C(z)$ is monotonically increasing, $C(\theta) < C(1) = 1$. □

Corollary 8.17. *Under the conditions of the lemma there exists one and only one* $z^* \in (0, \theta)$ *such that on* $(0, z^*)$, $C(z)$ *monotonically decreases from* $+\infty$ *to 1, on* $[z^*, \theta]$ *it monotonically decreases from 1 to* $C(\theta)$, *and on* $(\theta, 1)$ *it monotonically increases from* $C(\theta)$ *to 1.*

Proof. The existence and uniqueness of z^* follows from the fact that $C(z)$ is a monotonically decreasing function on $(0, z^*]$, and $C(0+) = +\infty$ and $C(\theta) < 1$.

Let $z(t)$ be the root of the equation

$$\frac{1}{t} = C(z(t))$$

on the interval $(0, 1)$. Substituting it into the numerator of the right-hand side of Eq. (8.27),

$$z^{\ell}(t) + P_{0\ell}(t)t\frac{1 - C(z(t))}{1 - z(t)} = z^{\ell}(t) + P_{0\ell}(t)t\frac{1 - \frac{1}{t}}{1 - z(t)}$$
$$= z^{\ell}(t) + P_{0\ell}(t)\frac{t - 1}{1 - z(t)} = 0,$$

whence

$$P_{0\ell}(t) = z^{\ell}(t)\frac{1 - z(t)}{1 - t}.$$

The chain is irreducible and aperiodic, so the equilibrium distribution does not depend on the initial state. By the Tauberian theorem [90]

$$P_{\ell} = \lim_{t \to 1}(1 - t)P_{0\ell}(t) = z^{*\ell}(1 - z^*),$$

where z^* is the root of the equation $C(z) = 1$ on the interval $(0, 1)$. If $C'(1) > 0$, then z^* lies between 0 and 1, so we get a nondegenerate distribution. In the case $C'(1) \leq 0$, we have $z^* = 1$, and the distribution is degenerate. □

Comment 8.18. *If* $A(x)$ *denotes the distribution function of interarrival times, then the generating function of the number of customers served is* $Q(z) = A^{\sim}(\mu - \mu z)$, *and from* $C(z^*) = 1$ *it follows that* z^* *is the only root of*

$$z^* = A^{\sim}(\mu - \mu z^*)$$

in the interval $0 < z^* < 1$ *(see Kleinrock [55]).*

Comment 8.19. *The condition* $C'(1) > 0$ *can easily be expressed with the help of the generating function* $Q(z)$ *since*

$$\left.\frac{Q'(z)z - Q(z)}{z^2}\right|_{z=1} = Q'(1) - 1 > 0,$$

from which the stability condition is $Q'(1) > 1$. *It has a simple meaning, namely, the mean value of the number of customers served between the entries of two successive customers must be more than one. This condition is equivalent to the inequality* $\mu/\lambda > 1$.

Comment 8.20. *We have examined the* $G/M/1$ *system at moments just before the arrivals and found the stability condition for these points. We mention that in contrast to the* $M/G/1$ *system, this distribution does not hold for the inner points.*

8.5 Exercises

Exercise 8.1. There is an $M/G/1$ queue. The arrival intensity is λ, and the service time is exponentially distributed with the parameter μ_2 with probability $1 - p$ and is the sum of two independent exponentially distributed random variables with the parameters μ_1 and μ_2 with probability p.

- Compute server utilization.
- Compute the coefficient of variation of the service time.
- Compute the mean system time of customers.
- Compute the mean number of customers in the buffer.

Exercise 8.2. Patients arrive at a dentist's office according to a Poisson process with intensity λ. Arriving patients enter the dentist's operatory if no one is there; otherwise, they wait in the waiting room. In the dentist's operatory there is a registration of time D (deterministic). With probability p a patient is directed to the dentist for treatment, which takes an exponentially distributed time with the parameter μ; with probability $1 - p$ the patient is rejected.

- Compute patients' mean time in the waiting room.
- Compute the probability that an arriving patient must wait.
- Compute the mean waiting time.

Exercise 8.3. $F_A(t)$ is the interarrival distribution in an $G/M/1$ queue whose service rate is μ. $N(t)$ is the number of customers in the system at time t, and $T_1, T_2, \ldots$ denote the arrival instances of the first, second, etc. customers. The mean of the stationary number of customers is $\bar{N} = \lim_{t\to\infty} E(N(t))$, and the mean of the stationary number of customers at arrival instants is $\check{N} = \lim_{n\to\infty} E(N(T_n-))$. Compute the relation of $\bar{N}$ and $\check{N}$ if

- The interarrival distribution is hyperexponential $[F_A(t) = 1-p\mathrm{e}^{\lambda_1 t}-(1-p)\mathrm{e}^{\lambda_2 t}]$,
- The interarrival distribution is deterministic,
- The interarrival distribution is exponential.

Exercise 8.4. Find the mean value of the number of customers in an $M/G/1$ system and in the waiting queue. Let us consider the cases of $M/M/1$ and $M/D/1$ systems.

Exercise 8.5. Using the Pollaczek–Khinchin transform equation show that in an $M/M/1$ system the equilibrium distribution is geometrical.

Exercise 8.6. Let us consider an $M/G/1$ system with bulk arrivals. An arriving group with probability g_i consists of i customers. Show that the generating function of the number of customers entering during time t is $\mathrm{e}^{-\lambda t(1-G(z))}$, where λ is the intensity of arrivals and $G(z) = \sum_{i=1}^{\infty} g_i z^i$.

Exercise 8.7. Show that in an $M/G/1$ system with bulk arrivals the generating function of the number of customers arriving for the service time of a customer is $b^{\sim}(\lambda(1 - G(z))$, where $b^{\sim}(s)$ is the Laplace–Stieltjes transform of the distribution function of this service time.

Chapter 9
Queueing Systems with Structured Markov Chains

In the previous chapters we studied queueing systems with different interarrival and service time distributions. Chapter 7 is devoted to the analysis of queueing systems with exponential interarrival and service time distributions. The number of customers in these queueing systems is characterized by CTMCs with a generally nonhomogeneous birth-and-death structure. In contrast, Chap. 8 is devoted to the analysis of queueing systems with nonexponential interarrival and service time distributions. It turns out that far more complex analysis approaches are required for the analysis of queues with nonexponential interarrival and service time distributions. In this chapter we introduce queueing systems whose interarrival and service time distributions are nonexponential, but they can be analyzed with CTMCs. Indeed in this chapter we demonstrate the use of the results of Chap. 5 for the analysis of queueing systems with phase-type (PH) distributed interarrival and service times or with arrival and service processes that are MAPs. The main message of this chapter is that in queueing models the presence of PH or MAP processes instead of exponential distributions results in a generalization of the underlying CTMCs from birth-and-death processes to quasi-birth-and-death (QBDs) processes.

9.1 $PH/M/1$ Queue

One of the simplest queueing systems with nonexponentially distributed interarrival time distribution is the $PH/M/1$ queue. We study this queue in detail in order to demonstrate the elementary steps needed to construct the matrix block structure of a CTMC describing the behavior of the queue.

We consider a queueing system whose arrival process is composed by independent and identically PH distributed interarrival periods characterized by initial probability vector $\boldsymbol{\tau}$ and transient generator matrix $\boldsymbol{T}$. Consequently, the arrival process is a PH renewal process with representation $(\boldsymbol{\tau}, \boldsymbol{T})$. The service time is exponentially distributed with parameter μ. The queue has one server and an unlimited buffer. The service discipline is FIFO.

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, DOI 10.1007/978-1-4614-5317-8_9,

9.1.1 QBD Process of PH/M/1 Queue

This queueing system can be analyzed as a $G/M/1$ queue using the results of the previous chapter because the PH distributed interarrival time is a special case of general, nonexponential interarrival time distributions. But using the fact that the PH distribution is characterized by a background Markov chain we can also analyze the $PH/M/1$ queue as a compound CTMC $\{N(t), J(t)\}$, where $N(t)$ is the number of customers in the queue and $J(t)$ is the state (phase) of the Markov chain characterizing the PH distributed arrivals at time t.

If at time t the state of a compound CTMC is $(N(t), J(t)) = (n, j)$, then the following state transitions are possible.

- There might be a phase transition from phase j to k $(k \neq j)$ in a background Markov chain of the PH distribution without an arrival. The rate of this transition from (n, j) to (n, k) is T_{jk}.
- The Markov chain of the PH distribution might move to the absorbing state and generate an arrival. In this case the number of customers in the system increases by one and a new PH distributed interval starts according to the initial phase distribution $\boldsymbol{\tau}$. Let $\boldsymbol{t}$ be a column vector containing transition rates to the absorbing state, $\boldsymbol{t} = -\boldsymbol{T}\mathbb{1}$. The transition rate from (n, j) to $(n+1, k)$ due to these steps is $\boldsymbol{t}_j \boldsymbol{\tau}_k$.
- If $n > 0$, then there is a customer in the server that which is served with an exponentially distributed service time with parameter μ. When the service completes, the number of customers in the system decreases by one and, due to the independence of the arrival and the service processes, the service completion does not affect the phase arrival process. Thus the transition rate from (n, j) to $(n-1, j)$ is μ and from (n, j) to $(n-1, k)$ $(k \neq j)$ it is 0.

These possible transitions define all nondiagonal elements of the generator matrix of the CTMC. In the case of a PH arrival process with two phases, the generator matrix has the form

$$\boldsymbol{Q} = \left[\begin{array}{cc|cc|cc|cc}
\bullet & \boldsymbol{T}_{12} & \boldsymbol{t}_1\boldsymbol{\tau}_1 & \boldsymbol{t}_1\boldsymbol{\tau}_2 & 0 & 0 & 0 & 0 \\
\boldsymbol{T}_{21} & \bullet & \boldsymbol{t}_2\boldsymbol{\tau}_1 & \boldsymbol{t}_2\boldsymbol{\tau}_2 & 0 & 0 & 0 & 0 \\
\hline
\mu & 0 & \bullet & \boldsymbol{T}_{12} & \boldsymbol{t}_1\boldsymbol{\tau}_1 & \boldsymbol{t}_1\boldsymbol{\tau}_2 & 0 & 0 \\
0 & \mu & \boldsymbol{T}_{21} & \bullet & \boldsymbol{t}_2\boldsymbol{\tau}_1 & \boldsymbol{t}_2\boldsymbol{\tau}_2 & 0 & 0 \\
\hline
0 & 0 & \mu & 0 & \bullet & \boldsymbol{T}_{12} & \boldsymbol{t}_1\boldsymbol{\tau}_1 & \boldsymbol{t}_1\boldsymbol{\tau}_2 \\
0 & 0 & 0 & \mu & \boldsymbol{T}_{21} & \bullet & \boldsymbol{t}_2\boldsymbol{\tau}_1 & \boldsymbol{t}_2\boldsymbol{\tau}_2 \\
\hline
 & & & & \ddots & \ddots & \ddots & \ddots \\
 & & & & \ddots & \ddots & \ddots & \ddots
\end{array}\right].$$

The diagonal elements are determined by the nondiagonal elements due to the fact that the row sum of the generator matrix is zero. This matrix already indicates that the generator matrix has a regular structure on the matrix block level, highlighted by

the horizontal and vertical lines. The matrix blocks are closely related to the vectors and the matrix characterizing the PH arrival process:

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{T} & \boldsymbol{t}\boldsymbol{\tau} & & & \\ \hline \mu\boldsymbol{I} & \boldsymbol{T}-\mu\boldsymbol{I} & \boldsymbol{t}\boldsymbol{\tau} & & \\ \hline & \mu\boldsymbol{I} & \boldsymbol{T}-\mu\boldsymbol{I} & \boldsymbol{t}\boldsymbol{\tau} & \\ \hline & & \mu\boldsymbol{I} & \boldsymbol{T}-\mu\boldsymbol{I} & \boldsymbol{t}\boldsymbol{\tau} \\ \hline & & & \ddots & \ddots \\ \hline \end{array}.$$

The nondiagonal elements of this generator matrix defined by matrix blocks are readily identified with those of a detailed generator matrix. For the validity of the diagonal element we evaluate the row sum of the elements in a row of matrix blocks. If $n = 0$, the row sum is $\boldsymbol{T}\mathbb{1} + \boldsymbol{t}\boldsymbol{\tau}\mathbb{1} = \boldsymbol{T}\mathbb{1} + \boldsymbol{t} = \boldsymbol{0}$ because $\boldsymbol{\tau}\mathbb{1} = 1$ and $\boldsymbol{t} = -\boldsymbol{T}\mathbb{1}$. If $n > 0$, the row sum is $\mu\boldsymbol{I}\mathbb{1} + (\boldsymbol{T} - \mu\boldsymbol{I})\mathbb{1} + \boldsymbol{t}\boldsymbol{\tau}\mathbb{1} = \boldsymbol{T}\mathbb{1} + \boldsymbol{t} = \boldsymbol{0}$.

The block level structure of the generator matrix shows that $\{N(t), J(t)\}$ is a QBD process with regular level 0. The forward, local, backward, and level 0 local matrices of this QBD process are $\boldsymbol{F} = \boldsymbol{t}\boldsymbol{\tau}$, $\boldsymbol{L} = \boldsymbol{T} - \mu\boldsymbol{I}$, $\boldsymbol{B} = \mu\boldsymbol{I}$, and $\boldsymbol{L}' = \boldsymbol{T}$.

9.1.2 Condition of Stability

From the $G/M/1$ interpretation of the $PH/M/1$ queue we already know that the queue is stable as long as the mean interarrival time is greater than the mean service time, that is, $\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1} > 1/\mu$. Now we analyze the relation of this condition to the stability condition of the QBD process. The phase process of the regular levels is a CTMC with generator $\boldsymbol{B} + \boldsymbol{L} + \boldsymbol{F} = \boldsymbol{T} + \boldsymbol{t}\boldsymbol{\tau}$. Let $\boldsymbol{\alpha}$ be the stationary distribution of the phase process (i.e., the solution of $\boldsymbol{\alpha}(\boldsymbol{F} + \boldsymbol{L} + \boldsymbol{B}) = \boldsymbol{0}, \boldsymbol{\alpha}\mathbb{1} = 1$).

Theorem 9.1.

$$\boldsymbol{\alpha} = \frac{\boldsymbol{\tau}(-\boldsymbol{T})^{-1}}{\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1}}.$$

Proof. The normalizing condition obviously holds:

$$\frac{\boldsymbol{\tau}(-\boldsymbol{T})^{-1}}{\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1}}\,\mathbb{1} = 1.$$

For the product of the stationary solution vector and the generator of the phase process we have

$$\boldsymbol{\tau}(-\boldsymbol{T})^{-1}(\boldsymbol{B}+\boldsymbol{L}+\boldsymbol{F}) = \boldsymbol{\tau}(-\boldsymbol{T})^{-1}(\boldsymbol{T}-\boldsymbol{T}\mathbb{1}\boldsymbol{\tau}) = -\boldsymbol{\tau}+\boldsymbol{\tau} = \mathbf{0},$$

where we neglect the normalizing constant $1/\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1}$. □

Based on the stationary distribution of the phase process, the condition of stability of the QBD process is $\boldsymbol{\alpha}\boldsymbol{B}\mathbb{1} > \boldsymbol{\alpha}\boldsymbol{F}\mathbb{1}$ where $\boldsymbol{\alpha}\boldsymbol{B}\mathbb{1} = \boldsymbol{\alpha}\mu\boldsymbol{I}\mathbb{1} = \mu$ and

$$\boldsymbol{\alpha}\boldsymbol{F}\mathbb{1} = \frac{\boldsymbol{\tau}(-\boldsymbol{T})^{-1}}{\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1}}(-\boldsymbol{T}\mathbb{1}\boldsymbol{\tau})\mathbb{1} = \frac{1}{\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1}}.$$

9.1.3 Performance Measures

The main performance measures of $PH/M/1$ queues are based on the stationary distribution of the $\{N(t), J(t)\}$ QBD process. According to Theorem 5.9, the row vector of the stationary probabilities with n customers can be computed as $\boldsymbol{\pi}_n = \boldsymbol{\pi}_0\boldsymbol{R}^n$, where matrix $\boldsymbol{R}$ is the solution (the only one whose eigenvalues are inside the unit disk) of

$$\boldsymbol{F} + \boldsymbol{R}\boldsymbol{L} + \boldsymbol{R}^2\boldsymbol{B} = \mathbf{0}$$

and vector π_0 is the solution of the linear system

$$\boldsymbol{\pi}_0(\boldsymbol{L}' + \boldsymbol{R}\boldsymbol{B}) = \mathbf{0}, \quad \boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\mathbb{1} = 1.$$

Below we compute the main performance measures assuming that the matrix-geometric stationary distribution is known.

Utilization

The only server of a queueing system is busy when the number of customers in the system is at least 1. Consequently, the utilization is

$$\rho = \lim_{t\to\infty}\mathbf{P}(N(t)\geq 1) = \sum_{n=1}^{\infty}\boldsymbol{\pi}_n\mathbb{1} = 1-\boldsymbol{\pi}_0\mathbb{1}.$$

Number of Customers

The distribution of the stationary number of customers in the queue is

$$p_n = \lim_{t\to\infty}\mathbf{P}(N(t)=n) = \boldsymbol{\pi}_n\mathbb{1} = \boldsymbol{\pi}_0\boldsymbol{R}^n\mathbb{1}.$$

The mean number of customers can be computed as

$$\begin{aligned}\mathbf{E}(N) &= \lim_{t\to\infty}\mathbf{E}(N(t)) = \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n\boldsymbol{\pi}_0\boldsymbol{R}^n\mathbb{1}\\ &= \sum_{n=0}^{\infty}\sum_{k=0}^{n}\boldsymbol{\pi}_0\boldsymbol{R}^n\mathbb{1} = \boldsymbol{\pi}_0\sum_{k=0}^{\infty}\sum_{n=k}^{\infty}\boldsymbol{R}^n\mathbb{1}\\ &= \boldsymbol{\pi}_0\sum_{k=0}^{\infty}\boldsymbol{R}^k(\boldsymbol{I}-\boldsymbol{R})^{-1}\mathbb{1} = \boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-2}\mathbb{1}.\end{aligned}$$

The distribution of the stationary number of customers right before a customer arrival is defined as $q_n = \lim_{k\to\infty}\mathbf{P}(N(T_k-) = n)$, where T_k denotes the arrival instant of the kth customer. We have

$$q_n = \frac{\text{stationary arrival rate from level } n}{\text{stationary customer arrival rate}} = \frac{\boldsymbol{\pi}_n\boldsymbol{t}}{\sum_{i=0}^{\infty}\boldsymbol{\pi}_i\boldsymbol{t}} = \frac{\boldsymbol{\pi}_0\boldsymbol{R}^n\boldsymbol{t}}{\boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{t}},$$

and similarly

$$\mathbf{E}(N_A) = \lim_{k\to\infty}\mathbf{E}(N(T_k-)) = \sum_{n=0}^{\infty} nq_n = \frac{\boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{t}}{\boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{t}}.$$

It is worth mentioning that the arrival process is not a Poisson process (in general) and the distribution of the stationary number of customers and that of the stationary number of customers at arrivals differ.

System Time

If a customer arrives at the queue when there are n customers in front of it, then its waiting time is the sum of the remaining service time of the customer in the server, if any (which is exponentially distributed with parameter μ), and the total service time of the customers waiting in front of the newly arrived one, if any (which is also exponentially distributed with parameter μ). The system time (T) is the sum of the waiting time (W) and the service time (S). All together, if a customer arrives when there are n other customers in the queue, then its system time is the sum of n independent exponentially distributed random variables with parameter μ. We describe the Laplace transform of the system time because the sum of independent random variables has a simple form in the Laplace domain. The Laplace transform of the exponentially distributed service time with parameter μ is $\mathbf{E}\left(\mathrm{e}^{sS}\right) = \frac{\mu}{\mu+s}$.

$$\begin{aligned}f_T^*(s) = \mathbf{E}\left(\mathrm{e}^{sT}\right) &= \sum_{n=0}^{\infty} q_n\left(\frac{\mu}{\mu+s}\right)^n = \sum_{n=0}^{\infty}\frac{\boldsymbol{\pi}_0\boldsymbol{R}^n\boldsymbol{t}}{\boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{t}}\left(\frac{\mu}{\mu+s}\right)^n\\ &= \boldsymbol{\pi}_0\frac{1}{\boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{t}}\left(\boldsymbol{I}-\frac{\mu}{\mu+s}\boldsymbol{R}\right)^{-1}\boldsymbol{t}\end{aligned}$$

Using the relation of the Laplace transform and the moments, the mean system time can be computed as

$$\begin{aligned}\mathbf{E}(T) &= -\frac{d}{ds} f_T^*(s)|_{s=0} = -\boldsymbol{\pi}_0 \frac{1}{\boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{t}} \frac{d}{ds}\left(\boldsymbol{I} - \frac{\mu}{\mu+s}\boldsymbol{R}\right)^{-1}\Bigg|_{s=0} \boldsymbol{t} \\ &= \boldsymbol{\pi}_0 \frac{1}{\boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{t}} \left(\boldsymbol{I} - \frac{\mu}{\mu+s}\boldsymbol{R}\right)^{-2} \frac{\mu}{(\mu+s)^2}\boldsymbol{R}\Bigg|_{s=0} \boldsymbol{t} \\ &= \frac{\boldsymbol{\pi}_0 (\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{R}\,\boldsymbol{t}}{\mu\,\boldsymbol{\pi}_0(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{t}},\end{aligned}$$

where we utilized that $(\boldsymbol{I}-\boldsymbol{R})$ and $\boldsymbol{R}$ commute. Further performance measures, like waiting time, can be computed in a similar manner.

9.2 $M/PH/1$ Queue

In this section we analyze the other simple queueing system with an underlying QBD process. This is the $M/PH/1$ queue, where the arrival process is a Poisson process at a rate λ, the service time is PH distributed with representation $\tau, \boldsymbol{T}$ of size J, and there is a single server and an infinite buffer. Similarly, column vector $\boldsymbol{t}$ contains the transition rates to the absorbing state ($\boldsymbol{t} = -\boldsymbol{T}\mathbb{1}$). The most important difference between the $PH/M/1$ queue and the $M/PH/1$ queue is the structure of the underlying QBD process. This was a QBD process with regular level zero in the previous section, and it will be a QBD with irregular level zero in this section. Another useful feature of the $M/PH/1$ queue, which is unique among the queueing systems with an underlying QBD process, is that matrix $\boldsymbol{R}$ can be expressed in closed form.

9.2.1 QBD of M/PH/1 Queue

As with the $PH/M/1$ queue, the behavior of the $M/PH/1$ queue is characterized by a compound CTMC $\{N(t), J(t)\}$, where $N(t)$ is the number of customers in the queue and $J(t)$ is the state (phase) of the Markov chain characterizing the PH distributed service time at t. One of the main differences between the $PH/M/1$ and the $M/PH/1$ queues comes from the fact that the service process is inactive (does not exist) when there is no customer in the queue. Consequently, level 0 of the underlying QBD process has a different structure than the higher levels. The QBD process has a single phase at level zero and J phases at higher levels. Accordingly, the structure of the transitions from and to level 0 is different from the regular ones.

If at time t the state of the QBD process is $(N(t), J(t)) = (n, j)$, then the following state transitions are possible.

- If $n \geq 1$, then there might be a phase transition from (n, j) to (n, k) $(k \neq j)$ in the background Markov chain of the PH distribution without a departure at a rate T_{jk}.
- If $n \geq 2$, then the Markov chain of the PH distribution might move to the absorbing state at a rate $\boldsymbol{t}_j$, which represents the service completion of the customer in the server and the departure of this customer. In this case, the number of customers in the system decreases by one and a new PH distributed service time starts according to the initial phase distribution $\boldsymbol{\tau}$. The transition rate from (n, j) to $(n-1, k)$ is $\boldsymbol{t}_j \boldsymbol{\tau}_k$.
- If $n = 1$ and the PH distribution moves to the absorbing state at a rate $\boldsymbol{t}_j$, then the only customer leaves the queue. In this case, the queue becomes idle and the service process becomes inactive. As a result, there might be a transition from $(1, j)$ to $(0, 1)$ at a rate $\boldsymbol{t}_j$.
- If $n \geq 1$, then there is one customer in the server and the service process is active. In this case an arrival at a rate λ increases the number of customers in the queue and maintains the phase of the service process. Thus the transition rate from (n, j) to $(n+1, j)$ is λ, and from (n, j) to $(n+1, k)$ $(k \neq j)$ it is 0.
- If $n = 0$, then the arrival of a new customer at a rate λ initiates the service of the newly arrived customer according to the initial phase distribution $\boldsymbol{\tau}$. In this case the transition rate from $(0, 1)$ to $(1, k)$ is $\lambda \boldsymbol{\tau}_k$.

These transitions define all nondiagonal elements of a generator matrix. If $J = 2$, then we have

$$\boldsymbol{Q} = \left[\begin{array}{cc|cc|cc|cc}
\bullet & \lambda\boldsymbol{\tau}_1 & \lambda\boldsymbol{\tau}_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
\boldsymbol{t}_1 & \bullet & \boldsymbol{T}_{12} & \lambda & 0 & 0 & 0 & 0 & 0 \\
\boldsymbol{t}_2 & \boldsymbol{T}_{21} & \bullet & 0 & \lambda & 0 & 0 & 0 & 0 \\
\hline
0 & \boldsymbol{t}_1\boldsymbol{\tau}_1 & \boldsymbol{t}_1\boldsymbol{\tau}_2 & \bullet & \boldsymbol{T}_{12} & \lambda & 0 & 0 & 0 \\
0 & \boldsymbol{t}_2\boldsymbol{\tau}_1 & \boldsymbol{t}_2\boldsymbol{\tau}_2 & \boldsymbol{T}_{21} & \bullet & 0 & \lambda & 0 & 0 \\
\hline
0 & 0 & 0 & \boldsymbol{t}_1\boldsymbol{\tau}_1 & \boldsymbol{t}_1\boldsymbol{\tau}_2 & \bullet & \boldsymbol{T}_{12} & \lambda & 0 \\
0 & 0 & 0 & \boldsymbol{t}_2\boldsymbol{\tau}_1 & \boldsymbol{t}_2\boldsymbol{\tau}_2 & \boldsymbol{T}_{21} & \bullet & 0 & \lambda \\
\hline
 & & & & & \ddots & \ddots & \ddots & \ddots \\
 & & & & & \ddots & \ddots & \ddots & \ddots
\end{array}\right],$$

and on the level of matrix blocks the generator matrix of the QBD process is

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline -\lambda & \lambda\boldsymbol{\tau} & & & \\ \hline \boldsymbol{t} & \boldsymbol{T}-\lambda\boldsymbol{I} & \lambda\boldsymbol{I} & & \\ \hline & \boldsymbol{t}\boldsymbol{\tau} & \boldsymbol{T}-\lambda\boldsymbol{I} & \lambda\boldsymbol{I} & \\ \hline & & \boldsymbol{t}\boldsymbol{\tau} & \boldsymbol{T}-\lambda\boldsymbol{I} & \lambda\boldsymbol{I} \\ \hline & & & \ddots & \ddots \\ \hline \end{array}.$$

That is, $\boldsymbol{F} = \lambda\boldsymbol{I}$, $\boldsymbol{L} = \boldsymbol{T} - \lambda\boldsymbol{I}$, $\boldsymbol{B} = \boldsymbol{t}\boldsymbol{\tau}$, and the special matrix blocks at the zero level are $\boldsymbol{F}' = \lambda\boldsymbol{\tau}$, $\boldsymbol{L}' = -\lambda$, $\boldsymbol{B}' = \boldsymbol{t}$. Using $\boldsymbol{t} = -\boldsymbol{T}\mathbb{1}$ it is easy to check that the row sum of each row is zero.

The condition of the stability of this QBD process can be computed in a very similar way as in the case of PH/M/1 queue. The QBD is stable if $\lambda < \frac{1}{\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1}}$.

9.2.2 Closed-Form Solution of Stationary Distribution

Let $\boldsymbol{\pi} = \{\pi_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots\}$ be the partitioned stationary probability vector of the QBD process. The partitioned form of the set of stationary equations $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$ is

$$-\pi_0\lambda + \boldsymbol{\pi}_1\boldsymbol{t} = 0, \tag{9.1}$$

$$\pi_0\lambda\boldsymbol{\tau} + \boldsymbol{\pi}_1(\boldsymbol{T}-\lambda\boldsymbol{I}) + \boldsymbol{\pi}_2\boldsymbol{t}\boldsymbol{\tau} = \boldsymbol{0}, \tag{9.2}$$

$$\boldsymbol{\pi}_{n-1}\lambda\boldsymbol{I} + \boldsymbol{\pi}_n(\boldsymbol{T}-\lambda\boldsymbol{I}) + \boldsymbol{\pi}_{n+1}\boldsymbol{t}\boldsymbol{\tau} = \boldsymbol{0} \quad \forall n \geq 2. \tag{9.3}$$

The solution of this set of equations can be expressed in a closed matrix-geometric form.

Theorem 9.2. *For* $n \geq 1$

$$\boldsymbol{\pi}_n = \pi_0\boldsymbol{\tau}\boldsymbol{R}^n,$$

where $\pi_0 = 1 - \lambda\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1}$ *and* $\boldsymbol{R} = \lambda(\lambda\boldsymbol{I} - \boldsymbol{T} - \lambda\mathbb{1}\boldsymbol{\tau})^{-1}$.

Proof. Substituting Eq. (9.1) into Eq. (9.2) gives

$$\boldsymbol{\pi}_1(\boldsymbol{t}\boldsymbol{\tau} + \boldsymbol{T} - \lambda\boldsymbol{I}) + \boldsymbol{\pi}_2\boldsymbol{t}\boldsymbol{\tau} = \boldsymbol{0}\,.$$

Multiplying this expression by $\mathbb{1}$ from the right we obtain $\boldsymbol{\pi}_1\lambda\mathbb{1} = \boldsymbol{\pi}_2\boldsymbol{t}$. Now we take Eq. (9.3) with $n = 2$, multiply it by $\mathbb{1}$ from the right, and substitute $\boldsymbol{\pi}_1\lambda\mathbb{1} = \boldsymbol{\pi}_2\boldsymbol{t}$. This results in $\boldsymbol{\pi}_2\lambda\mathbb{1} = \boldsymbol{\pi}_3\boldsymbol{t}$. Recursively multiplying Eq. (9.3) by $\mathbb{1}$

and substituting the previous result we obtain

$$\lambda \boldsymbol{\pi}_n \mathbb{1} = \boldsymbol{\pi}_{n+1} \boldsymbol{t} \quad \forall n \geq 1.$$

Substituting this expression into the third term of Eq. (9.3) gives

$$\lambda \boldsymbol{\pi}_{n-1} + \boldsymbol{\pi}_n (\boldsymbol{T} - \lambda \boldsymbol{I}) + \lambda \boldsymbol{\pi}_n \mathbb{1} \boldsymbol{\tau} = \boldsymbol{0} \quad \forall n \geq 2,$$

whence

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1} \underbrace{\lambda (\lambda \boldsymbol{I} - \boldsymbol{T} - \lambda \mathbb{1} \boldsymbol{\tau})^{-1}}_{\boldsymbol{R}} \quad \forall n \geq 2.$$

Additionally, from Eq. (9.2) we have $\boldsymbol{\pi}_1 = \pi_0 \boldsymbol{\tau} \boldsymbol{R}$. π_0, the probability that the server is idle, can be obtained from Little's law when it is applied to the server itself. It says that $\mathbf{E}(N_S) = \lambda \mathbf{E}(S)$ (the mean number of customers in the server equals the arrival rate times the mean service time). In our case $\mathbf{E}(S) = \boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1}$. In a single-server queue $\mathbf{E}(N_S)$ is the probability that the server is busy, i.e., $\mathbf{E}(N_S) = 1 - \pi_0$, indeed, it is the utilization in this case. □

9.2.3 *Performance Measures*

The computation of the main performance measures follows the same pattern as those of the $PH/M/1$ queue, but in the case of $M/PH/1$ queues we can utilize the closed form of the stationary distribution.

Number of Customers

The distribution of the stationary number of customers in a queue is

$$p_n = \lim_{t \to \infty} \mathbf{P}(N(t) = n) = \boldsymbol{\pi}_n \mathbb{1} = \pi_0 \boldsymbol{\tau} \boldsymbol{R}^n \mathbb{1}, \quad n \geq 1,$$

and $p_0 = \pi_0 = 1 - \lambda \boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbb{1}$. The mean number of customers can be computed in a similar way as in case of the $PH/M/1$ queue:

$$\mathbf{E}(N) = \lim_{t \to \infty} \mathbf{E}(N(t)) = \sum_{n=0}^{\infty} n p_n = \sum_{n=1}^{\infty} n \pi_0 \boldsymbol{\tau} \boldsymbol{R}^n \mathbb{1} = \pi_0 \boldsymbol{\tau} (\boldsymbol{I} - \boldsymbol{R})^{-2} \mathbb{1}.$$

The distribution of the stationary number of customers right before a customer arrival is

$$q_n = \frac{\text{stationary arrival rate from level } n}{\text{stationary customer arrival rate}} = \frac{\boldsymbol{\pi}_n \mathbb{1} \lambda}{\sum_{i=0}^{\infty} \boldsymbol{\pi}_i \mathbb{1} \lambda} = \boldsymbol{\pi}_n \mathbb{1} = p_n.$$

The Poisson arrival process ensures that the distribution of the stationary number of customers and that of the stationary number of customers at arrivals are identical.

System Time

The Laplace transform of the PH distributed service time is $\mathbf{E}\left(\mathrm{e}^{sS}\right) = \boldsymbol{\tau}(s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t}$. Using this we get that the Laplace transform of the system time is

$$\begin{aligned} f_T^*(s) = \mathbf{E}\left(\mathrm{e}^{sT}\right) &= \sum_{n=0}^{\infty} q_n \left(\boldsymbol{\tau}(s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t}\right)^n = \sum_{n=0}^{\infty} \pi_0 \boldsymbol{\tau} \boldsymbol{R}^n \mathbb{1} \left(\boldsymbol{\tau}(s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t}\right)^n \\ &= \pi_0 \boldsymbol{\tau} \left(\boldsymbol{I} - \boldsymbol{\tau}(s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t}\boldsymbol{R}\right)^{-1} \mathbb{1}. \end{aligned}$$

The mean system time can be computed from the Laplace transform as

$$\begin{aligned} \mathbf{E}(T) &= -\frac{\mathrm{d}}{\mathrm{d}s} f_T^*(s)|_{s=0} = \pi_0 \boldsymbol{\tau} \frac{\mathrm{d}}{\mathrm{d}s} \left(\boldsymbol{I} - \boldsymbol{\tau}(s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t}\boldsymbol{R}\right)^{-1}\Big|_{s=0} \mathbb{1} \\ &= \pi_0 \boldsymbol{\tau} \left(\boldsymbol{I} - \boldsymbol{\tau}(s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t}\boldsymbol{R}\right)^{-2} \boldsymbol{\tau}(s\boldsymbol{I} - \boldsymbol{T})^{-2}\boldsymbol{t}\boldsymbol{R}\mathbb{1} \\ &= \pi_0 \boldsymbol{\tau} \left(\boldsymbol{I} - \boldsymbol{R}\right)^{-2} \boldsymbol{\tau}(-\boldsymbol{T})^{-1}\boldsymbol{R}\mathbb{1}, \end{aligned}$$

where we utilized $\boldsymbol{t} = -\boldsymbol{T}\mathbb{1}$ in the last step.

9.3 Other Queues with Underlying QBD

9.3.1 MAP/M/1 Queue

The difference between the $PH/M/1$ queue and the $MAP/M/1$ queue is minor. We focus our attention mainly on the extension from $PH/M/1$ queues to $MAP/M/1$ queues. Let the arrival process be a MAP with representation $\boldsymbol{D_0}, \boldsymbol{D_1}$, and let the service time be exponentially distributed with parameter μ. The $MAP/M/1$ queue has a single server and an infinite buffer. These possible transitions of the $(N(t), J(t)))$ CTMC in the case of a MAP arrival process with two phases are as follows:

$$\boldsymbol{Q} = \begin{array}{|cc|cc|cc|cc|} \hline \bullet & \boldsymbol{D_{0_{12}}} & \boldsymbol{D_{1_{11}}} & \boldsymbol{D_{1_{12}}} & 0 & 0 & 0 & 0 \\ \boldsymbol{D_{0_{21}}} & \bullet & \boldsymbol{D_{1_{21}}} & \boldsymbol{D_{1_{22}}} & 0 & 0 & 0 & 0 \\ \hline \mu & 0 & \bullet & \boldsymbol{D_{0_{12}}} & \boldsymbol{D_{1_{11}}} & \boldsymbol{D_{1_{12}}} & 0 & 0 \\ 0 & \mu & \boldsymbol{D_{0_{21}}} & \bullet & \boldsymbol{D_{1_{21}}} & \boldsymbol{D_{1_{22}}} & 0 & 0 \\ \hline 0 & 0 & \mu & 0 & \bullet & \boldsymbol{D_{0_{12}}} & \boldsymbol{D_{1_{11}}} & \boldsymbol{D_{1_{12}}} \\ 0 & 0 & 0 & \mu & \boldsymbol{D_{0_{21}}} & \bullet & \boldsymbol{D_{1_{21}}} & \boldsymbol{D_{1_{22}}} \\ \hline & & & & \ddots & \ddots & \ddots & \ddots \\ & & & & \ddots & \ddots & \ddots & \ddots \\ \hline \end{array},$$

from which the block level structure is

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{D_0} & \boldsymbol{D_1} & & & \\ \hline \mu\boldsymbol{I} & \boldsymbol{D_0}{-}\mu\boldsymbol{I} & \boldsymbol{D_1} & & \\ \hline & \mu\boldsymbol{I} & \boldsymbol{D_0}{-}\mu\boldsymbol{I} & \boldsymbol{D_1} & \\ \hline & & \mu\boldsymbol{I} & \boldsymbol{D_0}{-}\mu\boldsymbol{I} & \boldsymbol{D_1} \\ \hline & & & \ddots & \ddots \\ \hline \end{array},$$

where the row sum is zero due to $(\boldsymbol{D_0} + \boldsymbol{D_1})\mathbb{1} = \boldsymbol{0}$. Comparing the QBD process of the $PH/M/1$ queue and the $MAP/M/1$ queues we have that a $PH/M/1$ queue is a special $MAP/M/1$ queue with $\boldsymbol{D_0} = \boldsymbol{T}$ and $\boldsymbol{D_1} = \boldsymbol{t}\boldsymbol{\tau}$.

9.3.2 M/MAP/1 Queue

The consecutive interevent times of a MAP are correlated (in general). That is, the consecutive service times of an $M/MAP/1$ queue are correlated (in general), and it is independent of whether or not the queue is idle after a departure of a customer. Due to this property, the phase of the service process is carried on also when the queue is idle. Consequently, the zero level of the QBD contains the same number of phases as the higher level. This feature of an $M/MAP/1$ queue is similar to that of a $MAP/M/1$ queue but differs significantly from that of an $M/PH/1$ queue.

If the arrival process is a Poisson process at a rate λ and the service process is MAP with representation $\boldsymbol{S_0}, \boldsymbol{S_1}$, then the block level structure of the QBD process is

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline -\lambda\boldsymbol{I} & \lambda\boldsymbol{I} & & & \\ \hline \boldsymbol{S_1} & \boldsymbol{S_0}{-}\lambda\boldsymbol{I} & \lambda\boldsymbol{I} & & \\ \hline & \boldsymbol{S_1} & \boldsymbol{S_0}{-}\lambda\boldsymbol{I} & \lambda\boldsymbol{I} & \\ \hline & & \boldsymbol{S_1} & \boldsymbol{S_0}{-}\lambda\boldsymbol{I} & \lambda\boldsymbol{I} \\ \hline & & & \ddots & \ddots \\ \hline \end{array}.$$

The zero level of this matrix indicates that the service process is "switched off" at the zero level, but the phase of the service MAP is maintained while the queue is idle, and the service MAP resumes its evolution from the same phase when a customer arrives at the system.

In the case of a $PH/M/1$ queue with $(\boldsymbol{\tau}, \boldsymbol{T})$ and a $MAP/M/1$ queue with $\boldsymbol{D_0} = \boldsymbol{T}$ and $\boldsymbol{D_1} = \boldsymbol{t\tau}$, the QBD process of the $PH/M/1$ and $MAP/M/1$ queues are identical because both of them contain J phases at the zero level. The sizes of the zero levels of $M/PH/1$ and $M/MAP/1$ queues differ.

Fortunately, the representation of an $M/PH/1$ queue with $(\boldsymbol{\tau}, \boldsymbol{T})$ (zero level with one phase) as a special $M/PH/1$ queue with $\boldsymbol{S_0} = \boldsymbol{T}$ and $\boldsymbol{S_1} = \boldsymbol{t\tau}$ (zero level with J phases) remains valid with respect to all queue-related parameters computed from the two different QBD processes. The behavior of an $M/PH/1$ queue with $\boldsymbol{S_0} = \boldsymbol{T}$ and $\boldsymbol{S_1} = \boldsymbol{t\tau}$ (zero level with J phases) can be interpreted as that of a customer that leaves the system idle and decides the initial phase of the next service time (independently of the fact that the queue becomes idle); this phase is preserved by the QBD process during the idle time of the queue. In summary, we emphasize that PH arrival and service processes can always be represented as special MAPs.

9.3.3 MAP/PH/1 Queue

If both the arrival and service processes are characterized by a background Markov chain, then the $(N(t), J(t))$ QBD process can still be used for the analysis of the queueing system, but the phase process $J(t)$ must represent the phase of both background Markov chains. That is, the phase process of the QBD process is the Cartesian product of the phase processes of the arrival and service processes. The Markov chain describing the independent evolution of the arrival and service processes can be expressed by Kronecker operators. If the arrival process is a MAP with representation $\boldsymbol{D_0}, \boldsymbol{D_1}$, and the service time is PH distributed with representation $\tau, \boldsymbol{T}$ $(t = -\boldsymbol{T}\mathbb{1})$, then the structure of the generator matrix is

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{D_0} & \boldsymbol{D_1} \otimes \boldsymbol{\tau} & & & \\ \hline \boldsymbol{I} \otimes t & \boldsymbol{D_0} \oplus \boldsymbol{T} & \boldsymbol{D_1} \otimes \boldsymbol{I} & & \\ \hline & \boldsymbol{I} \otimes \boldsymbol{t\tau} & \boldsymbol{D_0} \oplus \boldsymbol{T} & \boldsymbol{D_1} \otimes \boldsymbol{I} & \\ \hline & & \boldsymbol{I} \otimes \boldsymbol{t\tau} & \boldsymbol{D_0} \oplus \boldsymbol{T} & \boldsymbol{D_1} \otimes \boldsymbol{I} \\ \hline & & & \ddots & \ddots \\ \hline \end{array}.$$

That is, $\boldsymbol{F} = \boldsymbol{D_1} \otimes \boldsymbol{I}$, $\boldsymbol{L} = \boldsymbol{D_0} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T} = \boldsymbol{D_0} \oplus \boldsymbol{T}$, $\boldsymbol{B} = \boldsymbol{I} \otimes \boldsymbol{t}\boldsymbol{\tau}$ and $\boldsymbol{F}' = \boldsymbol{D_1} \otimes \tau$, $\boldsymbol{L}' = \boldsymbol{D_0}$, and $\boldsymbol{B}' = \boldsymbol{I} \otimes t$.

9.3.4 MAP/MAP/1 Queue

Similarly, if the arrival process is a MAP with representation $\boldsymbol{D_0}$, $\boldsymbol{D_1}$, and the service process is a MAP with representation $\boldsymbol{S_0}$, $\boldsymbol{S_1}$ then the structure of the generator matrix is

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|} \hline \boldsymbol{D_0} \oplus \boldsymbol{I} & \boldsymbol{D_1} \otimes \boldsymbol{I} & & \\ \hline \boldsymbol{I} \otimes \boldsymbol{S_1} & \boldsymbol{D_0} \oplus \boldsymbol{S_0} & \boldsymbol{D_1} \otimes \boldsymbol{I} & \\ \hline & \boldsymbol{I} \otimes \boldsymbol{S_1} & \boldsymbol{D_0} \oplus \boldsymbol{S_0} & \boldsymbol{D_1} \otimes \boldsymbol{I} \\ \hline & & \ddots & \ddots \\ \hline \end{array}.$$

That is, $\boldsymbol{F} = \boldsymbol{D_1} \otimes \boldsymbol{I}$, $\boldsymbol{L} = \boldsymbol{D_0} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{S_0} = \boldsymbol{D_0} \oplus \boldsymbol{S_0}$, $\boldsymbol{B} = \boldsymbol{I} \otimes \boldsymbol{S_1}$, and $\boldsymbol{L}' = \boldsymbol{D_0} \oplus \boldsymbol{I}$.

9.3.5 MAP/PH/1/K Queue

Finally, we demonstrate that the analysis of finite QBD processes can be used for the analysis of finite buffer queues. For example, if the arrival process is a MAP with representation $\boldsymbol{D_0}$, $\boldsymbol{D_1}$, the service time is PH distributed with representation τ, $\boldsymbol{T}$, and at most K customers can be present in the queue, then the structure of the QBD process describing the queue behavior is

$$\boldsymbol{Q} = \begin{array}{|c|c|c|c|c|} \hline \boldsymbol{L}' & \boldsymbol{F}' & & & \\ \hline \boldsymbol{B}' & \boldsymbol{L} & \ddots & & \\ \hline & \boldsymbol{B} & \ddots & \boldsymbol{F} & \\ \hline & & \ddots & \boldsymbol{L} & \boldsymbol{F} \\ \hline & & & \boldsymbol{B} & \boldsymbol{L}'' \\ \hline \end{array},$$

where $\boldsymbol{F} = \boldsymbol{D_1} \bigotimes \boldsymbol{I}$, $\boldsymbol{L} = \boldsymbol{D_0} \bigoplus \boldsymbol{T}$, $\boldsymbol{B} = \boldsymbol{I} \bigotimes \boldsymbol{t\tau}$, $\boldsymbol{F'} = \boldsymbol{D_1} \bigotimes \boldsymbol{\tau}$, $\boldsymbol{L'} = \boldsymbol{D_0}$, $\boldsymbol{B'} = \boldsymbol{I} \bigotimes \boldsymbol{T}$ and $\boldsymbol{L}^{"} = (\boldsymbol{D_0} + \boldsymbol{D_1}) \bigoplus \boldsymbol{T}$.

9.4 Exercises

Exercise 9.1. Define a MAP representation of the departure process of an $M/M/1/2$ queue with an arrival rate λ and service rate μ.

Exercise 9.2. Define a MAP representation of the departure process of a $MAP/M/1/1$ queue with arrival MAP $(\hat{\boldsymbol{D}}_{\boldsymbol{0}}, \hat{\boldsymbol{D}}_{\boldsymbol{1}})$ and service rate μ.

Exercise 9.3. Define a MAP representation of the customer loss process of a $MAP/M/1/1$ queue with arrival MAP $(\hat{\boldsymbol{D}}_{\boldsymbol{0}}, \hat{\boldsymbol{D}}_{\boldsymbol{1}})$ and service rate μ.

Exercise 9.4. Compute the generator of a CTMC that describes the number of customers and the phase of the arrival PH distribution in a $PH/M/1$ queue if the representation of the PH distributed interarrival time is (α, A), with $\alpha = (1, 0)$ and $A = \begin{pmatrix} -\alpha & \alpha/2 \\ 0 & -\gamma \end{pmatrix}$, and the service rate is μ.

Exercise 9.5. Compute the generator of a CTMC that describes the number of customers and the phase of the service PH distribution in an $M/PH/1$ queue if the arrival rate is λ and the representation of the PH distributed service time is $(\beta, \boldsymbol{B})$, with $\beta = (1/3, 2/3)$ and $\boldsymbol{B} = \begin{pmatrix} -\mu & \mu \\ 0 & -\gamma \end{pmatrix}$.

Exercise 9.6. A packet transmission is performed in two phases in a slotted time communication protocol. The first phase is the resource allocation and the second is the data transmission. The times of both phases are geometrically distributed with the parameters q_1 and q_2. In every time slot one packet arrives with probability p (and no packet arrives with probability $1 - p$). Compute the probability of packet loss if at most two packets can be in the system.

Exercise 9.7. Requests arrive at a computer according to a Poisson process at a rate λ. The service of these requests requires, first, a processor operation for an exponentially distributed amount of time with the parameter μ_1. Following this processor operation the request leaves the system with probability p or requires a consecutive disk operation with probability $1 - p$. The time of the disk operation is exponentially distributed with the parameter μ_2. Following the disk operation the request requires a processor operation because it is a new one. There can be several loops of processor and disk operations. The processor is blocked during the disk operation, and one request is handled at a time.

Compute the efficient utilization of the processor, and compute the request loss probability if there is no buffer in the system.

Compute the efficient utilization of the processor, and compute the system time of the requests if there is an infinite buffer in the system.

Chapter 10
Queueing Networks

10.1 Introduction of Queueing Networks

Up to now, we have overviewed the main methods for the analysis of individual queueing systems. But the analysis of large telecommunication systems or computer systems executing complex interrelated tasks (e.g., transaction processing systems, Web server farms) requires the application of systems models that contain several servers (potentially of different kinds) where customers are traveling among these servers for consecutive services.

Queueing network models are commonly used for the analysis of these kinds of systems. A queueing network is a graph with directed arcs whose nodes represent the kinds of queueing systems that we have studied till now. The arcs of the graph describe the potential transitions of customers among these queueing systems.

It is a commonly applied modeling assumption in queueing networks that the transition of a customer from one node to the next is memoryless and independent of the network state, i.e., it is independent of the past history of the network, the current number of customers at the network nodes, and the status of the servers. After being served at a network node a customer chooses the next node according to the weight (probability) associated with the outgoing arcs of the given node.

There are two main classes of queueing networks: open and closed queueing networks. In closed queueing networks, a fixed number of customers circulate in the network, and there is no arrival/departure from/to the environment. In open queueing networks customers arrive from the environment, obtain a finite number of services at the network nodes (nodes are potentially visited more than once), and leave the network eventually.

Queueing networks are classified also based on the structure of the directed arcs. Queueing networks without a loop (series of directed arcs forming a loop) are referred to as acyclic or feedforward queueing networks, and those with a loop are referred to as cyclic or feedback queueing networks. Acyclic networks are meaningful only in the case of open queueing networks. The nodes of acyclic

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, DOI 10.1007/978-1-4614-5317-8_10,

networks can be numbered such that arcs are always directed from a node with a lower index to a node with a higher index or to the environment. Henceforth we assume that the nodes of acyclic networks are numbered in this way.

10.2 Burke's Theorem

It is possible to analyze a class of open acyclic queueing networks based on the following theorem.

Theorem 10.1 ([17]). *The customer departure process of a stable $M/M/m$ queue is a Poisson process with the same rate as the arrival process of the queue.*

Proof. The number of customers in an $M/M/m$ queue is a *reversible* Markov chain (Sect. 3.3.6). The time reverse of the process is stochastically identical (according to all finite-dimensional joint probabilities) with the original process. In this way the departure instances of the original process (which are the arrival instants of the reverse process) are stochastically identical with the arrival instants of the original process (which are the departure instants of the reverse process) which is a Poisson process. □

An important consequence of the theorem is that in equilibrium the time till the next departure is exponentially distributed, i.e., memoryless.

Let $D^*(s)$ be the Laplace transform of the time till the next departure, $A^*(s)$ the Laplace transform of the interarrival time distribution, $B^*(s)$ the Laplace transform of the service time distribution, and p the probability that in equilibrium the queue will be idle; then

$$D^*(s) = p\ B^*(s) + (1-p)\ A^*(s)\ B^*(s).$$

Using that $B^*(s) = \frac{\mu}{s+\mu}$, $A^*(s) = \frac{\lambda}{s+\lambda}$, $p = \frac{\lambda}{\mu}$, we have

$$D^*(s) = \frac{\lambda}{\mu}\ \frac{\mu}{s+\mu} + \frac{\mu-\lambda}{\mu}\ \frac{\lambda}{s+\lambda}\ \frac{\mu}{s+\mu},$$

and after some algebra

$$D^*(s) = \frac{\mu}{s+\mu}\ \frac{s\lambda + \lambda^2 + \mu\lambda - \lambda^2}{\mu(s+\lambda)} = \frac{\lambda}{s+\lambda}.$$

This expression indicates that we often have exponentially distributed interarrival, interdeparture times in Markovian queueing networks.

10.3 Tandem Network of Two Queues

The simplest queueing network is the open tandem network (Fig. 10.1) composed of two $M/M/1$ queues in which customers arriving from the environment get in queue 1 and after being served in queue 1 get in queue 2, from where, after being served, they depart to the environment. Let the arrival rate from the environment to queue 1 be λ and the service rate at queue 1 and 2 be μ_1 and μ_2, respectively.

From Burke's theorem we have that the arrival intensity to both queues is λ, and in this way the condition of stability is

$$\frac{\lambda}{\mu_1} < 1 \qquad \frac{\lambda}{\mu_2} < 1$$

that is

$$\lambda < \min(\mu_1, \mu_2).$$

Let us consider a Markov chain describing the number of customers in both queues. We identify the states of this Markov chain by a vector of the number of customers in the first queue and the second queue. That is, state $\{i, j\}$ refers to the state where there are i customers in the first and j customers in the second queue. The transition rates of this Markov chain are as follows:

$$\begin{aligned}
&\{i, j\} \to \{i+1, j\} && : \lambda, \\
&\{i, j\} \to \{i-1, j+1\} && : \mu_1 \text{ when } i \geq 1, \\
&\{i, j\} \to \{i, j-1\} && : \mu_2 \text{ when } j \geq 1.
\end{aligned}$$

We denote the stationary probability of state $\{i, j\}$ by $p_{i,j}$. The balance equations of the Markov chains are

$$\begin{cases}
\lambda p_{0,0} &= \mu_2 p_{0,1}, & \\
(\lambda + \mu_2) p_{0,j} &= \mu_1 p_{1,j-1} + \mu_2 p_{0,j+1} & \text{when } j \geq 1, \\
(\lambda + \mu_1) p_{i,0} &= \lambda p_{i-1,0} + \mu_2 p_{i,1} & \text{when } i \geq 1, \\
(\lambda + \mu_1 + \mu_2) p_{i,j} &= \lambda p_{i-1,j} + \mu_1 p_{i+1,j-1} + \mu_2 p_{i,j+1} & \text{when } i, j \geq 1.
\end{cases}$$

According to Burke's theorem, in equilibrium the arrival process of queue 2 is a Poisson process with rate λ. Using this fact the stationary state probabilities are

$$p_{i,j} = p_i^{(1)} p_j^{(2)} = \left(1 - \frac{\lambda}{\mu_1}\right)\left(\frac{\lambda}{\mu_1}\right)^i \left(1 - \frac{\lambda}{\mu_2}\right)\left(\frac{\lambda}{\mu_2}\right)^j,$$

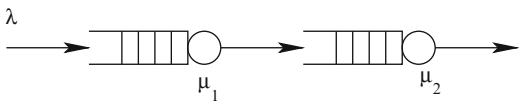


Fig. 10.1 Tandem network of two nodes

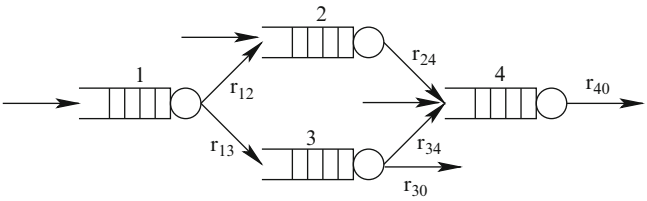


Fig. 10.2 Acyclic queueing network

where $p_i^{(1)}$ and $p_j^{(2)}$ are the stationary distributions of the corresponding $M/M/1$ queues.

Stationary solutions of this kind are referred to as *product-form solution* because the joint distribution is the product of two marginal distributions. It is important to note that despite the product-form stationary distribution the number of customers in the two queues is not independent. There is a very strong correlation between those processes, namely, a departure from the first queue results in an arrival at the second queue.

Based on the stationary distribution we can easily determine the important performance indices. For example, the mean number of customers in the system, the mean time spent in the network, and the mean waiting time spent in the network are

$$\mathbf{E}(N) = \sum_i \sum_j (i+j) p_{i,j} = \sum_i i p_i^{(1)} + \sum_j j p_j^{(2)} = \frac{\frac{\lambda}{\mu_1}}{1-\frac{\lambda}{\mu_1}} + \frac{\frac{\lambda}{\mu_2}}{1-\frac{\lambda}{\mu_2}},$$

$$\mathbf{E}(T) = \frac{\mathbf{E}(N)}{\lambda} = \frac{\frac{1}{\mu_1}}{1-\frac{\lambda}{\mu_1}} + \frac{\frac{1}{\mu_2}}{1-\frac{\lambda}{\mu_2}} = \frac{1}{\mu_1-\lambda} + \frac{1}{\mu_2-\lambda},$$

$$\mathbf{E}(W) = \mathbf{E}(T) - \frac{1}{\mu_1} - \frac{1}{\mu_2},$$

where we used Little's law to obtain the last two quantities.

10.4 Acyclic Queueing Networks

Acyclic queueing networks (Fig. 10.2) are queueing networks in which the outgoing arcs of the nodes are directed toward nodes with a higher index or to the environment. Consequently, in such queueing networks a customer visits each node at most once.

Based on Burke's theorem and the results on the superposition and filtering of independent Poisson processes [Property (h) of Poisson processes in Sect. 2.7.3],

we can apply the same approach as the one applied for the analysis of the tandem queueing network. That is, we can (explicitly) compute the arrival rate to each node of the network, and we can assume that the arrival process at the given node is a Poisson process with that arrival rate. Based on this assumption, the product-form solution remains valid, that is,

$$p_{k_1,k_2,\cdots,k_N} = \prod_{i=1}^{N} p_{k_i}^{(i)},$$

where $p_{k_i}^{(i)}$ is the stationary probability of the k_i state of an M/M/1 queue with a Poisson arrival process with the parameter λ_i and exponentially distributed service time with the parameter μ_i, which is

$$p_{k_i}^{(i)} = \left(1 - \frac{\lambda_i}{\mu_i}\right)\left(\frac{\lambda_i}{\mu_i}\right)^{k_i}.$$

10.5 Open, Jackson-Type Queueing Networks

In the previous subsections we discussed acyclic queueing networks and, based on Burke's theorem, we assumed that the arrival processes of the queues were independent Poisson processes. Based on this assumption we obtained product-form solutions. From now on we consider cyclic queueing networks and consequently we can no longer apply Burke's theorem due to the dependencies on the arrival processes of customers at a queue.

The main results of this kind of queueing networks were published by Jackson [44] in 1963. Since then, these kinds of networks have often been referred to as Jackson-type networks (Fig. 10.3). Jackson considered the following queueing network model:

- The network is composed of N nodes.
- There are m_i servers at node i.

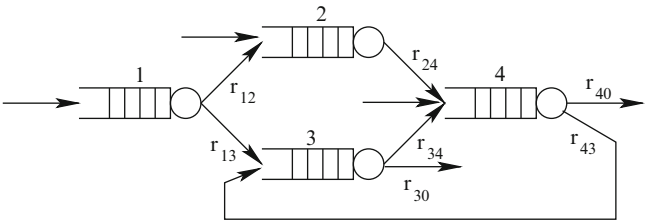


Fig. 10.3 Jackson-type queueing network

- The service time distribution at node i is exponentially distributed with the parameter μ_i.
- From the environment customers arrive at node i according to a Poisson process at rate γ_i.
- A customer getting served at node i goes to node j with probability $r_{i,j}$ $(i, j = 1, 2, \cdots, N)$, and the probability that the customer departs from the network is

$$r_{i,0} = 1 - \sum_{k=1}^{N} r_{i,k} \qquad i, j = 1, 2, \cdots, N.$$

Stability Condition of Jackson-Type Queueing Networks

The following *traffic equations* define the traffic rate at the nodes of the network:

$$\lambda_i = \gamma_i + \sum_{j=1}^{N} \lambda_j \, r_{j,i} \quad i = 1, 2, \cdots, N. \tag{10.1}$$

The left-hand side of the equation represents the aggregate traffic intensity arriving at node i. Due to the stability of the network nodes, the arriving traffic intensity is identical with the departing traffic intensity from node i. The right-hand side of the equation gives the traffic components arriving at node i. γ_i is the traffic component arriving from the environment, and $\lambda_j \, r_{j,i}$ is the traffic component that departs from node j and goes to node i.

Introducing the row vector $\lambda = \{\lambda_i\}$ and $\gamma = \{\gamma_i\}$ and matrix $\boldsymbol{R} = \{r_{ij}\}$ the traffic equation can be written in the following vector form:

$$\lambda = \gamma + \lambda \boldsymbol{R},$$

whence

$$\lambda = \gamma (\boldsymbol{I} - \boldsymbol{R})^{-1}$$

if $(\boldsymbol{I} - \boldsymbol{R})$ is nonsingular.

The elements of the matrix $(\boldsymbol{I} - \boldsymbol{R})^{-1}$ have a well-defined physical interpretation according to the following theorem. Let L_{ij} denote the number of visits to node j (before departing to the environment) by a customer arriving at node i:

Theorem 10.2.

$$\left[(\boldsymbol{I} - \boldsymbol{R})^{-1}\right]_{i,j} = \mathbf{E}\left(L_{i,j}\right),$$

where the left-hand side denotes the i, j element of the matrix $(\boldsymbol{I} - \boldsymbol{R})^{-1}$.

Proof. The number of visits to node j satisfies the following equation:

$$\mathbf{E}\left(L_{i,j}\right) = \delta_{i,j} + \sum_{k=1}^{N} r_{i,k} \mathbf{E}\left(L_{k,j}\right),$$

where $\delta_{i,j}$ is the Kronecker delta, that is, $\delta_{i,j} = 1$ if $i = j$, 0 otherwise. Introducing matrix $\boldsymbol{L}$ whose i, j element is $\mathbf{E}\left(L_{i,j}\right)$ we can rewrite the preceding equation in matrix form:

$$\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{R}\boldsymbol{L},$$

from which the theorem comes. □

The theorem gives a condition for the nonsingularity of the matrix $(\boldsymbol{I} - \boldsymbol{R})$. $(\boldsymbol{I} - \boldsymbol{R})$ is nonsingular if all customers leave the queueing network after a finite number of visits to the nodes of the network.

A queueing network is said to be stable if all queues are stable, which holds when

$$\lambda_i < m_i \mu_i, \qquad i = 1, 2, \cdots, N.$$

Stationary Distribution of Jackson-Type Queueing Networks

According to the properties of Jackson-type queueing networks, the number of customers at the nodes of the network is a continuous-time Markov chain. Let k_i denote the number of customers at node i, and let us introduce the following notations:

$$\begin{aligned}
\mathbf{N} &= (k_1, \cdots, k_i, \cdots, k_j, \cdots, k_N), \\
\mathbf{N}_{i,0} &= (k_1, \cdots, k_i + 1, \cdots, k_j, \cdots, k_N), \\
\mathbf{N}_{0,j} &= (k_1, \cdots, k_i, \cdots, k_j - 1, \cdots, k_N), \\
\mathbf{N}_{i,j} &= (k_1, \cdots, k_i + 1, \cdots, k_j - 1, \cdots, k_N),
\end{aligned}$$

where in the last two cases $k_j \geq 1$. Using these notations we can describe the possible transitions of Markov chains representing the number of customers at the network nodes.

- $\mathbf{N}_{0,j} \rightarrow \mathbf{N}$: a new customer arrives at node j from the environment, increasing the number of customers at node j from $k_j - 1$ to k_j. This happens at rate γ_j.
- $\mathbf{N}_{i,0} \rightarrow \mathbf{N}$: a customer departs to the environment from node j, decreasing the number of customers at node j from $k_j + 1$ to k_j. This happens at rate $r_{i,0}\alpha_i(k_i + 1)\mu_i$.
- $\mathbf{N}_{i,j} \rightarrow \mathbf{N}$: a customer gets served at node i and goes to node j. This transition decreases the number of customers at node i from $k_i + 1$ to i_j and increases the number of customers at node j from $k_j - 1$ to k_j. This happens at rate $r_{i,j}\alpha_i(k_i + 1)\mu_i$.

In the preceding expressions $\alpha_i(k_i) = \min\{k_i, m_i\}$ defines the coefficient of the service rate of node i when there are k_i customers at the node. When there are more customers at the node than servers, then all servers are working and the service rate is $m_i\mu_i$; when there are fewer customers than servers, then there are idle servers and the service rate is $k_i\mu_i$.

Theorem 10.3. *A Markov chain characterized by the previously defined state transitions has a product-form stationary distribution, that is,*

$$p_{\mathbf{N}} = p_{k_1,\cdots,k_N} = p_{k_1}^{(1)} p_{k_2}^{(2)} \cdots p_{k_N}^{(N)}, \tag{10.2}$$

where $p_{k_i}^{(i)}$ is the stationary distribution of an M/M/m_i queue with a Poisson arrival process at rate λ_i and exponentially distributed service time with the parameter μ_i. The stationary probabilities of such queues are given as a function of $p_0^{(i)}$:

$$p_{k_i}^{(i)} = \begin{cases} p_0^{(i)} \left(\dfrac{\lambda_i}{\mu_i}\right)^{k_i} \dfrac{1}{k_i!} & 0 \le k_i \le m_i, \\ p_0^{(i)} \left(\dfrac{\lambda_i}{\mu_i}\right)^{k_i} \dfrac{1}{m_i!} m_i^{m_i - k_i}, & k_i \ge m_i \end{cases} \tag{10.3}$$

and $p_0^{(i)}$ can be obtained from the normalizing equation $\sum_{k_i=0}^{\infty} p_{k_i}^{(i)} = 1$.

Proof. Based on the possible state transitions of a Markov chain, the balance equation of state $\mathbf{N}$ is as follows:

$$\begin{aligned} p_{\mathbf{N}} \left(\sum_{i=1}^{N} \gamma_i + \sum_{i=1}^{N} \alpha_i(k_i)\, \mu_i \right) &= \sum_{i=1}^{N} p_{\mathbf{N}_{i,0}} \alpha_i(k_i + 1)\, \mu_i\, r_{i,0} \\ &+ \sum_{j=1}^{N} p_{\mathbf{N}_{0,j}}\, \gamma_j\, \mathcal{I}_{\{k_j>0\}} + \sum_{i=1}^{N} \sum_{j=1}^{N} p_{\mathbf{N}_{i,j}}\, \alpha_i(k_i + 1)\, \mu_i\, r_{i,j}\,, \end{aligned} \tag{10.4}$$

where $\mathcal{I}_{\{k_j>0\}}$ is the indicator of $k_j > 0$, i.e., $\mathcal{I}_{\{k_j>0\}} = 1$ if $k_j > 0$ and $\mathcal{I}_{\{k_j>0\}} = 0$ otherwise.

The left-hand side of the equation is the rate at which the process departs from state $\mathbf{N}$ in equilibrium. It contains the state transitions due to a new customer arrival from the environment and due to a service completion. The right-hand side of the equation is the rate at which the process moves to state $\mathbf{N}$ in equilibrium. This can happen due to a service of a queue from which the customer leaves the network, due to the arrival of a new customer from the environment, or due to a service completion at node i from where the customer moves to node j.

If $\gamma_i > 0$ and $\mu_i > 0$, then the Markov chain is irreducible, the solution of the stationary equation is unique, and it is enough to show that the product-form solution (10.2) satisfies the balance Eq. (10.4). First we substitute the product-form solution into the right-hand side of the balance equation and use the fact that from Eq. (10.3) we have $p_{k_i+1}^{(i)} = p_{k_i}^{(i)} \frac{\lambda_i}{\mu_i \alpha_i(k_i+1)}$ and $p_{k_i-1}^{(i)} = p_{k_i}^{(i)} \frac{\mu_i \alpha_i(k_i)}{\lambda_i}$. We obtain that

$$
\begin{aligned}
&\sum_{i=1}^{N} p_{k_1}^{(1)} \cdots p_{k_i+1}^{(i)} \cdots p_{k_N}^{(N)} \, \alpha_i(k_i+1) \, \mu_i \, r_{i,0} \\
&\quad + \sum_{j=1}^{N} p_{k_1}^{(1)} \cdots p_{k_j-1}^{(j)} \cdots p_{k_N}^{(N)} \, \gamma_j \, I_{k_j>0} \\
&\quad + \sum_{i=1}^{N} \sum_{j=1}^{N} p_{k_1}^{(1)} \cdots p_{k_i+1}^{(i)} \cdots p_{k_j-1}^{(j)} \cdots p_{k_N}^{(N)} \, \alpha_i(k_i+1) \, \mu_i \, r_{i,j} \\
&= p_{k_1}^{(1)} \cdots p_{k_N}^{(N)} \left(\sum_{i=1}^{N} \lambda_i \, r_{i,0} + \sum_{j=1}^{N} \frac{\mu_j \alpha_j(k_j)}{\lambda_j} \, \gamma_j + \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\mu_j \alpha_j(k_j)}{\lambda_j} \, \lambda_i \, r_{i,j} \right) \\
&= p_{k_1}^{(1)} \cdots p_{k_N}^{(N)} \left(\sum_{i=1}^{N} \lambda_i \, r_{i,0} + \sum_{j=1}^{N} \frac{\mu_j \alpha_j(k_j)}{\lambda_j} \, \gamma_j + \sum_{j=1}^{N} \frac{\mu_j \alpha_j(k_j)}{\lambda_j} \underbrace{\sum_{i=1}^{N} \lambda_i \, r_{i,j}}_{\lambda_j - \gamma_j} \right) \\
&= p_{k_1}^{(1)} \cdots p_{k_N}^{(N)} \left(\sum_{i=1}^{N} \lambda_i \, r_{i,0} + \sum_{j=1}^{N} \mu_j \alpha_j(k_j) \right) \\
&= p_{k_1}^{(1)} \cdots p_{k_N}^{(N)} \left(\sum_{i=1}^{N} \gamma_i + \sum_{j=1}^{N} \mu_j \alpha_j(k_j) \right) . \qquad (10.5)
\end{aligned}
$$

In the third step of the derivation we used the traffic equation of queue j, Eq. (10.1), and in the fourth step we utilized that the intensity of customer arrivals from the environment $\sum_{i=1}^{N} \gamma_i$ is identical to the intensity of customer departures to the environment, $\sum_{i=1}^{N} \lambda_i \, r_{i,0}$, in equilibrium.

The obtained expression is the left-hand side of the balance equation assuming a product-form solution of the stationary distribution. □

There might be loops in a Jackson-type queueing network of which the arrival processes of the nodes are not independent Poisson processes and to which Burke's theorem is not applicable. Consequently, in this case we obtain a product-form solution despite the queues' dependent input processes. The reverse reasoning cannot be applied. The product-form solution has no implications for the dependencies of the arrival processes of the queues.

Traffic Theorem for Open Queueing Networks

Jackson-type queueing networks possess a traffic property similar to the PASTA (Poisson arrival sees time average) property of queueing systems with a Poisson arrival process.

Theorem 10.4. *The distribution of the number of customers in the queues at the arrival instants of node j is identical to the stationary distribution of the number of customers in the queues.*

Proof. We define an extended queueing network that contains one additional single-server node, node 0, with respect to the original queueing network. The traffic matrix is also similar to the original one. It is modified only such that customers going to node j are driven to node 0 and from node 0 to node j. The rest of the traffic matrix is unchanged. The extended queueing network is also of a Jackson type, and consequently its stationary distribution is product form: $p_{\mathbf{N}'} = p_{k_0}^{(0)} p_{k_1}^{(1)} p_{k_2}^{(2)} \cdots p_{k_N}^{(N)}$.

The service rate of node 0 is μ_0. As $\mu_0 \to \infty$, the behavior of the extended queueing network becomes identical to that of the original and the arrival instants of node j are the instants when there is one customer in node 0. In this way the distribution of the customers at an arrival instants of node j is

$$\begin{aligned}\mathbf{P}\left(K_1 = k_1, \cdots, K_N = k_N | K_0 = 1\right) &= \frac{\mathbf{P}\left(K_0 = 1, K_1 = k_1, \cdots, K_N = k_N\right)}{\mathbf{P}\left(K_0 = 1\right)} \\ &= p_{\mathbf{N}}.\end{aligned}$$

□

This theorem is important for computing the delays in a queueing system.

10.6 Closed, Gordon–Newell-Type Queueing Networks

The analysis of the closed queueing network counterpart of Jackson-type queueing networks was first published by Gordon and Newell in 1967 [40]. Since that time, this kind of queueing network has often carried their name. The node behavior of Gordon–Newell-type queueing networks is identical to that of Jackson-type networks. At node i there are m_i servers with exponentially distributed service time with parameters μ_i and an infinite buffer.

In contrast to the Jackson-type networks, there is no arrival from or departure to the environment in closed queueing networks. Thus, the number of customers in the network is constant, denoted by K. If k_i denotes the number of customers at node i, then in each state of the network we have

$$\sum_{i=1}^{N} k_i = K.$$

As with the Jackson-type network, the number of customers at the nodes of the network form a Markov chain. In a closed queueing network the only possible state transition in this Markov chain is the $\mathbf{N}_{i,j} \to \mathbf{N}$ transition, that is, a customer gets served at node i and moves to node j; the transition rate of this state transition is

$\alpha_i(k_i + 1)\mu_i r_{i,j}$. This state transition decreases the number of customers at node i from $k_i + 1$ to k_i and increases the number of customers at node j from $k_j - 1$ to k_j.

The aggregate arrival rate of the nodes are characterized by the traffic equation

$$\lambda_i = \sum_{j=1}^{N} \lambda_j \, r_{j,i} \quad i = 1, 2, \cdots, N. \tag{10.6}$$

Equation (10.6) indicates that customers arriving at node i are those customers that departed from node j and were directed to node i with probability r_{ij}. In a closed queueing network, $\sum_{j=1}^{N} r_{ij} = 1$ since there is no departure to the environment. The solution of the traffic equation of closed queueing networks is not unique. Multiplying an arbitrary solution by a constant gives another solution of the traffic equation.

Theorem 10.5. *The stationary distribution of the number of customers in a Gordon–Newell-type queueing network has product form. That is,*

$$p_{\mathbf{N}} = p_{k_1, \cdots, k_N} = \frac{1}{G} \prod_{i=1}^{N} h_{k_i}^{(i)}, \tag{10.7}$$

where λ_i is an arbitrary nonzero solution of the traffic equation,

$$h_{k_i}^{(i)} = \begin{cases} \left(\dfrac{\lambda_i}{\mu_i}\right)^{k_i} \dfrac{1}{k_i!} & 0 \le k_i \le m_i, \\ \left(\dfrac{\lambda_i}{\mu_i}\right)^{k_i} \dfrac{1}{m_i!} \, m_i^{m_i - k_i} & k_i \ge m_i, \end{cases} \tag{10.8}$$

and $G = \sum_{\mathbf{N}} \prod_{i=1}^{N} h_{k_i}^{(i)}$.

Proof. The proof follows the same pattern as that for the Jackson-type network. The balance equation for $\mathbf{N}$ is

$$p_{\mathbf{N}} \left(\sum_{i=1}^{N} \alpha_i(k_i) \, \mu_i \right) = \sum_{i=1}^{N} \sum_{j=1}^{N} p_{\mathbf{N}_{i,j}} \, \alpha_i(k_i + 1) \, \mu_i \, r_{i,j}, \tag{10.9}$$

where the left-hand side of the equation is the rate at which state $\mathbf{N}$ is left and the right-hand side is the rate at which state $\mathbf{N}$ is entered in equilibrium. Due to the irreducibility of a Markov chain, we assume a unique solution of the balance equations (together with the normalizing equation, $\sum_{\mathbf{N} \in \mathcal{S}} p_{\mathbf{N}} = 1$), and we only show that the product form satisfies the balance equation.

Substituting the product form into the right-hand side of the balance equation gives

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{N} p_{k_1}^{(1)}\cdots p_{k_i+1}^{(i)}\cdots p_{k_j-1}^{(j)}\cdots p_{k_N}^{(N)}\,\alpha_i(k_i+1)\,\mu_i\,r_{i,j}\\
&\quad= p_{k_1}^{(1)}\cdots p_{k_N}^{(N)}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\mu_j\alpha_j(k_j)}{\lambda_j}\,\lambda_i\,r_{i,j}\right)\\
&\quad= p_{k_1}^{(1)}\cdots p_{k_N}^{(N)}\left(\sum_{j=1}^{N}\frac{\mu_j\alpha_j(k_j)}{\lambda_j}\,\underbrace{\sum_{i=1}^{N}\lambda_i\,r_{i,j}}_{\lambda_j}\right)\\
&\quad= p_{k_1}^{(1)}\cdots p_{k_N}^{(N)}\left(\sum_{j=1}^{N}\mu_j\alpha_j(k_j)\right), \qquad (10.10)
\end{aligned}$$

which is identical to the left-hand side of the balance equation when the product-form solution is assumed. The normalizing constant, G, ensures that the normalizing equation is satisfied. □

The main difficulties of the analysis of closed queueing networks are that the solution of the traffic equation is not unique and that the normalizing constant cannot be computed in a node-based manner only for the whole network. The computation of G requires the evaluation of all system states, which gets very high even for reasonably small networks. When there are N nodes and K customers in a network, the number of system states is $\binom{N+K-1}{K}$ (e.g., for $N = 10, K = 25$ there are 52,451,256 states).

The commonly applied solution of the first problem is to add an additional equation to the set of traffic equations, $\lambda_1 = 1$, which makes its solution unique.

The second problem, the computation of the normalizing constant, G, is a real research challenge. Many proposals exist for computing the normalizing constant efficiently. Here we summarize the convolution algorithm [18] and the mean value analysis (MVA) algorithm [79].

Convolution Algorithm

The convolution algorithm was first published by Buzen [18]. In the original paper the nodes have a single server, but it is easy to extend the algorithm to Gordon–Newell-type queueing networks where the node i has m_i ($m_i \geq 1$) servers and an infinite buffer. We present the more general version of the algorithm.

Assuming that there are n nodes and k customers in the network, let the assumed normalizing constant be

$$g(k,n) = \sum_{(k_1,\ldots,k_n),\sum_j k_j=k} \prod_{i=1}^{n} h_{k_i}^{(i)},$$

and $g(0,n) = 1$. When $g(k,n)$ is known, we obtain the normalizing constant of the network with N nodes and K customers as $G = \sum_{\mathbf{N}} \prod_{i=1}^{N} h_{k_i}^{(i)} = g(K,N)$.

The following formula allows one to determine $g(k,n)$ in a recursive manner:

$$g(k,n) = \begin{cases} h_k^{(1)} & \text{ha } n = 1, \\ \sum_{j=0}^{k} h_j^{(n)} g(k-j, n-1) & \text{ha } n > 1. \end{cases} \tag{10.11}$$

In the case of one node ($n = 1$) and $k \geq 1$ customers, the recursive formula gives $h_k^{(1)}$, and in the case of more than one nodes we have

$$\begin{aligned} g(k,n) &= \sum_{(k_1,\ldots,k_n),\sum_j k_j=k} \prod_{i=1}^{n} h_{k_i}^{(i)} \\ &= \sum_{(k_1,\ldots,k_n),\sum_j k_j=k,k_n=0} h_0^{(n)} \prod_{i=1}^{n-1} h_{k_i}^{(i)} + \ldots \\ &\quad + \sum_{(k_1,\ldots,k_n),\sum_j k_j=k,k_n=k} h_k^{(n)} \prod_{i=1}^{n-1} h_{k_i}^{(i)} \\ &= h_0^{(n)} g(k,n-1) + \ldots + h_k^{(n)} g(0,n-1). \end{aligned}$$

This expression relates the normalizing constant of a network with n nodes to the normalizing constant of a network with $n-1$ nodes.

The convolution algorithm starts from $n = 1$, $k = 1,\ldots,K$, and increases n to N step by step according to Eq. (10.11). The computational complexity of this algorithm is proportional to N and K^2 [denoted by $O(NK^2)$], and its memory complexity is proportional to K [denoted by $O(K)$].

Another benefit of the convolution algorithm is that some interesting performance parameters are closely related to the $g(k,n)$ parameters. For example, the probability that there are ℓ customers in queue k is

$$\mathbf{P}(k_\ell = k) = \sum_{(k_1,\ldots,k_n),\sum_j k_j=K,k_\ell=k} \frac{1}{G} \prod_{i=1}^{n} h_{k_i}^{(i)} = h_k^{(\ell)} \frac{g(K-k, N-1)}{g(K,N)},$$

and from this the utilization of node ℓ is

$$U_\ell = 1 - \mathbf{P}(k_\ell = 0) = 1 - h_0^{(\ell)} \frac{g(K, N-1)}{g(K, N)}.$$

Traffic Theorem for Closed Queueing Networks

The MVA algorithm is based on the traffic theorem for closed queueing networks, so we present the theorem first.

Theorem 10.6. *In a closed Gordon–Newell-type queueing network containing K customers, the distribution of the number of customers upon a customer's arrival at node j is identical to the stationary distribution of the same network with $K-1$ customers.*

Proof. The proof is practically identical to that provided for open queueing networks. We extend the network with a single-server node 0 and redirect all customers going to node j to node 0 and from node 0 all customers go to node j. The rest of the network is left unchanged. The extended network is of a Gordon–Newell type as well; thus it has a product-form stationary distribution, $p_{k_0,k_1,\ldots,k_N,\sum_{i=0}^N k_i = K} = \frac{1}{G'} \prod_{i=0}^N h_{k_i}^{(i)}$.

The service rate of node 0 is μ_0. As $\mu_0 \to \infty$, the behavior of the extended network and that of the original networks are identical, and the arrival instances of node j are the instances when the number of customers in node 0 is 1. Thus,

$$\mathbf{P}\left(K_1 = k_1, \cdots, K_N = k_N, \sum_{i=0}^N k_i = K | K_0 = 1\right)$$

$$= \frac{\mathbf{P}\left(K_0 = 1, K_1 = k_1, \cdots, K_N = k_N, \sum_{i=0}^N k_i = K\right)}{\mathbf{P}(K_0 = 1)}$$

$$= \mathbf{P}\left(K_1 = k_1, \cdots, K_N = k_N, \sum\nolimits_{i=1}^N k_i = K - 1\right).$$

□

MVA Algorithm

In the convolution algorithm, the number of nodes increases in an iteration of the algorithm. The MVA algorithm is a kind of counterpart of the convolution algorithm in the sense that the MVA algorithm is also an iterative algorithm, but in this case

the number of customers increases in an iteration step. According to this approach, we analyze the involved quantities as a function of the number of customers in the network.

In contrast with the convolution algorithm, the applicability of the MVA algorithm is limited to the case of single servers at the network nodes, i.e., $m_i = 1, i = 1, \ldots, N$, and the algorithm yields mean performance measures, hence its name.

The mean time a customer spends at node i during a visit to node i is

$$\mathbf{E}\left(T_i(K)\right) = \left(1 + \mathbf{E}\left(N_i^*(K)\right)\right)\frac{1}{\mu_i},$$

where $\mathbf{E}\left(N_i^*(K)\right)$ denotes the mean number of customers present at node i upon the arrival of an observed customer. According to the traffic theorem, $\mathbf{E}\left(N_i^*(K)\right)$ is identical to the stationary number of customers at node i when the number of customers in the network is $K-1$, i.e., $\mathbf{E}\left(N_i(K-1)\right)$, whence

$$\mathbf{E}\left(T_i(K)\right) = \left(1 + \mathbf{E}\left(N_i(K-1)\right)\right)\frac{1}{\mu_i}.$$

On the other hand, the mean number of customers at node i in equilibrium is

$$\mathbf{E}\left(N_i(K)\right) = K\,\frac{\lambda_i \mathbf{E}\left(T_i(K)\right)}{\sum_{j=1}^{N}\lambda_j \mathbf{E}\left(T_j(K)\right)}$$

because the arrival rate at node i is proportional to an arbitrary nonzero solution of the traffic equation $\hat{\lambda}_i = \lambda_i c$, according to Little's law $\mathbf{E}\left(N_i(K)\right) = \hat{\lambda}_i \mathbf{E}\left(T_i(K)\right)$ and

$$K\,\frac{\lambda_i \mathbf{E}\left(T_i(K)\right)}{\sum_{j=1}^{N}\lambda_j \mathbf{E}\left(T_j(K)\right)} = K\,\frac{\hat{\lambda}_i \mathbf{E}\left(T_i(K)\right)}{\sum_{j=1}^{N}\hat{\lambda}_j \mathbf{E}\left(T_j(K)\right)} = K\,\frac{\mathbf{E}\left(N_i(K)\right)}{\sum_{j=1}^{N}\mathbf{E}\left(N_j(K)\right)}$$
$$= K\,\frac{\mathbf{E}\left(N_i(K)\right)}{K} = \mathbf{E}\left(N_i(K)\right).$$

Applying Little's law to another time we obtain

$$\hat{\lambda}_i = \frac{\mathbf{E}\left(N_i(K)\right)}{\mathbf{E}\left(T_i(K)\right)} = K\,\frac{\lambda_i}{\sum_{j=1}^{N}\lambda_j \mathbf{E}\left(T_j(K)\right)}.$$

With these expressions we have all the ingredients of the iterative algorithm:

Initial value:

$$\mathbf{E}\left(N_i(0)\right) = 0;$$

Iteration step:

$$\mathbf{E}\left(T_i(K)\right) = \left(1 + \mathbf{E}\left(N_i(K-1)\right)\right)\frac{1}{\mu_i},$$

$$\mathbf{E}\left(N_i(K)\right) = K\,\frac{\lambda_i \mathbf{E}\left(T_i(K)\right)}{\sum_{j=1}^{N} \lambda_j \mathbf{E}\left(T_j(K)\right)};$$

Closing step:

$$\hat{\lambda}_i = \frac{\mathbf{E}\left(N_i(K)\right)}{\mathbf{E}\left(T_i(K)\right)}.$$

The computational complexity and memory complexity of the algorithm are $O(KN^2)$ and $O(N)$. Compared to the convolution algorithm the MVA is more efficient when K is larger than N.

10.7 BCMP Networks: Multiple Customer and Service Types

The Jackson-type and Gordon–Newell-type queueing networks have a product-form stationary distribution. Thus, efficient computational methods are applicable for the analysis of systems modeled by this kind of network. For a long time, the performance analysis and the development of efficient computer systems were based on these kinds of simple and computable models. The analysis of increasingly complex system behavior required the introduction of more complex queueing behavior and the analysis of the obtained queueing network models. This resulted in fertile research in an effort to find the most general set of queueing networks with a product-form stationary distribution. The results of this effort are summarized in [9], and the set of most general queueing networks with a product-form solution is commonly referred to as BCMP networks, whose abbreviation comes from the initials of the coauthors: Baskett, Chandy, Muntz, and Palacios [9].

The set of BCMP networks generalizes the previous queueing networks in two main directions. In the previously discussed queueing networks, customers are indistinguishable and the service discipline is first come, first served (FCFS). In BCMP networks, customers belong to customer classes that are distinguished by the system because customers of different classes might arrive from the environment at the nodes at different rates, might obtain different services (service time distribution and service discipline) at the nodes, and might follow a different traffic routing probability upon completion of a service. Still, customers of the same class are indistinguishable.

The arrival of class r customers at node i occurs at rate γ_{ir}. When a class r customer is rendered a service at node i, the customer gets in the queue at node j as a class s customer with probability $P_{ir,js}$, i.e., customers might change their class right after the completion of a service. Let the number of customer classes be C. Then

$$\sum_{j=0}^{N} \sum_{s=1}^{C} P_{ir,js} = 1, \quad \forall i = 1, \ldots, N, \; r = 1, \ldots, C,$$

$P_{ir,0s}$ denotes the probability of departure to the environment.

A wide range of traffic models can be defined with an appropriate setting of the arrival rate γ_{ir} and traffic routing probability $P_{ir,js}$. Some examples are listed below.

- Customer classes are independent, and some classes behave as in open queueing networks and others as in closed queueing networks: $P_{ir,js} = 0$ if $r \neq s$, i.e., there is no class change. $\gamma_{ir} = 0$ if $r \leq C_z$, and for all $r > C_z$ there exists i such that $\gamma_{ir} > 0$, i.e., the first C_z classes of customers behave as in closed queueing networks and the rest as in open ones. The probability of departure to the environment is as follows, $P_{ir,0s} = 0$ for $r \leq C_z$, and for all $r > C_z$ there exists i such that $P_{ir,0s} > 0$.
- Background traffic at a subset of the network: Let $\gamma_{ir} = 0$ if $i > N_z$, $r \leq C_z$, and $P_{ir,js} = 0$ if $i \leq N_z$, $j > N_z$, $r, s \leq C_z$. In this case the class $r \leq C_z$ customers load only node $i \leq N_z$ and form a kind of background traffic for customers of class $r > C_z$ in that part of the network.
- Multiple service at a node: Customer classes can be used to obtain a fixed number of services, u, at node i during a single visit to node i by customers of class v. For example, if for $r = v, \ldots, v + u - 2$ we let $P_{ir,js} = 1$ if $s = r + 1$, $j = i$, and $P_{ir,js} = 0$ otherwise, and for $r = v + u - 1$ we let $P_{ir,js} \geq 0$ if $s = r$, $j \neq i$, and $P_{ir,js} = 0$ otherwise, then we have the following behavior. A class v customer arrives at node i and gets served sooner as a class v customer than as a class $v + 1$ customer and so on, while it departs as a class $v + u - 1$ customer from node i and goes to node j as a class v customer.

The service disciplines at a node of a BCMP network can be one of the following disciplines:

1. FCFS (first come, first served): Customers arrive at the server in the same order in which they arrived at the node. With this service discipline the service time of all customers is exponentially distributed with the same parameter, which is common to all customer classes. The service intensity might depend on the number of all customers at the node.
2. Processor sharing (PS): In this case, the service capacity of the server is divided into as many equal parts as there are customers at the node, and each part of the server capacity is assigned to a customer. That is, when there are n customers at the node, all of them are served by a $1/n$ portion of the full service capacity. In this case (if there are n customers at the node during the complete service of a customer), the service time of the customer is n times longer than it would

have been had the full service capacity been assigned to this customer. With this service discipline the service time distribution of different customer classes might be different and can be more general than exponentially distributed. Service time distributions with rational Laplace transforms (matrix exponential distributions) are allowed in this case.

3. LCFS–PR (last come first served–preemptive resume): The server serves one customer at a time, but in such a way that the last arrived customer interrupts the service of the customer currently being served (if any) and starts being served. If during this customer's service time a new customer arrives, the first customer is interrupted and waits while all of the customers arriving later get served. At this point, the first cusomter goes to the server again and resumes the service process starting at the point at which it was interrupted.

   Similar to the PS case, with this service discipline the service time distribution of different customer classes might be different and can be more general than exponentially distributed. Service time distributions with rational Laplace transforms (matrix exponential distributions) are allowed with this service discipline.
4. Infinite server (IS): There are infinitely many servers in this service discipline, and thus all arriving customers go to an idle server upon arrival. Similar to the PS and LCFS–PR cases, with this service discipline the service time distributions of different customer classes might be different and can be more general than exponentially distributed. Service time distributions with rational Laplace transforms (matrix exponential distributions) are allowed with this service discipline.

With the introduction of customer classes, the traffic equation only slightly modifies,

$$\lambda_{ir} = \gamma_{ir} + \sum_{j=1}^{N} \sum_{s=1}^{C} \lambda_{js} \, P_{js,ir}, \tag{10.12}$$

but to describe the product-form solution of BCMP networks, we need to introduce further cumbersome notations. To avoid this, we restrict our attention to exponentially distributed service times instead of matrix exponentially distributed ones, but we allow all other generalizations of BCMP service disciplines.

Let N_{ir} denote the number of class r customers at node i and define the vectors $\mathbf{N_i} = \{N_{i1}, \ldots, N_{iC}\}$ and $\mathbf{N} = \{\mathbf{N_1}, \ldots \mathbf{N_N}\}$. Thus, vector $\mathbf{N}$ defines the distribution of the different classes of customers at the network nodes. With this notation the stationary distribution has the form

$$p_{\mathbf{N}} = \frac{1}{G} \prod_{i=1}^{N} h^{(i)}_{\mathbf{N_i}}, \tag{10.13}$$

where

$$h_{\mathbf{N_i}}^{(i)} = \begin{cases} \dfrac{N_i!}{\mu_i^{N_i}} \displaystyle\prod_{r=1}^{C} \frac{1}{N_{ir}!} \lambda_{ir}^{N_{ir}} & \text{if node } i \text{ is FCFS type,} \\ N_i! \displaystyle\prod_{r=1}^{C} \frac{1}{N_{ir}!} \left(\frac{\lambda_{ir}}{\mu_{ir}}\right)^{N_{ir}} & \text{if node } i \text{ is PS or IS type,} \\ \displaystyle\prod_{r=1}^{C} \frac{1}{N_{ir}!} \left(\frac{\lambda_{ir}}{\mu_{ir}}\right)^{N_{ir}} & \text{if node } i \text{ is LCFS-PR type,} \end{cases}$$

and $N_i = \sum_{r=1}^{C} N_{ir}$. μ_{ir} denotes the service rate of a class r customer at node i.

10.8 Non-Product-Form Queueing Networks

Despite the fact that BCMP networks allow for a wide range of node behaviors, there are practical examples whose stationary solutions do not exhibit product-form solutions. The most common reasons for non-product-form solutions are

- Non-Poisson customer arrival process,
- Different exponentially distributed service time at FCFS-type node for different customer classes,
- Nonexponentially distributed service time at FCFS-type node,
- Nonmatrix exponentially distributed service time,
- Queueing nodes with finite buffer.

In general queueing networks, the stochastic behavior of the number of (different classes of) customers at the nodes is not a Markov chain (e.g., in the case of general interarrival or service time distributions). There are also cases where the number of (different classes of) customers at the nodes is a Markov chain but the stationary solution of this Markov chain does not possess product form (e.g., in the case of a Poisson arrival process and exponentially distributed service time distributions and finite-capacity FCFS-type nodes). In these cases no exact analysis methods are available, and we must resort to approximate analysis methods.

The majority of the approximate analysis methods are somewhat based on a product-form solution. They analyze a system as if its solution were of product form and adjust the result obtained from the product-form assumptions to better satisfy system equations.

From the set of approximate analysis methods of queueing networks we summarize traffic-based decomposition.

10.9 Traffic-Based Decomposition

One way to interpret the product-form solution is that the network nodes are independently analyzed based on the traffic load given by the solution of the traffic equation and the known service process (discipline and service time) of the node.

Traffic-based decomposition is an iterative procedure that analyzes the nodes of a network independently, and the traffic load of the node under evaluation is determined based on the departure processes of the network nodes previously analyzed.

The advantages of the procedure are its flexibility and low computational cost, while its disadvantages are the potential inaccuracy of the results and the lack of evidence about the convergence of the procedure. Despite its disadvantages, this is a very often applied approximate analysis method in practice because in the majority of cases it converges and gives reasonable agreement with simulation results.

The traffic-based decomposition procedure iteratively goes through all nodes of the network and performs the following steps for all nodes:

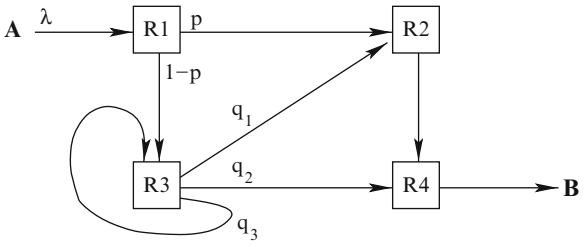
- Traffic aggregation: aggregates the traffic coming from the environment and from the departure processes of the other nodes (based on the preceding iterations).
- Node analysis and departure process computation: a single queueing system analysis step in which the parameters of the departure process are also computed.
- Departure process filtering: computation of traffic components going to other network nodes.

The complexity of an iteration step and the accuracy of the results depend on the applied traffic descriptors. The flexibility of the procedure is due to the wide range of potentially applicable traffic descriptors. The most commonly used traffic descriptor is the average intensity of the traffic such that a Poisson arrival process is assumed with a given intensity. Using this traffic model with more than one traffic class results in a nontrivial analysis problem itself. If a more sophisticated traffic model is applied to, e.g., higher moments or correlation parameters of the interarrival time distribution are considered, then the complexity of the analysis steps increases and the overall accuracy improves.

10.10 Exercises

Exercise 10.1. In the depicted queueing network the requests of input A are forwarded to output B according to the following traffic routing probabilities: $p = 0.3, q_1 = 0.2, q_2 = 0.5, q_3 = 0.3$.

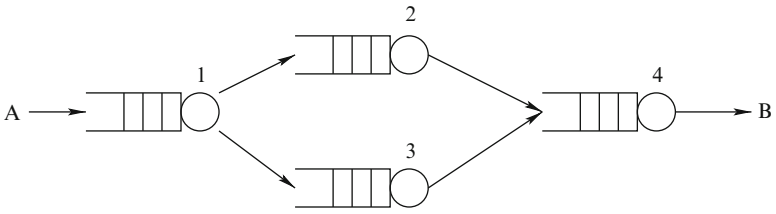
Requests from input A arrive according to a Poisson process at a rate $\lambda = 50$. The service times are exponentially distributed in nodes R1, R2, and R3 with the parameters $\mu_1 = 90$, $\mu_2 = 35$, and $\mu_3 = 100$, respectively. The service time in



R4 is composed of two phases. The first phase is exponentially distributed with the parameter $\mu_4 = 400$, and the second phase is deterministic with $D = 0.01$.

- Compute the traffic load of the nodes.
- Compute the mean and the coefficient of variation of the service time at node R4.
- Compute the system time at each node.
- Compute $\lambda_{\max}$ at which the system is at the limit of stability.

Exercise 10.2. In the depicted queueing network the requests of input A are forwarded to output B according to the following traffic routing probabilities: $p_{12} = 0.3$, $p_{13} = 0.7$.



The requests from input A arrive according to a Poisson process at a rate $\lambda = 50$. In nodes 1, 2, and 3 there are single servers and infinite buffers, and the service times are exponentially distributed with the parameters $\mu_1 = 80$, $\mu_2 = 45$, and $\mu_3 = 50$, respectively. There are two servers and two additional buffers at node R4. Both servers can serve requests with exponentially distributed service time with the parameter $\mu_4 = 40$.

- Characterize the nodes using Kendall's notation.
- Compute the traffic load of the nodes.
- Compute the system time at each node.
- Compute the server utilization at node 4.
- Compute the packet loss probability.
- Compute the mean time of a request from A to B.
- Which node is the bottleneck of the system? Which node saturates first when λ increases?

Chapter 11
Applied Queueing Systems

11.1 Bandwidth Sharing of Finite-Capacity Links with Different Traffic Classes

Traditional telephone networks were designed to implement a single type of communication service, i.e., the telephone service. Today's telecommunication networks implement a wide range of communication services. In this section we introduce Markov models of communication services that compete for the bandwidth of a finite-capacity communication link.

11.1.1 Traffic Classes

There are several important features of traffic sources of communication services that allow for their classification. Here assume that the traffic sources require the setting up of a connection for a finite period of time during which data communication is carried out between the parties of the communication service. We classify the traffic sources based on the bandwidth of the data transmission during a connection. The simplest case is where data are transmitted with a fixed bandwidth during a connection. This case is commonly referred to as constant bit rate (CBR). A more general traffic behavior is obtained when the bandwidth of data transmission varies during a connection. This case is commonly referred to as variable bit rate (VBR). The most common form of bandwidth variation is when the bandwidth alternates between 0 and a fixed bandwidth. These VBR sources are referred to as ON-OFF sources, and we restrict our attention to the ON-OFF case. The most complex traffic sources adjust their bandwidth according to the available capacities of the network resources. There are two classes of this kind of source. *Adaptive* traffic sources set up a connection for a given period of time and transmit data according to the available bandwidth in the network. If the network resources are occupied during the connection of an adaptive traffic

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, 303
DOI 10.1007/978-1-4614-5317-8_11, © Springer Science+Business Media, LLC 2013

source, then the source transmits data with a low bandwidth, and the overall amount of transmitted data during a connection is low. *Elastic* traffic sources set up a connection for transmitting a given amount of data. The bandwidth of the data transmission depends on the available bandwidth in the network. If the network resources are occupied during the connection of an elastic connection, then the period of the connection is extended in such a way that the source transmits the required amount of data.

In this section we assume that the traffic sources demonstrate a memoryless time-homogeneous stochastic behavior and, consequently, the arrival processes are Poisson processes and the connection times are exponentially distributed, except for the elastic class, where the amount of data to transmit is exponentially distributed. Additionally, the traffic sources are characterized by their bandwidth parameters. In the case of CBR and ON-OFF VBR sources, the bandwidth parameter is the bandwidth of the active period. In the case of adaptive and elastic sources, the bandwidth parameters are the minimal and maximal bandwidth at which the source can transmit data.

Consequently, in the case of the different kinds of traffic sources, a class k traffic source is characterized by the following parameters:

- CBR connection: connection arrival intensity λ_k, bandwidth requirement c_k, parameter of exponentially distributed connection holding time μ_k;
- VBR connection: connection arrival intensity λ_k, bandwidth requirement in ON state c_k, parameters of exponentially distributed connection holding time, ON time, and OFF time μ_k, α_k, and β_k, respectively.
- Adaptive connection: connection arrival intensity λ_k, minimal bandwidth $c_{\min}^{(k)}$, maximal bandwidth $c_{\max}^{(k)}$, parameter of exponentially distributed connection holding time μ_k;
- Elastic connection: connection arrival intensity λ_k, minimal bandwidth $c_{\min}^{(k)}$, maximal bandwidth $c_{\max}^{(k)}$, parameter of exponentially distributed amount of transmitted data δ_k.

These parameters define the arrival process and the bandwidth needs of the traffic sources but they do not define completely the service procedure as the common resource (the finite capacity link) is shared among the traffic types and classes. In the case of traditional telephone services, the procedure for a new telephone call is obvious: accept as many calls as possible with the given finite-capacity link. In the case of different traffic classes, more complex procedures are required to properly utilize the resources and to provide the desired service features to each traffic class. The set of rules concerning the acceptance or rejection of a new connection is referred to as call admission control (CAC). CAC defines the acceptance or rejection of a new connection of all types under all possible traffic conditions. We will see some typical CACs and their properties.

The most common performance parameters of interest in these kinds of traffic models are

- Per-class connection-dropping probability (at arrival connection arrival),
- VBR connection-dropping probabilities (during ongoing connection at an OFF to ON transition),
- Per-class mean bandwidth of adaptive and elastic connections,
- Sojourn time of elastic connections.

Different dimensioning methods apply for different traffic classes. In the following sections we investigate the simple Markov models of these traffic classes, which form the bases of the complex dimensioning methods used in practice.

11.1.2 Bandwidth Sharing by CBR Traffic Classes

One of the first generalizations of traditional telecommunication models is due to the coexistence of communication services with different bandwidth requirements. When a link is utilized by different kinds of CBR connections with the previously detailed Markovian properties, then the overall system behavior can be described by a CTMC. The main problem of analyzing the performance parameters through this CTMC is the potentially very high number of states. If a finite-capacity link of bandwidth C is utilized by I different kinds of CBR connections, then a state of the CTMC should represent the number of ongoing connections of each class, and the number of states is proportional to the product $\prod_{i=1}^{I}(\frac{C}{c_i}+1)$.

To overcome this practical problem, an efficient numerical procedure was proposed by two researchers independently [50, 80]; the procedure is often referred to as the Kaufman–Roberts method. It is based on the fact that a large CTMC, which represents the number of ongoing connections of each class, satisfies the local balance equations

$$\lambda_i \; p(n_1,\ldots,n_i-1,\ldots,n_I) = n_i\mu_i \; p(n_1,\ldots,n_i,\ldots,n_I),$$

where $p(n_1,\ldots,n_i,\ldots,n_I)$ denotes the stationary probability of the state where the number of class i connections is n_i for $i = 1,\ldots,I$. The local balance equation represents that the stationary state-transition rate due to an arriving class i connection is in balance with the stationary state-transition rate due to a departing class i connection. The main idea of the Kaufman–Roberts method is to unify those states of a large Markov chain that represent the same bandwidth utilization of a link. In the state $(n_1,n_2,\ldots,n_I)$, the bandwidth utilization is $c=\sum_{i=1}^{I} n_i c_i$. Summing up the local balance equations for the states where the bandwidth utilization on the right-hand side is c we have

$$\sum_{i=1}^{I} \lambda_i \; P(c - c_i)\, \mathcal{I}_{\{c_i \geq c\}} = \sum_{i=1}^{I} n_i \mu_i \; P(c)$$

$$\sum_{i=1}^{I} \frac{\lambda_i c_i}{\mu_i} \; P(c - c_i)\, \mathcal{I}_{\{c_i \geq c\}} = \underbrace{\sum_{i=1}^{I} n_i c_i}_{c} \; P(c)$$

$$\sum_{i=1}^{I} \frac{\lambda_i c_i}{\mu_i c} \; P(c - c_i)\, \mathcal{I}_{\{c_i \leq c\}} = \; P(c),$$

where $P(c)$ denotes the sum of the stationary probabilities of the states where the bandwidth utilization is c, that is, $P(c) = \sum_{(n_1,n_2,\ldots,n_I):\sum_{i=1}^{I} n_i c_i = c} p(n_1, n_2, \ldots, n_I)$. The last equation is the core of the Kaufman–Roberts method, which computes the relative (nonnormalized) probabilities of the link utilization levels first and then normalizes probabilities as follows.

1. Let $\tilde{P}(0) = 1$, and for $c = 1, 2, \ldots, C$ compute

$$\tilde{P}(c) = \sum_{i} \frac{\lambda_i c_i}{\mu_i c} \; \tilde{P}(c - c_i)\, \mathcal{I}_{\{c_i \leq c\}}.$$

2. Compute $\tilde{P} = \sum_{c=0}^{C} \tilde{P}(c)$.
3. Normalize the probabilities by $P(c) = \tilde{P}(c)/\tilde{P}$.

There is an implicit technical assumption that is necessary for the application of the Kaufman–Roberts method. There must be a bandwidth unit such that each c_i is an integer multiple of this bandwidth unit. (The method remains applicable if C is not an integer multiple of the bandwidth unit.) Fortunately, in important applications such a bandwidth unit exists.

Having the stationary probabilities of the utilization levels we can compute the loss probabilities. If the CAC allows all connections entering the link as long as the available bandwidth is not less than the bandwidth of the entering connection, then the loss probability of class i connections is

$$b_i = \sum_{c > C - c_i} P(c).$$

It is a straightforward consequence of the CAC that connections with higher bandwidth requirements have a higher loss probability. If a kind of fairness is required among the different classes such that each class experiences the same loss probability, then the CAC needs to be modified. Let us assume that the traffic class with the highest bandwidth is class I. If the CAC is modified such that each incoming connection is rejected when the available bandwidth is less than c_I, then the distribution of the link utilization changes, but each class is accepted and rejected at the same time at the different link utilization levels, and consequently they have

the same loss probability. If the CAC depends only on the link utilization level (as in the case of a modified CAC with identical dropping probabilities), then the Kaufman–Roberts method remains applicable. In this case the main iteration step of the procedure changes to

$$\tilde{P}(c) = \sum_i \frac{\lambda_i c_i}{\mu_i c} \tilde{P}(c - c_i) \, \mathrm{CAC}(i, c - c_i),$$

where $\mathrm{CAC}(i, c)$ is one if a class i connection is accepted at link utilization c, and zero otherwise. The link utilization-level-dependent CAC can also be generalized to probabilistic CACs. In this case the main iteration step of the procedure remains the same as for the deterministic one, and $\mathrm{CAC}(i, c)$ indicates the probability that a class i connection is accepted at link utilization c.

11.1.3 Bandwidth Sharing with VBR Traffic Classes

When a link is utilized by different kinds of VBR connections and each of them is characterized by the previously described Markovian properties, the overall system behavior can be described by a CTMC. The states of this CTMC represent the number of ongoing VBR connections of each class and the number of connections in the ON phase, $(n_1, m_1, n_2, m_2, \ldots, n_I, m_I)$. Note that $n_i \geq m_i$, $i = 1, \ldots, I$, and $\sum_{i=1}^{I} m_i c_i \leq C$, where the second inequality means that the utilized bandwidth should not exceed the link capacity. The state space of this CTMC is even larger than that for CBR connections, which represents only the number of ongoing connections of each class, but unfortunately there is no more efficient computation method available for this model than to solve the CTMC. This is due to the fact that this CTMC does not satisfy the local balance equations. At any rate, the numerical solution of this CTMC is still possible for a limited number of VBR classes and connections.

Generally, the CAC for VBR connections is more complex than that for CBR connections. A conservative CAC does not allow more VBR connections than the link can serve, assuming that all VBR connections are in the ON state. That is, $\sum_{i=1}^{I} n_i c_i \leq C$ holds for each state. Unfortunately, a conservative CAC results in a very low resource utilization, especially when the length of the ON period is short with respect to the length of the OFF period. In these cases, it is worth allowing more VBR connections than a conservative CAC in order to increase the link utilization. The drawback of nonconservative CACs is that accepted ongoing VBRs can be dropped due to insufficient capacity with positive probability at an OFF to ON phase transmission. In practice, it is usually required that the dropping probability of ongoing VBR connections be much lower than that of newly arriving ones.

The possible state transitions of Markov chains are

(a) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i + 1, m_i + 1, \ldots, n_I, m_I)$ at rate λ_i if $\mathrm{CAC}(i, \{n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I\}) = 1$;
(b) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i - 1, m_i - 1, \ldots, n_I, m_I)$ at rate $m_i \mu_i$;
(c) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i - 1, m_i, \ldots, n_I, m_I)$ at rate $(n_i - m_i)\mu_i$;
(d) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i, m_i - 1, \ldots, n_I, m_I)$ at rate $m_i \alpha_i$;
(e) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i, m_i + 1, \ldots, n_I, m_I)$ at rate $(n_i - m_i)\beta_i$ if $\sum_{j=1}^{I} m_j c_j + c_i \le C$;
(f) $(n_1, m_1, \ldots, n_i, m_i, \ldots, n_I, m_I) \to (n_1, m_1, \ldots, n_i - 1, m_i, \ldots, n_I, m_I)$ at rate $(n_i - m_i)\beta_i$ if $\sum_{j=1}^{I} m_j c_j + c_i > C$,

where $\mathrm{CAC}(i, \{n_1, m_1, \ldots, n_I, m_I\})$ denotes the CAC decision in state $(n_1, m_1, \ldots, n_I, m_I)$ for an incoming class i connection, and state transitions with a zero rate are impossible. According to the bandwidth limit of a link,

$$\mathrm{CAC}(i, \{n_1, m_1, \ldots, n_I, m_I\}) = 0 \text{ if } \sum_{j=1}^{I} m_j c_j + c_i > C.$$

The transitions represent the following events:

(a) New class i connection arrival.
(b) Departure of a class i connection that is in the ON phase.
(c) Departure of a class i connection which is in the OFF phase.
(d) A class i connection switches from ON to OFF phase.
(e) A class i connection switches from OFF to ON phase.
(f) A class i connection is lost due to insufficient bandwidth for OFF to ON phase transition.

With the stationary probabilities of this CTMC, denoted by $p(n_1, m_1, \ldots, n_I, m_I)$, the dropping probability of class i incoming and ongoing connections can be computed as follows:

$$\begin{aligned} b_i^{\text{new}} &= \frac{\text{number of class } i \text{ incoming connections dropped upon arrival}}{\text{number of class } i \text{ incoming connections}} \\ &= \sum_{n_1, m_1, \ldots, n_I, m_I} p(n_1, m_1, \ldots, n_I, m_I)(1 - \mathrm{CAC}(i, \{n_1, m_1, \ldots, n_I, m_I\})), \end{aligned}$$

$$b_i^{\text{ongoing}} = \frac{\text{number of class } i \text{ dropped ongoing connections}}{\text{number of class } i \text{ incoming connections}}$$
$$= \sum_{\mathcal{S}_i} \frac{(n_i - m_i)\beta_i}{\lambda_i} p(n_1, m_1, \ldots, n_I, m_I),$$

where $\mathcal{S}_i$ denotes the set of states for which $\sum_{j=1}^{I} m_j c_j + c_i > C$. The link utilization is

$$\rho = \sum_{n_1, m_1, \ldots, n_I, m_I} p(n_1, m_1, \ldots, n_I, m_I) \sum_{j=1}^{I} m_j c_j .$$

If the state space of the CMTC is such that a stationary analysis is feasible, then the computation of the performance parameters is straightforward, but the inverse problem, the design of a CAC that satisfies blocking probability constraints and maximizes link utilization, is still an interesting research problem.

11.1.4 Bandwidth Sharing with Adaptive Traffic Classes

In the case of adaptive traffic classes, connections can adapt their bandwidth to the available bandwidth of the link between the class-specific bandwidth limits $c_{\min}^{(i)}$ and $c_{\max}^{(i)}$. If the link is not completely utilized, then each connection receives its maximal bandwidth. If the sum of the maximal bandwidth needs is larger than the link capacity, then the link is completely utilized and a bandwidth reduction affects the bandwidth of all classes according to the following rule. If the actual bandwidth of a class i connection is c and c is less than $c_{\max}^{(i)}$, then for any other class j the bandwidth is c if $c \leq c_{\max}^{(j)}$ or $c_{\max}^{(j)}$ if $c > c_{\max}^{(j)}$. This means that the bandwidth of each class is reduced to the same level c if the class-specific maximal bandwidth $c_{\max}^{(j)}$ is not less than c. Consequently, the main features of bandwidth sharing with adaptive traffic classes are as follows:

- The departure rate of connections is proportional to the number of active connections and is independent of the instantaneous bandwidth of the connections.
- The bandwidth of the connections varies according to the link capacity and the number of active connections.
- An arriving class i connection is rejected when the minimal required bandwidth $c_{\min}^{(i)}$ cannot be granted.
- The transmitted data of a connection depends on the instantaneous bandwidth during the connection.

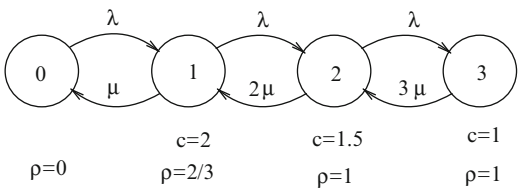


Fig. 11.1 Markov chain of the number of adaptive connections on a finite-capacity link

Example 11.1. We demonstrate the behavior of adaptive connections on a finite-capacity link in the case of a single adaptive class with link bandwidth $C = 3$ Mbps, bandwidth limits $c_{\min} = 1$ Mbps, $c_{\max} = 2$ Mbps, and connection arrival and departure rates λ [1/s] and μ[1/s], respectively. Due to the memoryless arrival and departure processes, the number of active connections $X(t)$ is a Markov chain and it is depicted in Fig. 11.1. The figure also indicates the bandwidth of the ongoing connections. If there are three ongoing connections, then the arriving connections are rejected because in the case of four connections the common bandwidth $c = 3/4$ Mbps would be smaller than the minimal bandwidth requirement $c_{\min} = 1$ Mbps.

The main performance measures of this system are the mean bandwidth of connections

$$\bar{c} = \sum_{i=0}^{3} i\, c(i)\, p_i = 2p_1 + 2 \cdot 1.5 p_2 + 3 \cdot 1 p_3,$$

the link utilization

$$\rho = \sum_{i=0}^{3} \rho_i\, p_i = 2/3 p_1 + 1 p_2 + 1 p_3,$$

and the blocking probability

$$b = p_3,$$

where p_i, ρ_i, and $c(i)$ denote the stationary probability, the utilization, and the bandwidth of a connection in state i, respectively.

Example 11.2. The approach applied to the single class model can be used for the analysis of models with multiple adaptive classes. In the case of two adaptive classes with link bandwidth $C = 5$, bandwidth limits $c_{\min}^{(1)} = 1.5$, $c_{\max}^{(1)} = 3$, $c_{\min}^{(2)} = 1$, $c_{\max}^{(2)} = 2$, connection arrival and departure rates λ_1, λ_2 and μ_1, μ_2, the Markov chain describing the number of active connections of class 1 and 2 is depicted in Fig. 11.2. The figure indicates the bandwidth of the ongoing connections by bold characters. Arriving connections of both classes are rejected in states $(0, 5)$, $(1, 2)$, $(2, 1)$, $(3, 0)$, and additionally arriving connections of class 1 are rejected in states $(0, 3)$, $(0, 4)$. Considering only the minimal bandwidth constraints and the link bandwidth, the state $(1, 3)$ would be feasible $(1.5 + 3 \cdot 1 < 5)$, but the identical bandwidth sharing of the classes makes this state infeasible because it violates the minimal bandwidth requirement of class 1 $(5/4 < c_{\min}^{(1)})$. In contrast, in the state $(1, 1)$ the bandwidth is

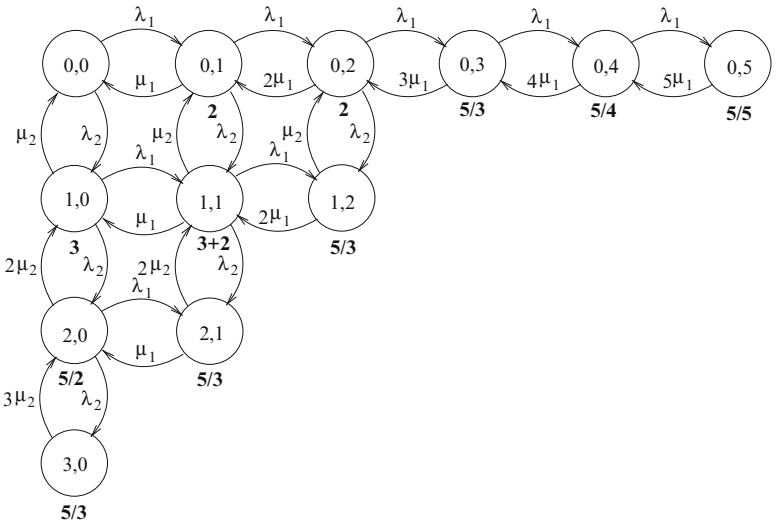


Fig. 11.2 Markov chain of the number of adaptive connections with two adaptive classes

unevenly divided. This is possible because a class 2 connection obtains its maximal bandwidth and the remaining bandwidth is utilized by the class 2 connection. The performance measures can be computed in a similar way as in the case of a single adaptive class.

11.1.5 Bandwidth Sharing with Elastic Traffic Classes

In the case of elastic traffic classes, the connections can adapt their bandwidth to the available bandwidth similar to the adaptive class, but the amount of data transmitted through a connection is fixed. Thus, during a period when the bandwidth is low, the sojourn time of the elastic connections is longer. The bandwidth of elastic connections is also bounded by class-specific bandwidth limits $c_{\min}^{(i)}$ and $c_{\max}^{(i)}$, and the bandwidth sharing between traffic classes follows the same role as in the case of adaptive connections. The main features of bandwidth sharing with elastic traffic classes are as follows:

- The departure rate of a connection of class i depend on the instantaneous bandwidth of the class i connections. Thus, the length of the connections varies according to the link capacity and the number of active connections.
- An arriving class i connection is rejected when the minimal required bandwidth $c_{\min}^{(i)}$ cannot be granted.
- The amount of transmitted data of a connection is a class-specific random variable that does not depend on the instantaneous bandwidth during the connection.

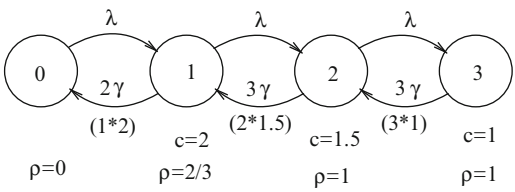


Fig. 11.3 Markov chain of the number of elastic connections on a finite-capacity link

Example 11.3. We demonstrate the behavior of elastic connections with the same model as in Example 11.2 but assuming that the connections are elastic. That is, the link bandwidth is $C = 3$ Mbps, the bandwidth limits are $c_{\min} = 1$ [Mb/s] and $c_{\max} = 2$ [Mb/s], the connection arrival rate is λ[1/s], and the amount of transmitted data of an elastic connection is exponentially distributed with the parameter γ[1/Mb]. Due to the memoryless arrival process and the exponential distribution of the amount of transmitted data of the elastic connections, the number of active connections $X(t)$ is a Markov chain (Fig. 11.3). The figure indicates the bandwidth of the ongoing connections (parameter c) and the computation of the departure rate of connections in brackets. For example, in state 2 there are two ongoing connections with bandwidth 1.5 [Mb/s]. The rate at which one of them completes the data transmission is $1.5[Mb/s] \times \gamma[1/Mb] = 1.5\gamma[1/s]$ and the sum of the two identical departure rates is $2 \times 1.5[1/s] = 3[1/s]$. Apart from these differences, the bandwidth sharing, the link utilization, and the rejection of arriving connections are the same as in the case of adaptive connections.

11.1.6 Bandwidth Sharing with Different Traffic Classes

In the previous sections we discussed the bandwidth sharing of a finite-capacity link by traffic classes of the same type. All of the discussed traffic classes have a memoryless stochastic behavior, and thus the performance of the models can be analyzed by CTMCs. Unfortunately, practical limitations arise when the size of the state space gets large, which is often the case in practically interesting situations. The case where only CBR-type connections are present at a link allows for the efficient analysis method referred to as the Kaufman–Roberts method. If any other types of connections appear, then this method will no longer be applicable. The Markov-chain-based framework of the previous sections is also applicable to the analysis of bandwidth sharing by traffic classes of different types. Interested readers may find further details in [68, 77, 78, 81].

11.2 Packet Transmission Through Slotted Time Channel

In this section we focus on a peculiar detail of modeling slotted time systems with discrete-time Markov chains (DTMCs) – the definition of time slots, more precisely, the positioning of the beginning of a time slot on continuous-time axes. The modeler

has some freedom in this respect, and consequently different DTMC models can be obtained for describing the same system behavior. It turns out that these different models result in the same performance parameters if the performance parameters are independent of the slot definition, which is the case with the majority of the practically important queueing parameters. Below we evaluate two models of a simple packet transmitter, which can be seen as discrete-time counterparts of an M/M/1 queue.

Consider a packet transmitter with the following properties:

- Packet arrival process: in each time slot, 1 packet arrives with probability p and 0 packets arrive with probability $1 - p$ independently of past history.
- Service (packet transmission) process: if there is at least one packet to transmit, then 1 packet is transmitted with probability q and 0 packets are transmitted with probability $1-q$ independently of past history, which means that the service time of a packet is geometrically distributed with parameter q [Pr service time $= k = (1 - q)q^{k-1}$].
- Service discipline: FIFO.
- Buffer size: infinite.

Let X_n denote the number of packets in a system at the beginning of the nth time slot. X_n is a DTMC with a special birth-and-death structure and infinite state space. Depending on the definition of the beginning of a time slot, the following two cases arise.

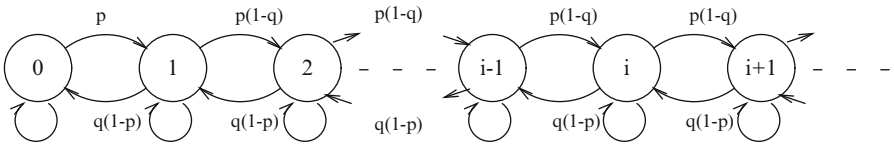
- **Case I.**: A time slot starts with packet transmission (if any), and after that packet, arrivals can happen.
- **Case II.**: A time slot starts with packet arrival (if any), and after that packet, transmission can happen.

These two cases result in different Markov chains, as detailed below.

Case I. In case I, the X_n Markov chain can be described by the following evolution equation:

$$X_{n+1} = (X_n - V_{n+1})^+ + Y_{n+1},$$

where the random variable Y_{n+1} is the number of packet arrivals during time slot $n + 1$ and the random variable V_{n+1} is the number of packets that can be transmitted during time slot $n + 1$. Y_n and V_n are Bernoulli distributed with parameters p and q, respectively. The state-transition graph of this Markov chain is



The stationary distribution of this Markov chain is

$$p_0 = \frac{q-p}{q},\ p_i = \left(\frac{p(1-q)}{q(1-p)}\right)^i \frac{q-p}{(1-q)q},\ i \geq 1.$$

The numerator of the stationary probabilities already indicates that the condition of stability is $p < q$. This result can also be obtained from the evolution equation $\mathbf{E}(V) > \mathbf{E}(Y) \rightarrow p < q$ and from the Foster criterion (Theorem 3.42) $q(1-p) > p(1-q) \rightarrow p < q$. The basic performance measures can be computed from the stationary distribution. The utilization is

$$\rho = 1 - p_0 = 1 - \left(1 - \frac{p}{q}\right) = \frac{p}{q},$$

the mean of the stationary number of packets in the queue is

$$\mathbf{E}(X) = \sum_{i=1}^{\infty} i\, p_i = \frac{p(1-p)}{q-p},$$

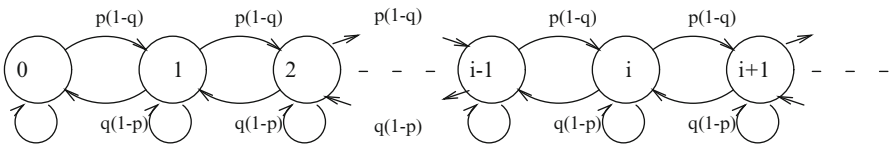
and the mean of the stationary system time of a packet is

$$\mathbf{E}(T) = \frac{1}{q} p_0 + \sum_{i=1}^{\infty} p_i \left(\frac{i+1}{q} - 1\right) = \frac{1-p}{q-p}.$$

Case II. In this case the evolution equation has the form

$$X_{n+1} = (X_n - V_{n+1} + Y_{n+1})^+,$$

and the transition graph of the Markov chain is



Due to the different transition probabilities around state 0, we have a different stationary distribution

$$p_i = \left(\frac{p(1-q)}{q(1-p)}\right)^i \frac{q-p}{(1-p)q},\ i \geq 0.$$

The computation of some performance measures is identical in this case. for example, the condition of stability is $\mathbf{E}(V) > \mathbf{E}(Y) \rightarrow p < q$ based on the

evolution equation and $q(1-p) > p(1-q) \rightarrow p < q$ based on the Foster criterion. The computation of some other performance measures is different in case II. For example the utilization is computed as

$$\rho = 1 - p_0(1-p) = 1 - \frac{(q-p)(1-p)}{(1-p)q} = \frac{p}{q}$$

because the server can be utilized by a packet that arrives with probability p when it is idle at the beginning of a time slot. The mean of the stationary system time can be computed as

$$\mathbf{E}(T) = \sum_{i=0}^{\infty} p_i \left(\frac{i+1}{q} - 1 \right) = \frac{1-q}{q-p},$$

and the results are identical with those in case I. In contrast, the mean of the stationary number of packets in the queue is

$$\mathbf{E}(X) = \sum_{i=1}^{\infty} i\, p_i = \frac{p(1-q)}{q-p},$$

which is different from the results of case I. It reflects the fact that a different number of packets is in the system before and after an arrival.

The evaluated performance measures validate the intuitive expectations that there are performance measures that are dependent on the definition of time slot and others that are independent of that definition. A modeler can choose the time slot freely if the required performance measures are time slot definition independent, but the time slot definition should be related to the required performance measures otherwise.

11.3 Analysis of an Asynchronous Transfer Mode Switch

11.3.1 Traffic Model of an Asynchronous Transfer Mode Switch

In this section we consider the behavior of an asynchronous transfer mode (ATM) [28] switch with N input and N output ports and set up a traffic model of this behavior. An ATM switch transmits packets of fixed size (53 bytes), referred to as a cell. The input and output ports work in a slotted synchronized manner. The length of a time slot is the transmission time of a cell.

We assume that the arrival processes of cells to the input ports of the switch are independent and memoryless. The processes of cell arrival at input ports are

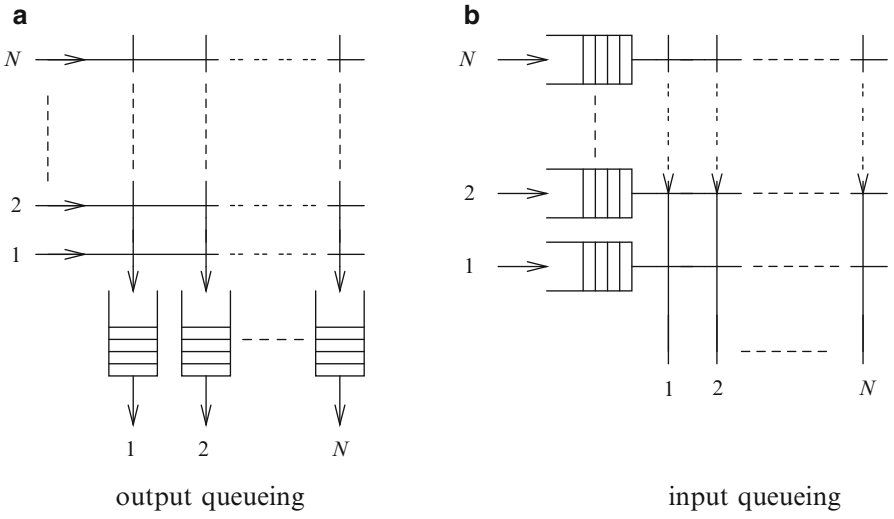


Fig. 11.4 Input and output buffering in packet switching

characterized by a vector $q = \{q_i\}$, $i = 1, \ldots, N$, where q_i is the probability that 1 cell arrives at input port i in a time slot. Consequently, the probability that 0 cells arrive at input port i in a time slot is $1 - q_i$.

An incoming cell is directed to one of the N output ports. We assume that a cell from input port i is directed to output port j with probability w_{ij} independent of the past and the state of the system. The matrix composed by these probabilities $W = \{w_{ij}\}$ is referred to as a traffic matrix. The traffic is a stochastic matrix, that is, $w_{ij} \geq 0$ and $\sum_j w_{ij} = 1$.

The bandwidth of the input and output ports are identical. If more than one cell is directed to a given output port in a given time slot, then only one of them can be transmitted and the others are buffered. As is quantified below, the location of the buffers where the colliding cells are stored has a significant effect on the performance of the switch. We consider two cases: buffering at the input ports and buffering at the output ports. Figure 11.4 depicts the structure of these cases.

Real systems contain buffers both at the input and at the output ports. The performance characteristics and the design of the switch determine the proper model of the system. In the worst case, one cell arrives at each input port and each cell is directed to the same output. If the switch is designed such that it can transfer all of the N cells to the buffer of the given output port in a single time slot, then the output buffer model describes the system properly. If the switch is designed such that it can transfer only one of the conflicting cells to the output buffer and the remaining $N-1$ cells are left at the input buffers, then the input buffer model is the proper model of the system.

Between the input and output buffer models the output buffer seems to provide better performance because in the case of an input buffer model it can happen

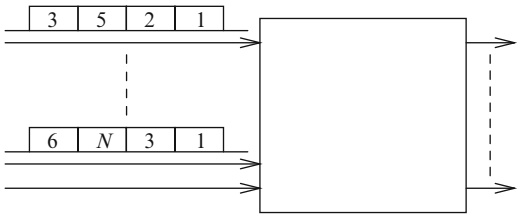


Fig. 11.5 Head-of-line blocking with input buffering

that a given input port is blocked due to the conflict of the first cell in the queue, while other cells waiting in the buffer are directed to idle output ports (Fig. 11.5). This phenomenon is often referred to as head-of-line blocking. This very intuitive qualitative comparison of the two buffer models will be quantified below for some special symmetric configurations.

11.3.2 Input Buffering

In this section we consider the simplest input buffering case, where $N = 2$. If two cells at the heads of the two input ports are directed to the same output port, then one of them is chosen with even probability $(1/2)$ and the chosen one is transferred to the output port and the other one is left in the buffer of the input port. With the preceding modeling assumptions the number of cells in the two input buffers is a DTMC. We assume that the time slots are such that if a cell arrives at an idle buffer and does not collide with any other cell, then it leaves the input port in the same time slot.

Due to the fact that the system state is described by two discrete variables (the number of cells at the two input buffers) it is worth it to depict the state space as a two-dimensional one (Fig. 11.6). The state space can be divided into four parts: both queues are idle, queue 1 is idle and queue 2 is busy, queue 2 is idle and queue 1 is busy, both queues are busy. The state-transition probabilities follow the same structure in these four parts.

Figure 11.7 shows the environment of (0, 0). It is the state where both buffers are idle, and state transitions starting from this state are depicted. In this and the following figures S denotes the probability of conflict. Conflict occurs when two cells from the head of the two buffers are directed to the same output. Its probability is $S = w_{11}w_{21} + w_{12}w_{22}$, where the first term stands for the case where both cells go to input 1 and the second term stands for the case where they go to input 2. Starting from (0, 0), there are the following cases:

- The next state is (1, 0) if there is a conflict and the cell from input 2 is chosen for transmission.
- The next state is (1, 0) if there is a conflict and the cell from input 1 is chosen for transmission.

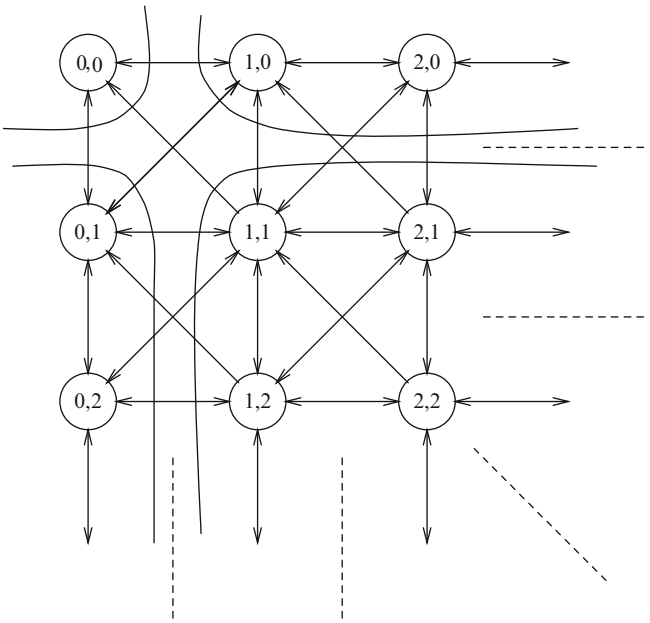
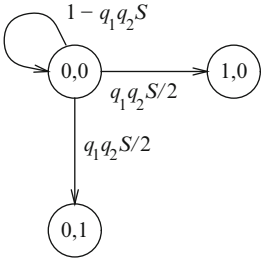


Fig. 11.6 Markov chain with input buffering

Fig. 11.7 Idle buffers



- If there is no conflict (zero or one cell arrives or two cells arrive but the cells are directed to a different output), then the next state is (0, 0).

Figure 11.8 shows the state transitions when buffer 2 is idle and there is at least one cell in buffer 1. Denoting the starting state by $(x, 0)$, $x \geq 1$, the following state transitions can occur:

- $(x, 0) \rightarrow (x - 1, 0)$ if
  - No new cells arrive,
  - A cell arrives at input 2.
- $(x, 0) \rightarrow (x, 0)$ if
  - A cell arrives at input 1 and no cell arrives at input 2;

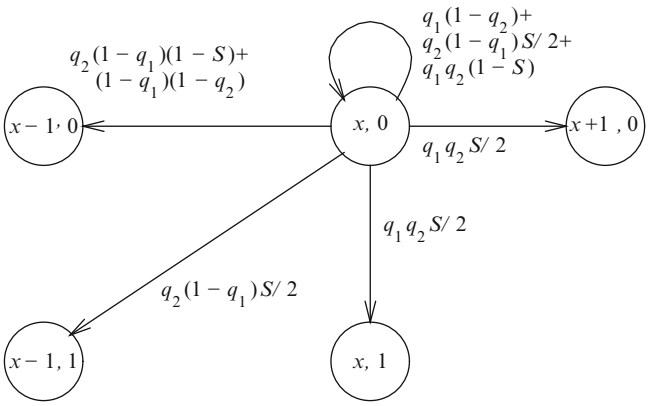


Fig. 11.8 Buffer 2 is idle

  - A cell arrives at input 2 and no cell arrives at input 1, it is in conflict with the one at the head of buffer 1, and the cell in buffer 2 is chosen for transmission;
  - Cells arrive at both inputs and there is no conflict (the cells at the head of the buffer are directed to different outputs).

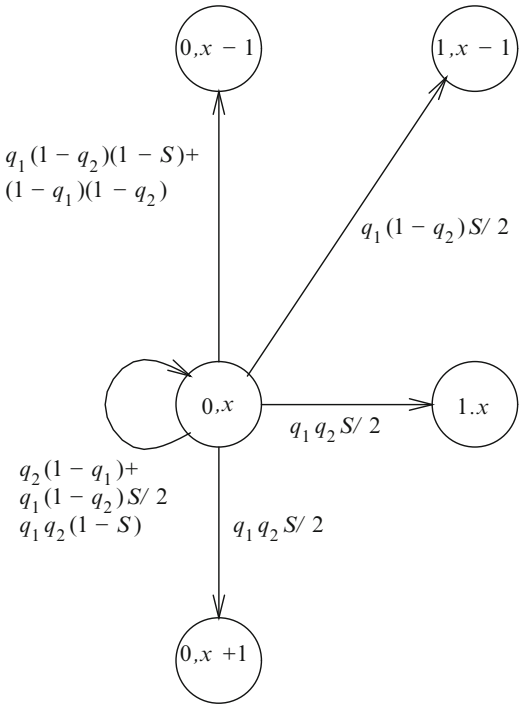
- $(x, 0) \rightarrow (x + 1, 0)$ if
  - Cells arrive at both inputs, there is a conflict, and the cell in buffer 2 is chosen for transmission.
- $(x, 0) \rightarrow (x - 1, 1)$ if
  - A cell arrives at input 2 and no cell arrives at input 1, there is a conflict, and the cell in buffer 1 is chosen for transmission.
- $(x, 0) \rightarrow (x, 1)$ if
  - Cells arrive at both inputs, there is a conflict, and the cell in buffer 1 is chosen for transmission.

States where buffer 1 is idle and buffer 2 is not idle is depicted in Fig. 11.9. The state transitions of these cases follow a similar pattern as those in Fig. 11.8 by replacing the role of the buffers.

Figure 11.10 presents a case where cells are waiting in both buffers. The figure does not show transition $(x, y) \rightarrow (x, y)$ whose probability is 1 minus the sum of the depicted transition probabilities. The following state transitions are possible.

- $(x, y) \rightarrow (x - 1, y - 1)$ if
  - No new cells arrive and there is no conflict.
- $(x, y) \rightarrow (x, y - 1)$ if
  - A cell arrives at input 1, no cells arrive at input 2, and there is no conflict;

Fig. 11.9 Buffer 1 is idle



  - No new cells arrive, there is a conflict, and the cell in buffer 2 is chosen for transmission.

- $(x, y) \to (x + 1, y - 1)$ if
  - A cell arrives at input 1, no cells arrive at input 2, there is a conflict, and the cell in buffer 2 is chosen for transmission.
- $(x, y) \to (x - 1, y)$ if
  - A cell arrives at input 2, no cells arrive at input 1, and there is no conflict;
  - No new cells arrive, there is a conflict, and the cell in buffer 1 is chosen for transmission.
- $(x, y) \to (x, y)$ if
  - A cell arrives at input 1, no cells arrive at input 2, there is a conflict, and the cell in buffer 1 is chosen for transmission (with probability $q_1(1 - q_2)S/2$);
  - A cell arrives at input 2, no cell arrives at input 1, there is a conflict, and the cell in buffer 2 is chosen for transmission (with probability $q_2(1 - q_1)S/2$);
  - New cells arrive at both buffers, and there is no conflict [with probability $q_1q_2(1 - S)$].
- $(x, y) \to (x + 1, y)$ if

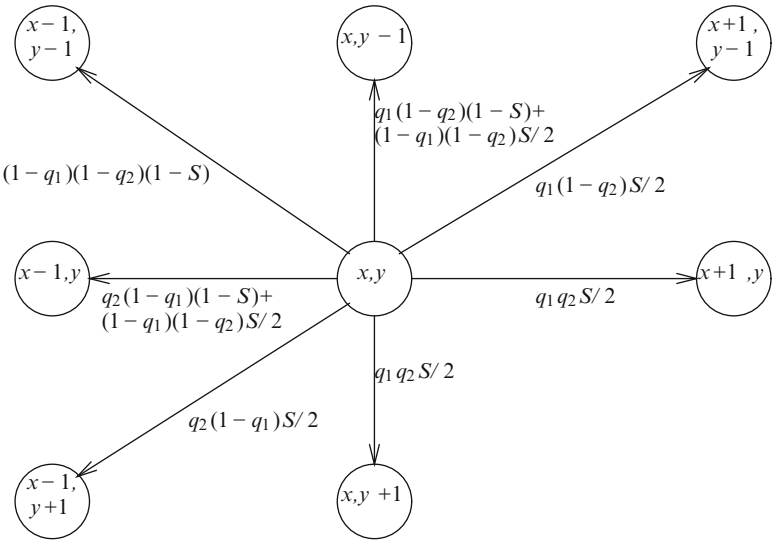


Fig. 11.10 There are cells in both buffers

- New cells arrive at both buffers, there is a conflict, and the cell in buffer 2 is chosen for transmission.

- $(x, y) \rightarrow (x - 1, y + 1)$ if
  - A cell arrives at input 2, no cells arrive at input 1, there is a conflict, and the cell in buffer 1 is chosen for transmission.
- $(x, y) \rightarrow (x, y + 1)$ if
  - New cells arrive at both buffers, there is a conflict, and the cell in buffer 1 is chosen for transmission.

11.3.3 Output Buffering

The analytical description of the switch with output buffering is easier than that with input buffering because in this case the number of cells in a buffer depends only on the properties of the arriving cells and is independent of the number of cells in the other buffer. Consequently, it is possible to analyze one output buffer in isolation.

Figure 11.11 presents the Markov chain of buffer 1 with output buffering. There are two possible state transitions if the buffer is idle:

- $0 \rightarrow 1$ if cells arrive at both inputs and both cells are directed to output 1.
- $0 \rightarrow 0$ otherwise.

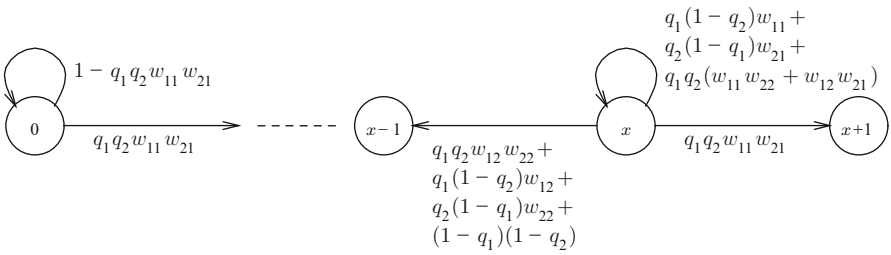


Fig. 11.11 Markov chain of buffer 1 with output buffering

There are three possible state transitions if the buffer is not idle:

- $x \to x - 1$ if a cell arrives at output 1.
- $x \to x$ if one cell arrives at output 1.
- $x \to x + 1$ if two cells arrive at output 1.

The probabilities of these state transitions are provided in Fig. 11.11.

11.3.4 Performance Parameters

In this section we compute some performance parameters in the case of input and output buffering assuming that the buffers are finite.

Mean Number of Cells in Buffers

Let P_{ij}, $i, j \geq 0$, be the steady-state probability of state (i, j) of a Markov chain describing a switch with input buffers. The mean number of cells in buffers 1 and 2 can be computed as

$$E_1 = \sum_{i \geq 0} \sum_{j \geq 0} i P_{ij},$$

$$E_2 = \sum_{i \geq 0} \sum_{j \geq 0} j P_{ij}.$$

Similarly, let $P_i^{(1)}$ and $P_i^{(2)}$, $i \geq 1$, be the steady-state probability of having i cells in buffers 1 and 2, respectively, in Markov chains describing a switch with output buffers. The mean number of cells in buffers 1 and 2 can be computed as

$$E_1 = \sum_{i \geq 1} i P_i^{(1)},$$

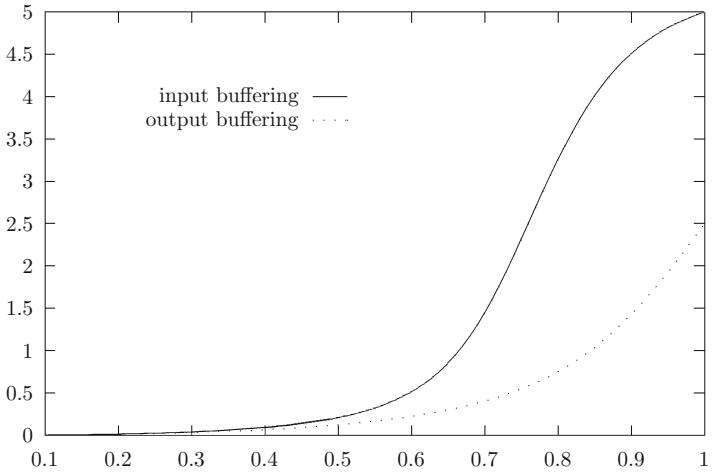


Fig. 11.12 Average number of cells with input and output buffering

$$E_2 = \sum_{i \geq 1} i P_i^{(2)}.$$

Figure 11.12 plots the average buffer content as a function of the arrival probability, $q = q_1 = q_2$, for the input and output buffer models, where the buffer length is limited to 5 and $w_{11} = w_{21} = 0.5$.

Throughput

The throughput (δ) is the mean number of cells the switch transmits in a time slot. In the case of input buffering, we can compute the throughput following the same division of the states of the Markov chain. Denoting the stationary probability of the four parts by P_{00}, P_{x0}, P_{0y}, P_{xy}, the throughput is

$$\begin{aligned}\delta = {} & P_{00}[1 \times (q_1(1-q_2) + q_2(1-q_1) + q_1 q_2 S) + 2 \times q_1 q_2 (1-S)] \\ & + \sum_{x \geq 1} P_{x0}[1 \times ((1-q_2) + q_2 S) + 2 \times q_2 (1-S)] \\ & + \sum_{y \geq 1} P_{0y}[1 \times ((1-q_1) + q_1 S) + 2 \times q_1 (1-S)] \\ & + \sum_{x \geq 1, y \geq 1} P_{xy}[1 \times S + 2 \times (1-S)],\end{aligned}$$

where we detail the cases with one-cell and with two-cell transmission.

In the case of output buffering, the throughput is

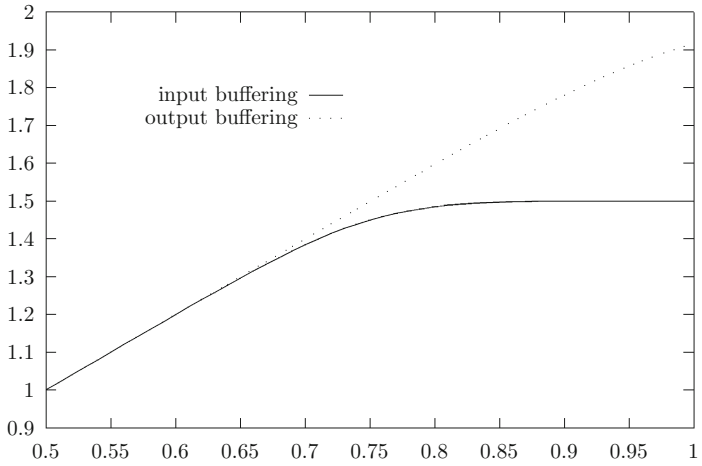


Fig. 11.13 Throughput

$$\delta = P_0^{(1)}\left(q_1(1-q_2)w_{11} + q_2(1-q_1)w_{21} + q_1q_2(1-w_{12}w_{22})\right) + \sum_{x\geq 1} P_x^{(1)}$$
$$+P_0^{(2)}\left(q_1(1-q_2)w_{12} + q_2(1-q_1)w_{22} + q_1q_2(1-w_{11}w_{21})\right) + \sum_{x\geq 1} P_x^{(2)}.$$

Figure 11.13 plots the throughput as a function of cell arrival probability, $q = q_1 = q_2$, for input and output buffering when the buffer length is limited to 5 and $w_{11} = w_{21} = 0.5$

In accordance with intuitive expectations, the throughput with input buffering is less than that with output buffering. As the arrival probability tends to 1, the throughput tends to 1.5 in the case of input buffering. A quick intuitive explanation of this property is as follows. If packets arrive at each time slot, the buffers are always busy, $\sum_{x\geq 1, y\geq 1} P_{xy}$ tends to 1, and δ tends to $1 \times S + 2 \times (1 - S)$ where $S = 1/2$.

11.3.5 Output Buffering in N × N Switch

Let us consider a single output of an $N \times N$ switch with output buffering and assume that cells arrive at the N input ports according to N independent identical Bernoulli processes. The probability that a packet arrives at a time slot is p. The arriving cells are directed to the output ports according to independent uniform distributions.

The number of cells directed to a tagged output port is binomially distributed:

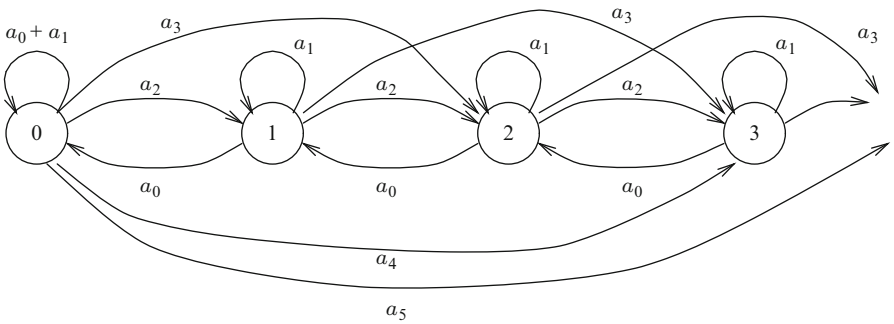


Fig. 11.14 Markov chain modeling a switch with output buffering

$$a_i = \mathbf{P}(i \quad \text{cells arrived in the time slot}) = \binom{N}{i} (p/N)^i (1 - p/N)^{N-i}.$$

Figure 11.14 shows the transition probability graph of a Markov chain describing the number of cells in a tagged output port. This Markov chain can also be described by an evolution equation of type II:

$$X_{n+1} = (X_n - 1 + Y_{n+1})^+,$$

and its state transition probability matrix is

$$\mathbf{\Pi} = \begin{bmatrix} a_0 + a_1 & a_2 & a_3 & a_4 & \cdots & a_{k+1} & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots & a_k & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots & a_{k-1} & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots & a_{k-2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}. \tag{11.1}$$

The stationary probabilities satisfy the following linear equations:

$$p_0 = p_0 a_0 + p_0 a_1 + p_1 a_0, \tag{11.2}$$

$$p_k = p_k a_1 + p_{k+1} a_0 + \sum_{i=0}^{k-1} p_i a_{k+1-i} = \sum_{i=0}^{k+1} p_i a_{k+1-i}. \tag{11.3}$$

Let us introduce the following z-transform functions

$$P(z) = \sum_{k=0}^{\infty} p_k z^k, \quad A(z) = \sum_{k=0}^{\infty} a_k z^k.$$

From Eqs. (11.2) and (11.3) we have

$$\begin{aligned}
P(z) &= p_0a_0 + p_0a_1 + p_1a_0 + \sum_{k=1}^{\infty}\sum_{i=0}^{k+1} p_i a_{k+1-i} z^k \\
&= p_0a_0 + \sum_{k=0}^{\infty}\sum_{i=0}^{k+1} p_i a_{k+1-i} z^k \\
&= p_0a_0 + \sum_{k=0}^{\infty}\sum_{i=1}^{k+1} p_i a_{k+1-i} z^k + \sum_{k=0}^{\infty} p_0 a_{k+1} z^k \\
&= p_0a_0 + \sum_{i=1}^{\infty} p_i \sum_{k=i-1}^{\infty} a_{k+1-i} z^k + z^{-1}\sum_{k=0}^{\infty} p_0 a_{k+1} z^{k+1} \\
&= p_0a_0 + z^{-1}\sum_{i=1}^{\infty} p_i z^i \sum_{l=0}^{\infty} a_l z^l + z^{-1} p_0 \sum_{m=1}^{\infty} a_m z^m \\
&= p_0a_0 + z^{-1}(P(z) - p_0)A(z) + p_0 z^{-1}(A(z) - a_0),
\end{aligned}$$

whence

$$P(z) = \frac{(1 - z^{-1})p_0a_0}{1 - z^{-1}A(z)} = p_0a_0\frac{(z-1)}{z - A(z)}.$$

Considering that $\lim_{z\to 1} P(z) = 1$ and applying l'Hospital's rule we have

$$1 = p_0a_0\frac{1}{1 - A'(z)}\,|_{z=1},$$

where $A'(1)$, the mean number of cells arriving at the tagged output in a time slot, can be computed;

$$p_0a_0 = 1 - A'(z) = 1 - p,$$

and using this

$$P(z) = \frac{(1-p)(z-1)}{z - A(z)} = \frac{(1-p)(1-z)}{A(z) - z}.$$

To check the obtained results we set $N = 2$. In this case the probability that i cells arrive at the tagged output post is

$$a_i = \binom{2}{i}\left(\frac{p}{2}\right)^i \left(1 - \frac{p}{2}\right)^{2-i},$$

whence

$$A(z) = \left(1 - \frac{p}{2} + z\frac{p}{2}\right)^2$$

and

$$P(z) = \frac{(1-p)(1-z)}{\left(1-\frac{p}{2}+z\frac{p}{2}\right)^2 - z}.$$

The probability that the buffer is idle, p_0, is easily obtained from the transform domain expression

$$p_0 = P(z) \mid_{z=0} = \frac{1-p}{\left(1-\frac{p}{2}\right)^2}.$$

This result can be checked easily by considering that for $N = 2$ the Markov chain has a birth-death structure with forward probability a_2 and backward probability a_0. From this Markov chain we have

$$p_0 = \frac{1}{1+\sum_{k=1}^{\infty}\left(\frac{a_2}{a_0}\right)^k} = \frac{1-p}{\left(1-\frac{p}{2}\right)^2}.$$

11.3.6 Throughput of N × N Switch with Input Buffering

The end of Sect. 11.3.4 shows the throughput computation of a 2×2 switch with input buffering. In this section we compute the throughput of larger switches, $N > 2$, with input buffering.

We assume that the cells are indistinguishable, the switch chooses one of the cells in conflict with independent uniform distribution, and the cells are directed to output ports in a uniformly distributed manner.

To obtain the maximal throughput of the system, we assume that cells are waiting in each buffer of the switch, i.e., none of the input buffers is idle.

We use the following notations:

- Let R_m^i be the number of cells that are at the head of a buffer at time m, are directed to output i, and are not forwarded due to collision.
- Let A_m^i be the number of cells that arrive at the head of a buffer at time m and are directed to output i.

R_m^i is a DTMC. It can be described by the evolution equation

$$R_m^i = (0, R_{m-1}^i + A_m^i - 1)^+,$$

where the sum on the right-hand side is reduced by 1 due to the cell that is transmitted to output i. A_m^i follows a binomial distribution

$$\mathbf{P}\left(A_m^i = k\right) = \binom{F_{m-1}}{k}(1/N)^k(1-1/N)^{F_{m-1}-k}, \qquad k = 0, 1, \cdots, F_{m-1}, \tag{11.4}$$

Table 11.1 Per output port throughput as a function of N

N	Throughput
1	1.0000
2	0.7500
3	0.6825
4	0.6553
5	0.6399
6	0.6302
7	0.6234
8	0.6184
∞	0.5858

where the number of new cells at the head of the buffer is

$$F_{m-1} = N - \sum_{i=1}^{N} R_{m-1}^{i}. \tag{11.5}$$

Equation (11.5) is based on the assumption that none of the input buffers is idle. Due to the same assumption the number of new cells arriving at the head of the buffers is equal to the number of cells successfully transmitted in a time slot. Consequently, the throughput output i is $\delta^i = \lim_{m\to\infty} E(A_m^i)$.

The parameters of the binomial distribution are F_{m-1} and $1/N$ since there are F_{m-1} new cells at the heads of the buffers and they choose their destination according to a uniform distribution.

With all elements of the evolution equation defined it is possible compute the stationary distribution of the Markov chain numerically. From the stationary distribution we also have

$$E(R^i) = \lim_{m\to\infty} E(R_m^i).$$

Taking the expectation of A^i based on Eq. (11.4) and both sides of Eq. (11.5) we get

$$E(A^i) = E(F)/N \text{ and } E(F) = N - \sum_{i=1}^{N} E(R^i) = E(F) = N - NE(R),$$

where we utilized the symmetry of the uniform output selection in the last step. Introducing $\delta = \delta^i$ $E(A) = E(A^i)$ $E(R) = E(R^i)$ we get

$$\delta = E(A) = 1 - E(R).$$

For any finite N the previously discussed numerical method results in the throughput of one output port. Table 11.1 presents the throughput as a function of N. For $N = 2$ we already computed the result in Sect. 11.3.4, but there we computed the throughput of the whole switch, not for one output port.

When $N \to \infty$, the same evolution equation holds, but A_m^i tends to be Poisson distributed with the parameter δ. At the limit we obtain a Markov chain with the same structure as that in Eq. (11.1). Following the same transform domain analysis and using $A(z) = e^{\delta(z-1)}$ we obtain $E(R)$ and from $\delta = 1 - E(R)$ the limiting throughput of the switch.

11.4 Conflict Resolution Methods of Random Access Protocols

One of the main functions of medium access control (MAC) is to share common resources between randomly arriving customer requests. Different random access protocols are developed for this purpose. In the case of random access protocols, several users try to communicate through a common transmission channel. The users do not know the activity of the others. In this kind of environment stable communication requires the application of a protocol that under a system-dependent load level ensures

- Stable communication (with finite mean delay),
- Transmission of all packets,
- Fairness (users obtain the same service).

These protocols work based on the information available about the state of the users. The following procedures are different members of the set of random access protocols, which differ (1) in the way users are informed about the status of the common channel and, indirectly, the activity of the other users and (2) in the design goals to adopt to the alternation of the traffic load.

11.4.1 ALOHA Protocol

The ALOHA protocol [2] is the simplest random access protocol. It was developed for simple radio communication between radio terminals and a central station. It uses two radio channels, one of which is used by the terminals for communication to the central station and the other for communication from the central station to all terminals (Fig. 11.15).

If more than one terminal sends a message to the central station, then the signals interfere and the central station cannot receive any correct messages. This case is referred to as a collision of messages. Collision can only happen in the first radio channel since the second radio channel is used only by the central station. Successfully received messages are acknowledged by the central station. The terminals are informed about the success of their message transmission by these acknowledgements. If no acknowledgement arrives within a given deadline, then the

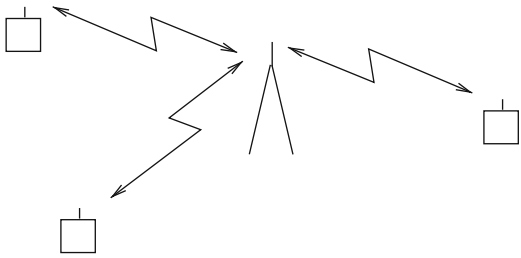


Fig. 11.15 Radio terminal system

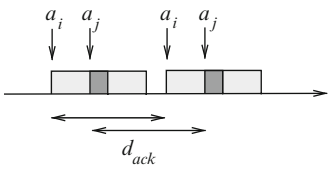


Fig. 11.16 Packet retransmission without random delay

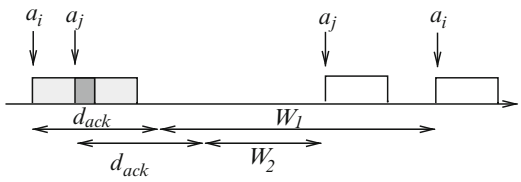


Fig. 11.17 Packet retransmission with random delay

terminal assumes that the transmission failed. The ALOHA protocol is designed to ensure communication in this system.

ALOHA functions as follows. As soon as a terminal has a new packet to transmit it starts sending it away without any attention to the activity of the other terminals. If no acknowledgement arrives within a given deadline, the terminal assumes that the message collided and it switches to message retransmission mode. Terminals in message retransmission mode are referred to as blocked terminals. In this mode the terminal retransmits the message until it receives an acknowledgement about successful transmission.

If after an unsuccessful transmission a terminal were to retransmit the message immediately, then there would be multiple collisions among the same set of terminals (Fig. 11.16).

To avoid these multiple collisions in the same set of terminals, the terminals wait for a random amount of time before retransmitting messages (Fig. 11.17). The cited figures distinguish between the period of collision (dark gray) and the additional period that is wasted due to the collision (light gray).

The quantitative behavior of a system is straightforward. If the terminals choose a large random delay, then the probability of consecutive collisions with the same set of terminals is low, but the time to successful transmission is high due to the long delay till retransmission. In the opposite case, if the delay is short, then the probability of subsequent collisions is higher, and this could cause a longer time to successful transmission due to the high number of repeated transmission attempts. The optimal behavior of the system is somewhere between these extremes.

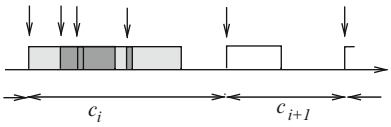


Fig. 11.18 Cycles of busy and idle periods

The modeling and analysis of ALOHA systems has been a well-studied area since 1970s. It is still an interesting research area because several random access protocols that were subsequently introduced contain elements of the basic ALOHA protocol (as is detailed in the following sections).

There exists a wide variety of performance studies. These studies differ in their assumptions about the behavior of users and systems. It is practically impossible to analyze the simplest ALOHA protocol in all its minute technical details. To reduce the complexity of the models, several simplifying assumptions are used. The obtained simplified models often closely approximate real system behavior.

In the following sections we introduce some of the simplest models of the basic ALOHA system and their analysis.

Continuous Time ALOHA System

We adopt the following modeling assumptions:

- The aggregate arrival process of new and retransmitted messages is a Poisson process with parameter λ.

  This model is not a correct model of the aggregate arrival process (in general) but there are several cases (e.g., the number of blocked terminals is negligible compared to the number of all terminals) when it properly approximates the real system behavior. This kind of model, where the arrivals of the new and the retransmitted messages are considered in an aggregate flow, is referred to as a zero-order model.
- The length of the messages is fixed and the time of a message transmission is T.

With the zero-order model we evaluate which portion of the new and repeated messages, which arrive according to a Poisson process with parameter λ, is transmitted successfully, and what is the related transmission delay and collision probability.

As shown in Fig. 11.18 we divide the time axes according to the busy and idle periods of the common channel.

The probability of a successful message transmission in a busy period equals the probability that after the beginning of a busy period the next message arrives later than T. Its probability is $e^{-\lambda T}$.

In order to determine the long-term idle ratio, successful busy and unsuccessful busy periods, we determine the average length of these periods. The interarrival time in a Poisson process with parameter λ is exponentially distributed with parameter λ. The length of an idle period is the remaining time of an exponentially distributed

interarrival time, which is exponential again with the same parameter. Thus the mean length of an idle period is $1/\lambda$.

The length of a successful busy period is T. The difficult question is the length of the unsuccessful period. An unsuccessful busy period is composed of $N-1$ ($N \geq 2$) interarrival intervals shorter than T and a final interval of length T. The case where $N = 1$ is the successful busy period. Due to the memoryless property of Poisson processes we can compute the number of colliding messages during the unsuccessful busy period independently of the length of the interarrival times,

$$Pr(N = n) = (1 - e^{-\lambda T})^{n-1} e^{-\lambda T}.$$

The CDF of the length of an interarrival interval shorter than T, denoted as U, is

$$F_U(t) = Pr(U < t) = Pr(\tau < t | \tau < T) = \begin{cases} \dfrac{1 - e^{\lambda t}}{1 - e^{\lambda T}} & 0 < t < T, \\ 1 & T < t, \end{cases}$$

whence $E(U) = \dfrac{1 - e^{-\lambda T} - \lambda T e^{-\lambda T}}{\lambda(1 - e^{-\lambda T})}$. Consequently, in a cycle composed of a busy and an idle period

- The mean length of the idle period is $E(I) = 1/\lambda$,
- The probability of a successful message transmission is $E(S) = e^{-\lambda T} T$, and
- The mean length of an unsuccessful busy period is

$$E(L) = \sum_{n=2}^{\infty} Pr(N = n) \Big((n-1)E(U) + T \Big) = \frac{1 - e^{-\lambda T} - \lambda T e^{-\lambda T}}{\lambda e^{-\lambda T}}.$$

System utilization is characterized by the portion of time associated with successful message transmission:

$$\rho = \frac{E(S)}{E(I) + E(S) + E(L)} = \lambda T e^{-2\lambda T}.$$

It can be seen that utilization depends only on the λT product. The maximum of utilization is obtained through the derivative of ρ as a function of λT. The maximum is found at $\lambda T = 0.5$ and is $\rho = 1/2e \sim 0.18394$. Figure 11.19 shows that utilization decreases significantly as the load increases above 0.5; consequently these systems should be operated with a load lower than 0.5.

The mean number of arriving messages in a Δ long interval is $\lambda\Delta$. In the same interval the mean time of successful message transmission is $\rho\Delta$. During this interval the mean number of successfully transmitted messages is $\rho\Delta/T$. The ratio between the number of successfully transmitted messages and the number of all message transmission attempts, which is the mean number of transmission attempts per message, is $E(R) = \lambda T/\rho = e^{2\lambda T}$.

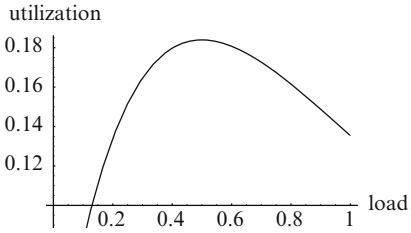


Fig. 11.19 Utilization ρ as a function of load λT

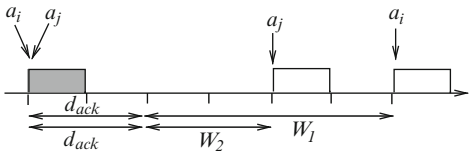


Fig. 11.20 Slotted ALOHA system

Having the mean number of transmission attempts per message we can compute the message transmission delay:

$$E(D) = E\left(\sum_{r=1}^{\infty} Pr(R=r)\left(rT + \sum_{i=1}^{r-1} d_{\text{ack}} + W_i\right)\right)$$
$$= E(R)T + (E(R)-1)(d_{\text{ack}} + E(W)),$$

where d_{ack} is the time a terminal waits for message acknowledgement and W is the random delay spent before message retransmission.

Discrete-Time (Slotted) ALOHA System

The main disadvantage of continuous-time ALOHA systems is that the wasted time when messages collide is large. This phenomenon can be seen in Figs. 11.16 and 11.18. The dark gray period denotes the overlapping intervals of colliding messages, while light gray periods are additional wasted time intervals that cannot be used for useful message transmission.

With a simple modification of the ALOHA system this additional wasted time interval can be avoided. If all terminals work in a synchronized manner and initiate message transmission only at the beginning of time slots, then the length of time the colliding messages occupy the common channel reduces to T. Naturally in this system the delay of a message retransmission should be an integer multiple of the time slot, T. This system is commonly referred to as a slotted ALOHA system (Fig. 11.20).

The zero-order model of a slotted ALOHA system assumes that the terminals generate a Poisson distributed number of new and repeated messages in a time slot, where the parameter of the Poisson distribution is λT. This model of message arrivals is similar to the zero-order model of continuous-time ALOHA systems

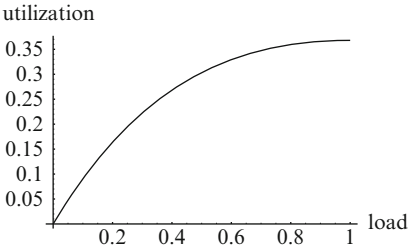


Fig. 11.21 Utilization of slotted ALOHA system

assuming that the messages are generated continuously according to a Poisson process, but the messages generated during a time slot are delayed till the beginning of the next time slot.

With these assumptions, utilization of the zero-order model of a slotted ALOHA system can be computed based on the analysis of a single time slot. Let N be the number of packets generated in a time slot. In this case,

$$\rho = Pr(\text{successful message transmission}) = Pr(N = 1) = \lambda T \mathrm{e}^{-\lambda T}.$$

Maximum utilization is obtained at $\lambda T = 1$, and it is $\rho = 1/e \sim 0.367879$. Compared to the continuous-time ALOHA system, the optimal throughput doubles and the aggregated load (new and repeated messages) can be increased to the capacity of the system ($\lambda T = 1$), as is plotted in Fig. 11.21.

The mean number of retransmission attempts, R, can be computed as the ratio between the successfully transmitted and all messages:

$$E(R) = \frac{E(N)}{E(\text{successfully transmitted messages})} = \frac{\lambda T}{\lambda T \mathrm{e}^{-\lambda T}} = \mathrm{e}^{\lambda T}.$$

Similar to the continuous-time case, the message transmission delay is

$$\begin{aligned} E(D) &= E\left(\sum_{r=1}^{\infty} Pr(R = r) \left(rT + \sum_{i=1}^{r-1} d_{\text{ack}} + W_i \right) \right) \\ &= E(R)T + (E(R) - 1)(d_{\text{ack}} + E(W)). \end{aligned}$$

The more complex models of the ALOHA system distinguish the states of the terminals (message generation, new message transmission attempt, waiting random delay, message retransmission attempt) and characterize the arrival of new and repeated messages according to those states [27].

In contrast to the terminology of ALOHA systems, the general terminology of random access protocols refers to stations instead of terminals and packets instead of messages.

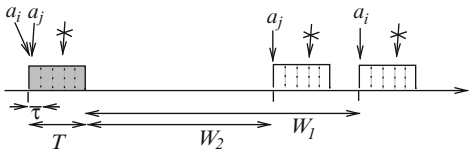


Fig. 11.22 CSMA system

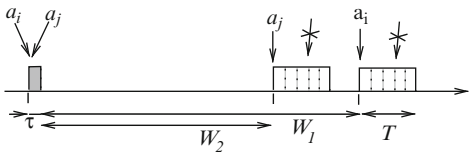


Fig. 11.23 CSMA/CD system

11.4.2 CSMA and CSMA/CD Protocols

The more advanced random access protocols aim to enhance channel utilization based on the information available for the stations by the given physical media.

In the introduction of the slotted ALOHA system we saw that the reduction of the time period while a collision makes the channel unavailable enhances the performance of the protocol. In the case of radio terminal systems where the terminals can be outside of each other's propagation range, it is hard to further reduce the ineffective time of the channel. In the case of wired systems, the stations can sense each other's signals, but not immediately; there is a propagation delay of the medium. This direct sensing of the stations can be used to enhance the performance of the multiple access protocol in the following two ways.

- If one station senses that another station is sending a packet when it has a packet to send, then the first station does not start sending the packet.
- If by accident the packets of two stations collide (because they are sent within the propagation delay), then the stations can recognize that the packets collide and finish the useless packet transmission immediately.

The first way is referred to as *carrier sense multiple access* (CSMA) and the second as *collision detection* (CD).

Figures 11.22 and 11.23 demonstrate the behavior of the CSMA and the CSMA/CD systems. In these systems the time is slotted and the time unit is the maximal propagation delay between the most remote stations, τ. Collision can happen only among packets transmitted within the same slot because in the next time slot all stations are aware of the busy state of the channel. A station can initiate packet transmission only if the channel is idle. In the case of CSMA without CD, colliding packets are transmitted completely. Thus a significant portion of the channel capacity is lost (Fig. 11.22). In the case of CSMA with CD, the collision is recognized within one time slot and packet transmission is finished immediately (Fig. 11.23).

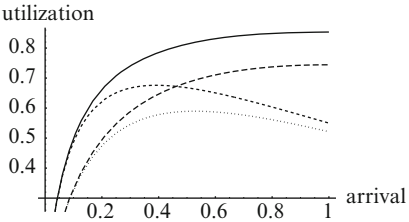


Fig. 11.24 Utilization of slotted CSMA and CSMA/CD systems

Performance of Slotted CSMA System

We analyze the zero-order model of a system by the analysis of the intervals between consecutive packet transmission attempts. The beginning of these intervals is indicated by the arrows below the time axes in Figs. 11.22 and 11.23. It can also happen that, in contrast to the figures, there is no idle period between two consecutive packet transmission attempts. According to the zero-order model of a system, we assume that after an idle time slot or the last time slot of a successful packet transmission there is a Poisson distributed number of (new and repeated) packet transmissions initiated with parameter $\tau\lambda$. That is, in contrast with the previously discussed zero-order models, the state of the channel affects the arrival process of the packets. Packets can arrive in the aforementioned time slots and not otherwise. The success of the packet transmission depends on the number of arriving packets, N:

$$Pr(\text{succesfull packet transmission}) = Pr(N = 1|N \geq 1) = \frac{\lambda\tau e^{-\lambda\tau}}{1 - e^{-\lambda\tau}}.$$

After a successful or colliding packet transmission the channel remains idle until the next packet arrives. Let I denote the number of idle time slots until the next packet arrives. Due to the memoryless property of the arrival process, I is geometrically distributed. $Pr(I = i) = e^{-\lambda\tau i}(1 - e^{-\lambda\tau})$. Consequently, in an interval between consecutive packet transmission attempts

- The mean length of the idle period is $E(I)\tau = \frac{\tau\, e^{-\lambda\tau}}{1-e^{-\lambda\tau}}$,
- The mean length of successful packet transmission is $E(S) = \frac{T\ \lambda\tau e^{-\lambda\tau}}{1-e^{-\lambda\tau}}$, and
- The mean length of unsuccessful packet transmission is $E(L) = \frac{T\ (1-(1+\lambda\tau)e^{-\lambda\tau})}{1-e^{-\lambda\tau}}$.

Finally, utilization is obtained as

$$\rho = \frac{E(S)}{E(I) + E(S) + E(L)} = \frac{T\ \lambda\tau e^{-\lambda\tau}}{T(1 - e^{-\lambda\tau}) + \tau e^{-\lambda\tau}}.$$

Figure 11.24 plots utilization as a function of $\lambda\tau$ when $\tau/T = 0.2$ (dotted line) and when $\tau/T = 0.1$ (short dashed line). It can be seen that the probability of

collision is lower and utilization is higher in the case of shorter propagation time ($\tau/T = 0.1$). In any case, utilization reaches an optimum and starts decreasing when the load is increasing. The optimal load level depends on the τ/T ratio.

Performance of Slotted CSMA/CD System

The zero-order model of the CSMA/CD system is very similar to that of the CSMA system. It differs only in the time of unsuccessful packet transmission, which is shorter due to collision detection: $E(L) = \frac{\tau\ (1-(1+\lambda\tau)e^{-\lambda\tau})}{1-e^{-\lambda\tau}}$. As a result, utilization is

$$\rho = \frac{E(S)}{E(I) + E(S) + E(L)} = \frac{T\ \lambda\tau e^{-\lambda\tau}}{T\lambda\tau e^{-\lambda\tau} + \tau(1-\lambda\tau e^{-\lambda\tau})}.$$

Figure 11.24 plots the utilization of a CSMA/CD system as a function of load, $\lambda\tau$, together with that of the CSMA system. The propagation delay is $\tau/T = 0.2$ (long dashed line) and $\tau/T = 0.1$ (solid line). Also, in these cases the shorter propagation delay increases utilization. In contrast with the CSMA system, utilization is continuously increasing with the load due to the efficient utilization of the channel.

Slotted Persistent CSMA/CD System

Up to now we have not discussed the behavior of a station when it has a packet to transmit but the channel is busy. Indeed we implicitly assumed that these stations assumed that their packet collided and delayed the next packet retransmission attempt accordingly. This behavior is referred to as nonpersistent station behavior.

The stations sense the channel and know the history of the channel state from which they can compute when the packet under transmission finishes. Knowing this information, a station with a packet to transmit can reduce the packet retransmission time by attempting a packet transmission immediately when the channel becomes idle next. This behavior is referred to as persistent station behavior.

In the zero-order model of persistent CSMA systems we assume that in each τ long time slot stations generate a Poisson distributed number of new and repeated packets, and those stations that generate packets during a packet transmission attempt to transmit packets when the channel becomes idle next.

The analysis of this system is based on the analysis of successful (S), colliding (L), and idle (I) intervals because the system behavior is memoryless at the beginning of these periods. The mean length of these intervals is as follows: $E(S) = T$, $E(L) = \tau$, and $E(I) = \frac{1}{1-e^{-\lambda\tau}}$.

To compute the utilization we also need to know how often these intervals occur. The following transition probability matrix defines the probability of the occurrence of various consecutive intervals:

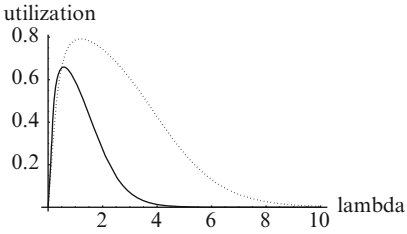


Fig. 11.25 Utilization of slotted persistent CSMA/CD system

$$\Pi = \begin{array}{|cccc|} \hline S & L & I & \\ \hline P(\lambda T,1) & P(\lambda T,>1) & P(\lambda T,0) & S \\ P(\lambda\tau,1) & P(\lambda\tau,>1) & P(\lambda\tau,0) & L \\ \dfrac{P(\lambda\tau,1)}{P(\lambda\tau,>0)} & \dfrac{P(\lambda\tau,>1)}{P(\lambda\tau,>0)} & 0 & I \\ \hline \end{array}$$

where $P(a,i) = \mathrm{e}^{-a}a^i/i!$ and $P(a,>i) = \sum_{j=i+1}^{\infty} P(a,j)$. This is a DTMC whose stationary solution is obtained by the solution of the linear system of equations $\pi\Pi = \pi$, $\pi_S + \pi_L + \pi_I = 1$, where $\pi = \pi_S, \pi_L, \pi_I)$. Given the stationary probabilities π_S, π_L, and π_I, the utilization is

$$\rho = \frac{\pi_S E(S)}{\pi_S E(S) + \pi_L E(L) + \pi_I E(I)} = \frac{\lambda\tau\ \lambda T}{1 - \lambda T\mathrm{e}^{-\lambda T} + \lambda\tau\mathrm{e}^{-\lambda T}\left(1 - \lambda\tau + \lambda T(1-\mathrm{e}^{\lambda\tau})\right) + \lambda\tau\left(\mathrm{e}^{\lambda\tau} + \lambda T - 1\right)}.$$

Figure 11.25 plots utilization as a function of load, λ, with propagation delay $\tau/T = 0.1$ (dotted line) and with $\tau/T = 0.2$ (solid line). As with the previous cases, shorter propagation delays result in a lower probability of collision and better utilization. Utilization decreases when the load is high. This is because the probability of collision after a successful packet transmission becomes very high due to persistent station behavior.

In summary, nonpersistent behavior is beneficial when the delay due to a collision at the end of a successful packet transmission is less than the normal message retransmission delay. At low load levels, persistent behavior decreases the delay (because the probability of collision is low), while at high load levels nonpersistent behavior performs better.

There is a continuous transition between persistent and nonpersistent behaviors. It is obtained when a station follows persistent behavior with probability p and nonpersistent behavior with probability $1 - p$. This behavior is referred to as p-persistent behavior. Obviously $p = 1$ results in persistent and $p = 0$ nonpersistent behavior. For a given load level we can optimize the system utilization by setting p to an optimal value.

11.4.3 IEEE 802.11 Protocol

One of the most commonly used ways to wirelessly access computer networks currently is the wireless fidelity (WF or wifi), which is defined in the IEEE 802.11 standard. The core of this rather complicated protocol is also an enhanced version of the slotted ALOHA protocol. The IEEE 802.11 protocol [1] is designed to meet the following requirements:

- The random delay for packet transmission is bounded.
- The protocol operates in a wide range of traffic load and adapts the actual level of traffic load.
- If large packets are transmitted, priority is given to the completion of the already started packets.

According to these requirements the ALOHA protocol is modified as follows: [11, 32].

- At a collision a station draws a random number uniformly distributed between 1 and M_i and retransmits the collided packet after the expiration of that many slots.
- The upper limit of the uniformly distributed delay depends on the number of unsuccessful transmission trials. After the first collision this value is $M_1 = 8$ and it doubles after each consecutive collision $M_i = 8 * 2^{i-1}$ until a predefined upper limit, $M_{\max}$, is reached.
- Large packets are transmitted in several small segments. In the case of an ALOHA system these segments are transmitted one by one and each of them can collide with other segments and got delayed by the collision resolutoin procedure. Thus the packet transmission delay, which is determined by the largest delay of the segments, can be very high. IEEE 802.11 reduces the packet transmission delay by giving priority to the consecutive segments of a packet under transmission. Thus only the first segment of a packet participates in the contention and other segments are transmitted with high priority. The protocol implements this feature by the introduction of two different delays. The stations in contention consider the medium available if it is idle for a *distributed interframe space* (DIFS) period, while packet segments can be sent within a *short interframe space* (SIFS) period, which is shorter than a DIFS period.

The IEEE 802.11 protocol is built on a so-called basic access method, which is practically identical with the ALOHA protocol and combines with various reservation methods. One of these reservation methods is the aforementioned DIFS- and SIFS-based packet transmission. The mathematical description of these complex reservation methods is rather complex. In this section we present an analytical model of the basic access method, which is introduced in [11]. This model is also based on a simplifying assumption. It assumes that there are so many independently working stations in the system that a packet transmission trial will be unsuccessful with probability p in each time slot independent of the past history of the system. Furthermore, to compute the maximal throughput we assume that stations always have packets to transmit.

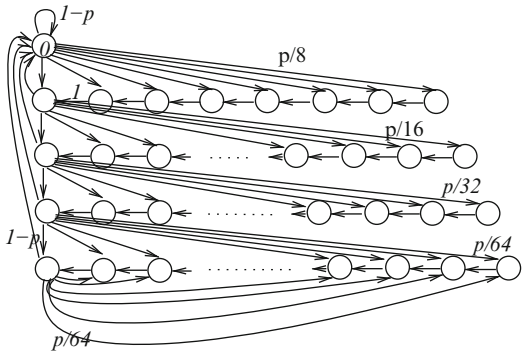


Fig. 11.26 Markov chain describing basic access method of IEEE 802.11 standard

With this assumption we can describe the behavior of a station with a DTMC. The state of the Markov chain describes the phase of the actual packet transmission attempt. Figure 11.26 shows the state-transition graph of this Markov chain. State 0 indicates that the station just finished transmitting a packet and is trying to transmit the next one in the next time slot. If it is unsuccessful, which will happen with probability p, then it draws a uniformly distributed random sample between 1 and $M_1 = 8$. Transitions to the right with probability 1 describe a situation where the station waits until the given delay expires. When the chain arrives at the leftmost state, it attempts to transmit the packet again and go back to state 0, or it moves to the next row, etc. In this Markov chain the retransmission delay is limited to $M_{\max} = 64$. The time between two consecutive visits to state 0 represents the packet transmission delay. The throughput of the station is p_0 and the mean packet transmission delay is $1/p_0$ if p_0 is the stationary probability of state 0 in this Markov chain.

11.5 Priority Service Systems

Priority systems appear in different fields [37,39,45,91]. Several aspects of telecommunications, data management, planning of computer networks, organization of health services, and automatization of production processes could be mentioned. For example, in mobile cellular networks the coverage area is partitioned into cells, and each cell can serve at most c simultaneous communications and use some channels from other cells. There are calls initiated by subscribers from the cell and handover calls from others. Handover calls already use the network resources and should be prioritized with respect to new calls. Different approaches are possible, e.g., a special channel or a priority queue of handover calls.

The problem may be formulated as follows. Customers of different types enter the service system, each of them belonging to a priority class, indexed by a subscript. We assume that customers with a lower index have higher priority in service; this way customers with a lower index can leave the queue earlier than customers with a higher index which were already in the queue at the arrival of customers with a lower index. There are two possible cases: either the entry of customers of higher priority

does not interrupt the actual service with lower priority customers or immediately starts its service (in the first case we speak of relative, in the second case of absolute priority). In the second case we have again two choices, whether or not the work up to this moment will be taken into account in the future. With respect to the first possibility, one must complete the residual service, for the second one must complete the residual service, for the second one must complete the whole service later, when the higher proirity customers are served. Both of these possibilities occur in computer systems. For example, the results of computations are either regularly saved or not. In the first case results are not lost at a system error. Similar situations appear in other fields. When a disaster occurs, one must first to divert the danger and after that to deal with less urgent tasks. For example, a dentist must first see patients who are in pain; other patients can wait.

We will consider service systems with two Poisson arrival processes where the service time will have exponential and general distributions. In the exponential case we will follow the usual method – find the system of differential equations describing the functioning of system, solve it, and at $t \to \infty$ determine the equilibrium distribution. In the general case, we examine the virtual waiting time by means of the Laplace–Stieltjes transform; the approach is mainly based on the Pollaczek–Khinchin formula concerning waiting time (8.19) (e.g., [70]).

11.5.1 Priority System with Exponentially Distributed Service Time

Let us consider the following problem. We have m homogeneous servers, and two types of customers. Type i customers arrive to the system according to a Poisson process with parameter $\lambda_i (i = 1, 2)$. If upon the entry of a type 1 customer all servers are occupied, but some servers handle customers of the second type, then a server will change its service and the type 2 customer will be lost. Thus, customers of the second type may be lost not only if a type 2 customer arrives and all servers are occupied, but if customers of the first type show up as well. First type customers are refused only when there are customers of the same type.

The service times of type 1 and type 2 customers are exponentially distributed with parameters μ_1 and μ_2, respectively.

It is quite clear that the service of type 1 customers is denied if all servers were busy and there were no type 2 customers. Thus, the probability of loss of type 1 customers clearly equals

$$p_v = \frac{\frac{\rho_1^m}{m!}}{\sum_{i=0}^{m} \frac{\rho_1^i}{i!}}, \qquad \rho_1 = \frac{\lambda_1}{\mu_1}.$$

Let $p_{ij}(t)$ be the probability of the event that at moment t there are i type 1 and j type 2 customers being served ($0 \le i + j \le m$). Furthermore, let

$$p_{i.}(t) = \sum_{j=0}^{m-i} p_{ij}(t) \qquad \text{and} \qquad p_{.j}(t) = \sum_{i=0}^{m-j} p_{ij}(t).$$

The sum $\sum_{i+j=m} p_{ij}(t)$ is the probability of loss of a type 2 customer at moment t. The probability of loss of a type 2 customer during its service is

$$\sum_{i+j=m} p_{ij}(t) - p_{m0}(t).$$

11.5.2 Probabilities $p_{ij}(t)$

The differential equations determining $p_{ij}(t)$ are

$$p'_{00}(t) = -(\lambda_1 + \lambda_2)p_{00}(t) + \mu_1 p_{10}(t) + \mu_2 p_{01}(t); \tag{11.6}$$

if $1 \le i < m$, then

$$p'_{i0}(t) = -(\lambda_1+\lambda_2 + i\mu_1)p_{i0}(t)+\lambda_1 p_{i-1,0}(t) + (i+1)\mu_1 p_{i+1,0}(t) + \mu_2 p_{i1}(t), \tag{11.7}$$

$$p'_{m0}(t) = -m\mu_1 p_{m0}(t) + \lambda_1[p_{m-1,0}(t) + p_{m-1,1}(t)]; \tag{11.8}$$

in the case of $1 \le j < m$,

$$p'_{0j}(t) = -(\lambda_1+\lambda_2+j\mu_2)p_{0j}(t)+\lambda_2 p_{0,j-1}(t)+\mu_1 p_{1j}(t)+(j+1)\mu_2 p_{0,j+1}(t), \tag{11.9}$$

$$p'_{0m}(t) = -(\lambda_1 + m\mu_2)p_{0m}(t) + \lambda_2 p_{0,m-1}(t); \tag{11.10}$$

in the case of $i \ge 1$, $j \ge 1$, $i + j < m$,

$$\begin{aligned} p'_{ij}(t) = &-(\lambda_1 + \lambda_2 + i\mu_1 + j\mu_2)p_{ij}(t) + \lambda_1 p_{i-1,j}(t) \\ &+\lambda_2 p_{i,j-1}(t) + (i+1)\mu_1 p_{i+1,j}(t) + \mu_2 p_{i,j+1}(t); \end{aligned} \tag{11.11}$$

in the case of $i > 0$, $j > 0$, $i + j = m$, $i \ne m$, $j \ne m$,

$$p'_{ij}(t) = -(\lambda_1 + i\mu_1 + j\mu_2)p_{ij}(t) + \lambda_1[p_{i-1,j}(t) + p_{i-1,j+1}(t)] + \lambda_2 p_{i-1,j}(t). \tag{11.12}$$

Summing up Eqs. (11.6), (11.9), and (11.12) by j from 0 to m we obtain

$$p'_{0.}(t) = -\lambda_1 p_{0.}(t) + \mu_1 p_{1.}(t). \tag{11.13}$$

Summing up Eqs. (11.7), (11.11), and (11.12) by j from 0 to m, in the case $1 \le i < m$,

$$p'_{i.}(t) = -(\lambda_1 + i\mu_1)\, p_{i.}(t) + \lambda_1 p_{i-1,.}(t) + (i+1)\mu_1 p_{i+1,.}(t). \tag{11.14}$$

Equation (11.8) may be rewritten in the form

$$p'_{m.}(t) = -m\mu_1 p_{m.}(t) + \lambda_1 p_{m-1,.}(t). \tag{11.15}$$

The summation of Eqs. (11.6)–(11.8) by i leads to

$$p'_{.0}(t) = -\lambda_2[p_{.0}(t) - p_{m0}(t)] + \lambda_1 p_{m-1,1}(t) + \mu_2 p_{.1}(t).$$

Summing up Eqs. (11.9), (11.11), and (11.12) by i at $1 \le j < m$:

$$\begin{aligned} p'_{.j}(t) = &-(\lambda_2 + j\mu_2)[p_{.j}(t) - p_{m-j,j}(t)] + \lambda_2[p_{.,j-1}(t) - p_{m-j+1,j-1}(t)] \\ &+(j+1)\mu_2 p_{.,j+1}(t) - j\mu_2 p_{m-j,j}(t) - \lambda_1 p_{m-j,j}(t) + \lambda_1 p_{m-j-1,j+1}(t). \end{aligned}$$

Equation (11.10) may be written in the form

$$p'_{.m}(t) = -(\lambda_1 + m\mu_2)\, p_{.m}(t) + \lambda_2[p_{.m}(t) - p_{1,m-1}(t)].$$

From these equations one can see that for type 2 customers the situation is more complicated; type 1 customers play an essential role in the service process.

Let us consider the case $m = 1$. Then Eqs. (11.6)–(11.12) lead to the equations

$$\begin{aligned} p'_{00}(t) &= -(\lambda_1 + \lambda_2)\, p_{00}(t) + \mu_1 p_{10}(t) + \mu_2 p_{01}(t), \\ p'_{10}(t) &= -\mu_1 p_{10}(t) + \lambda_1[p_{00}(t) + p_{01}(t)], \\ p'_{01}(t) &= -(\lambda_1 + \mu_2)\, p_{01}(t) + \lambda_2 p_{00}(t). \end{aligned}$$

This system may be solved easily; the initial conditions are

$$p_{00}(0) = 1, \qquad p_{10}(0) = 0, \qquad p_{01}(0) = 0.$$

We have

$$p_{10}(t) = \frac{\lambda_1}{\lambda_1 + \mu_1}\left(1 - \mathrm{e}^{-(\lambda_1+\mu_1)t}\right),$$

$$\begin{aligned} p_{01}(t) = &\frac{\lambda_2\mu_1}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)} + \frac{\lambda_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2-\mu_1)}\mathrm{e}^{-(\mu_1+\mu_2)t} \\ &-\left(\frac{\lambda_2\mu_1}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)} + \frac{\lambda_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2-\mu_1)}\right)\mathrm{e}^{-(\lambda_1+\lambda_2+\mu_2)t}, \end{aligned}$$

$$p_{00}(t) = \frac{\mu_1(\lambda_1+\mu_2)}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)} + \frac{\lambda_1(\mu_2-\mu_1)}{(\lambda_1+\mu_1)(\lambda_2+\mu_2-\mu_1)}\mathrm{e}^{-(\lambda_1+\mu_1)t}$$
$$+\frac{\lambda_2}{\lambda_1+\mu_1}\left(\frac{\mu_1}{\lambda_1+\lambda_2+\mu_2}+\frac{\lambda_1}{\lambda_2+\mu_2-\mu_1}\right)\mathrm{e}^{-(\lambda_1+\lambda_2+\mu_2)t}.$$

Consequently,

$$p_{0.}(t) = p_{00}(t) + p_{01}(t) = \frac{\mu_1}{\lambda_1+\mu_1} + \frac{\lambda_1}{\lambda_1+\mu_1}\mathrm{e}^{-(\lambda_1+\mu_1)t},$$

$$p_{1.}(t) = p_{10}(t) = \frac{\lambda_1}{\lambda_1+\mu_1}\left(1-\mathrm{e}^{-(\lambda_1+\mu_1)t}\right),$$

$$p_{.0}(t) = \frac{\lambda_1(\lambda_1+\lambda_2+\mu_2)+\mu_1(\lambda_1+\mu_2)}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)} - \frac{\lambda_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2-\mu_1)}\mathrm{e}^{-(\lambda_1+\mu_1)t}$$
$$+\frac{\lambda_2}{\lambda_1+\mu_1}\left(\frac{\mu_1}{\lambda_1+\lambda_2+\mu_2}+\frac{\lambda_1}{\lambda_2+\mu_2-\mu_1}\right)\mathrm{e}^{-(\lambda_1+\lambda_2+\mu_2)t},$$

$$p_{.1}(t) = p_{01}(t) = \frac{\lambda_2\mu_1}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)} + \frac{\lambda_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2-\mu_1)}\mathrm{e}^{-(\lambda_1+\mu_1)t}$$
$$-\frac{\lambda_2}{\lambda_1+\mu_1}\left(\frac{\mu_1}{\lambda_1+\lambda_2+\mu_2}+\frac{\lambda_1}{\lambda_2+\mu_2-\mu_1}\right)\mathrm{e}^{-(\lambda_1+\lambda_2+\mu_2)t}.$$

The stationary probabilities at $t \to \infty$ are

$$p_{00} = \mu_1(\lambda_1+\mu_2)/\left((\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)\right),$$
$$p_{01} = p_{.1} = \lambda_2\mu_1/\left((\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)\right),$$
$$p_{10} = p_{1.} = \lambda_1/(\lambda_1+\mu_1),$$
$$p_{0.} = \mu_1/(\lambda_1+\mu_1),$$
$$p_{.0} = \frac{\lambda_1(\lambda_1+\lambda_2+\mu_2)+\mu_1(\lambda_1+\mu_2)}{(\lambda_1+\mu_1)(\lambda_1+\lambda_2+\mu_2)}.$$

It is interesting to note the probability of the event that a service in process at a given moment will not be interrupted. This happens if during the service no type 1 customers enter, namely,

$$\int_0^\infty \mathrm{e}^{-\lambda_1 x}\mu_2\mathrm{e}^{-\mu_2 x}\,\mathrm{d}x = \frac{\mu_2}{\lambda_1+\mu_2},$$

then service will be interrupted with probability $\lambda_1/(\lambda_1+\mu_2)$.

11.5.3 Priority System with General Service Time

Now we come to priority systems with generally distributed service time. We will consider three cases:

1. If a type 1 customer enters, then the service of a type 2 customer is interrupted and is continued after all type 1 customers have been served. The performed work is taken into account, and the service time is decreased with the work done.
2. The service is realized as above, but when a type 2 customer is served, the performed work will not be taken into account; the service decreases with time spent.
3. When a type 1 customer enters, the actual service is interrupted, and the customer is lost.

In all three cases we assume the entering customers constitute Poisson processes with parameters λ_1 and λ_2, the service times are arbitrarily distributed random variables with distribution functions $B_1(x)$ and $B_2(x)$, respectively. The Laplace–Stieltjes transforms of the service time is

$$b_i^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx}\,\mathrm{d}B_i(x), \qquad i = 1, 2. \tag{11.16}$$

Let us denote the mean values of service times by τ_1 and τ_2, let $V_i(t)$ be the waiting time of a type i customer on the condition that it entered at moment t, and let $\hat{V}_i(t)$ be the time till completion of service. Let

$$F_i(x) = \lim_{t\to\infty} \mathbf{P}\left(V_i(t) < x\right), \qquad i = 1, 2, \ldots,$$

$$\hat{F}_i(x) = \lim_{t\to\infty} \mathbf{P}(\hat{V}_i(t) < x), \qquad i = 1, 2, \ldots,$$

be the distribution of the waiting time and the time till service completion and let their Laplace–Stieltjes transforms according to Eq. (11.16) be $f_i^{\sim}(s)$ and $\hat{f}_i^{\sim}(s)$.

Type 1 customers are served independently of type 2 customers, so by Eq. (8.19) (if the condition $\lambda_1\tau_1 < 1$ is fulfilled)

$$f_1^{\sim}(s) = \frac{1 - \lambda_1\tau_1}{1 - \lambda_1 \frac{1 - b_1^{\sim}(s)}{s}}.$$

The time interval till completion consists of two parts: the waiting time and the service time. They are independent random variables, so for $\hat{V}_1(t)$ we obtain

$$\hat{f}_1^{\sim}(s) = \frac{(1 - \lambda_1\tau_1)b_1^{\sim}(s)}{1 - \frac{\lambda_1}{s}(1 - b_1^{\sim}(s))}.$$

At the service of type 2 customers the service of type 1 customers may be interpreted as the breakdown of a server. Let L be a random variable denoting the time from the beginning of service of a type 2 customer till the beginning of service of the next one and

$$b_L^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx}\, \mathrm{d}\mathbf{P}(L < x).$$

At a fixed moment we have two possibilities: a type 1 customer is absent in the system with probability $1 - \lambda_1\tau_1$ and present with probability $\lambda_1\tau_1$, and in its presence according to the service discipline it is being served. If there are no type 1 customers, then by Eq. (8.19) the Laplace–Stieltjes transform of the remaining service time is

$$\frac{1 - \lambda_2 \mathbf{E}(L)}{1 - \frac{\lambda_2}{s}(1 - b_L^{\sim}(s))}.$$

In the presence of a type 1 customer, we must first finish the serving existing and entering type 1 customers, then serve type 2 customers (taking into account type 1 customers that enter in the meantime). The Laplace–Stieltjes transform of the service time for existing and entering type 1 customers is

$$\frac{1 - b_0^{\sim}(s)}{s\frac{\tau_1}{1-\lambda_1\tau_1}},$$

where $b_0^{\sim}(s)$ is the solution of the functional equation

$$b_0^{\sim}(s) = b_1^{\sim}(s + \lambda_1 - \lambda_1 b_0^{\sim}(s)),$$

i.e., the Laplace–Stieltjes transform of a busy period for type 1 customers. After having served the type 1 customers we come to the previous situation. Thus, the Laplace–Stieltjes transform of the time period till the service of type 2 customers entering at a given moment is

$$\begin{aligned}(1 - \lambda_1\tau_1)\frac{1 - \lambda_2\mathbf{E}(L)}{1 - \frac{\lambda_2}{s}(1 - b_L^{\sim}(s))} + \lambda_1\tau_1 \frac{1 - b_0^{\sim}(s)}{s\frac{\tau_1}{1-\lambda_1\tau_1}} \cdot \frac{1 - \lambda_2\mathbf{E}(L)}{1 - \frac{\lambda_2}{s}(1 - b_L^{\sim}(s))} \\ = \frac{(1 - \lambda_1\tau_1)(1 - \lambda_2\mathbf{E}(L))[s + \lambda_1(1 - b_0^{\sim}(s))]}{s - \lambda_2(1 - b_L^{\sim}(s))} = f_2^{\sim}(s). \qquad (11.17)\end{aligned}$$

In this expression $b_L^{\sim}(s)$ is still unknown, but we will find it for our three models.

1. From the point of view of a type 2 customer we can interpret the system behavior such that the entry of a type 1 customer is a failure and the end of the busy period generated by this type 1 customer as maintenance. Based on this interpretation our model can be considered a system with server breakdowns. Thus,

$$b_L^{\sim}(s) = b_2^{\sim}(s + \lambda_1 - \lambda_1 b_0^{\sim}(s)).$$

2. Let us consider the sequences of independent random variables $\{U_n\}$, $\{H_n\}$ and $\{A_n\}$, which have the following meaning:

   U_i: service time of a type 2 customer [with Laplace–Stieltjes transform $b_2^{\sim}(s)$];
   H_i: length of busy period for type 1 customers [the corresponding Laplace–Stieltjes transform is $b_0^{\sim}(s)$].
   A_i: interarrival time for type 1 customers (exponentially distributed random variable with parameter λ_1).

   If $U_1 \le A_1$, then $L = U_1$ (during the service of type 2 customers no type 1 customers enter, so the type 2 customer leaves after U_1 time from the beginning of service).

   If $A_1 < U_1$, $U_2 \le A_2$, then $L = H_1 + A_1 + U_2$ (during the service of a type 2 customer after A_1 time a type 1 customer enters, and for its and the entering customers' service we try time H_1, then the service of a type 2 customer is realized for U_2 without interruption). Similarly, if $A_1 < U_1$, $A_2 < U_2, \ldots, A_n < U_n$, $U_{n+1} \le A_{n+1}$, then $L = A_1 + H_1 + A_2 + H_2 + \ldots + A_n + H_n + U_{n+1}$.

   By the formula of total probability,

$$\begin{aligned}\mathbf{P}(L < x) = \sum_{n=0}^{\infty} \mathbf{P}(&A_i < U_i, 1 \le i \le n; U_{n+1} \\ &\le A_{n+1}; A_1 + H_1 + A_2 + H_2 + \ldots + A_n + H_n + U_{n+1} < x).\end{aligned}$$

   Since

$$\begin{aligned}\int_0^{\infty} \mathrm{e}^{-sx}\, \mathrm{d}_x P\{A_i < x,\ A_i < U_i\} &= \lambda_1 \int_0^{\infty} \mathrm{e}^{-sx}(1 - B_2(x))\mathrm{e}^{-\lambda_1 x}\, \mathrm{d}x \\ &= \frac{\lambda_1}{s + \lambda_1}[1 - b_2^{\sim}(s + \lambda_1)]\end{aligned}$$

   and

$$\begin{aligned}\int_0^{\infty} \mathrm{e}^{-sx}\, \mathrm{d}_x P\{U_i < x,\ U_i \le A_i\} &= \int_0^{\infty} \mathrm{e}^{-(s+\lambda_1)x}\, \mathrm{d}B_2(x) \\ &= b_2^{\sim}(s + \lambda_1),\end{aligned}$$

   we obtain

$$\begin{aligned}b_L^{\sim}(s) &= \sum_{n=0}^{\infty} \left\{ \frac{\lambda_1}{s + \lambda_1}[1 - b_2^{\sim}(s + \lambda_1)]b_0^{\sim}(s) \right\}^n b_2^{\sim}(s + \lambda_1) \\ &= \frac{(s + \lambda_1)b_2^{\sim}(s + \lambda_1)}{s + \lambda_1 - \lambda_1[1 - b_2^{\sim}(s + \lambda_1)]b_0^{\sim}(s)}.\end{aligned}$$

3. Using the random variables U_i, H_i, A_i we have

$$L = \begin{cases} U_1, & \text{ha } U_1 \le A_1, \\ A_1 + H_1, & \text{ha } U_1 > A_1. \end{cases}$$

Consequently,

$$b_L^{\sim}(s) = b_2^{\sim}(s+\lambda_1) + \frac{\lambda_1}{s+\lambda_1}[1 - b_2^{\sim}(s+\lambda_1)]b_0^{\sim}(s).$$

Now let us find the functions $\hat{f}_2^{\sim}(s)$. In the first two cases the time from the moment t till the end of service is $U_2(t) + L$. They are independent random variables, so in both cases

$$\hat{f}_2^{\sim}(s) = f_2^{\sim}(s) b_L^{\sim}(s).$$

In the third case we can lose the type 2 customer; this happens if during the service of a type 2 customer a type 1 customer appears, and the probability of loosing the type 2 customer is

$$\mathbf{P}(A_1 < U_1) = 1 - b_2^{\sim}(\lambda_1).$$

Obviously,

$$\hat{U}_2(t) = U_2(t) + \min(A_2, U_1).$$

Since

$$\int_{x=0}^{\infty} \mathrm{e}^{-sx}\, \mathrm{d}\mathbf{P}(\min(A_1, U_1) = x) = b_2^{\sim}(s+\lambda_1) + \frac{\lambda_1}{s+\lambda_1}[1 - b_2^{\sim}(s+\lambda_1)],$$

then

$$\hat{f}_2^{\sim}(s) = f_2^{\sim}(s) \left\{ b_2^{\sim}(s+\lambda_1) + \frac{\lambda_1}{s+\lambda_1}[1 - b_2^{\sim}(s+\lambda_1)] \right\}.$$

These formulas are true if the process has an equilibrium distribution. On the basis of Eq. (11.17), this means that the inequalities $\lambda_1\tau_1 < 1$ and $\lambda_2\mathbf{E}(L) < 1$ hold.

- In the first model $\mathbf{E}(L) = \tau_2/(1-\lambda_1\tau_1)$, from which the condition of equilibrium is $\lambda_2\tau_2 < 1 - \lambda_1\tau_1$.
- In the second model $\mathbf{E}(L) = [1 - b_2^{\sim}(\lambda_1)]/\lambda_1(1-\lambda_1\tau_1)b_2^{\sim}(\lambda_1)$, from which the condition of equilibrium is $\lambda_2[1 - b_2^{\sim}(\lambda_1)] < \lambda_1(1-\lambda_1\tau_1)b_2^{\sim}(\lambda_1)$.
- In the third model $\mathbf{E}(L) = [1 - b_2^{\sim}(\lambda_1)]/\lambda_1(1-\lambda_1\tau_1)$, from which the condition of equilibrium is $\lambda_2[1 - b_2^{\sim}(\lambda_1)] < \lambda_1(1-\lambda_1\tau_1)$.

11.6 Systems with Several Servers and Queues

11.6.1 Multichannel Systems with Waiting and Refusals

Let $(X_n, Y_n),\ n = 1, 2, \dots$ be a sequence of i.i.d. random vector variables, where $X_1, X_2, \dots$ are the interarrival periods of successive customers (the nth one enters at the moment $t_n = X_1 + \cdots + X_n,\ n = 1, 2, \dots$), and Y_n is the service time of nth customer.

Let us consider a $G/G/m$ system. We introduce the *waiting time vector* of the nth customer:

$$W_n = (W_{n,1}, \dots, W_{n,m}),\ n = 1, 2, \dots,$$

where $W_{n,i}$ means the random time interval the nth customer (entering at t_n) has to wait till i servers become free from all earlier (with numbers $1, \dots, n-1$) customers.

If the initial random vector variable W_0 (on the same probability space) is given, then the sequence $W_n,\ n \geq 0$, is uniquely determined and a recurrence relation is valid for W_n, i.e., $\{W_n,\ n \geq 0\}$ is a recurrent process. For the arbitrary $x = (x_1, \dots, x_m) \in \mathbb{R}^m$ let

$$x^+ = (x_1^+, \dots, x_m^+), \quad \text{where} \quad s^+ = \max(s, 0),\ s \in \mathbb{R},$$

$$R(x) = (x_{i_1}, \dots, x_{i_m}), \quad x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_m},$$

i.e., the function $R(x)$ arranges the components of vector x in increasing order. We introduce the vectors

$$\overset{(1)}{e} = (\ 1,\ \overset{(2)}{0}, \dots,\ \overset{(m)}{0}\ ), \qquad \overset{(1)}{i} = (\ 1, \dots,\ \overset{(m)}{1}\ ).$$

Theorem 11.4. *For the sequence $W_n,\ n \geq 0$, the recurrence relation*

$$W_{n+1} = R\left([(W_n + Y_n e) - X_n i]^+\right), \quad n \geq 0. \tag{11.18}$$

holds.

Proof. Using the definition of included quantities, this is trivial. □

In the investigation of queueing systems the existence of a limit distribution for the basic characteristics is an important question. Using results from the theory of recurrence processes one can prove a theorem valid in the more general case where instead of total independence we assume stationarity in a narrower sense and the ergodicity of the process $\{(X_n, Y_n),\ n \geq 1\}$ (see [14]).

Theorem 11.5. *Let* $\{(X_n, Y_n),\ n \geq 1\}$ *be a sequence of i.i.d. random variables and* $\mathbf{E}(Y_1 - mX_1) < 0,\ W_0 = 0$*; then there exists a stationary, in a narrow sense, process* $\{W^{(n)},\ n \geq 0\}$ *satisfying Eq.* (11.19)*, and the distribution function of* W_n *monotonically converges to the distribution function of* $W^{(0)}$.

***G/G/m/0* Systems with Refusals** Since we are considering a system with refusals, one can speak of waiting only in a virtual sense. Thus, the component $W_{n,i}$ of W_n means the possible waiting time of customers entering at moment t_n till i servers become free from all earlier (with numbers $1, \ldots, n-1$) customers (if $W_{n,1} > 0$, then the nth one will not be serviced). We can write the recurrence relation

$$W_{n+1} = R\left(\left[(W_n + Y_n e\mathcal{I}_{\{W_{n,1}=0\}}) - X_n i\right]^+\right).$$

The sufficient condition similar to the previous theorem is

$$\mathbf{P}(Y_1 \leq mX_1) > 0,\quad \mathbf{E}(Y_1) < \infty.$$

If $(X_n, Y_n),\ -\infty < n < \infty$ is not a sequence of independent random variables with the same distribution but a stationary (stationary in a narrower sense), ergodic sequence, even in this case we can give a sufficient condition for the existence of a limit distribution, namely,

$$\mathbf{P}(Y_0 \leq X_0 + \ldots + X_{m-1}, Y_{-1} \leq X_{-1} + X_0 + \ldots + X_{m-2}, \ldots, Y_{-m+1} \leq X_{-m+1} + X_{-m+2} + \ldots + X_0) > 0,$$

$$\mathbf{E}(Y_1) < \infty.$$

If we consider instead of the virtual waiting time the queue length L_n at the arrival moment of the nth customer, then it also has a nondegenerate limiting distribution.

***G/G/∞* system** Now we have an infinite number of servers, so one cannot speak of queueing or losses. In this case the basic characteristic is the queue length: $L_k,\ k \geq 1$, denotes the number of customers at the arrival moment of the kth customer [at an arbitrary moment t the number of customers present $L(t)$ is left continuous]. Actually, it is the number of occupied servers. At the beginning the system is empty, i.e., $L_1 = 0$.

For the sake of simplicity let $X_n,\ n \geq 1$, denote the interarrival time of the nth and $(n+1)$st customers, $Y_n,\ n \geq 1$, the service time of the nth customer. Then

$$L_{k+1} = \mathcal{I}_{\{Y_k > X_k\}} + \mathcal{I}_{\{Y_{k-1} > X_{k-1} + X_k\}} + \cdots + \mathcal{I}_{\{Y_1 > X_1 + \cdots + X_k\}},\quad k \geq 1.$$

Theorem 11.6. *If* $\{(X_n, Y_n),\ -\infty < n < \infty\}$ *is a sequence of i.i.d. random variables and* $0 < \mathbf{E}(Y_1) < \infty$ *is fulfilled, then*

$$L = \sum_{k \geq 1} \mathcal{I}_{\{Y_{-k} > X_{-k} + \cdots + X_{-1}\}}$$

defines a finite random variable with probability 1, the random variables

$$L_{-n} = \sum_{k=1}^{n} \mathcal{I}_{\{Y_{-k} > X_{-k} + \cdots + X_{-1}\}}, \quad n = 1, 2, \ldots$$

and L_n have the same distribution, and this distribution monotonically converges to the distribution of L.

Proof. For the proof it is enough to show that L is finite with probability 1. We need the following lemma; from it with probability 1 follows the finiteness of L. □

Lemma 11.7. *Let $U_1, U_2, \ldots$ be a sequence of i.i.d. random variables for which $\mathbf{P}(U_1 \geq 0) = 1$ and $h = \mathbf{E}\left(\mathrm{e}^{-U_1}\right) < 1$, i.e., the distribution of U_i is not concentrated at the point 0. Let V be an arbitrary random variable (not necessarily independent of U_i) for which $\mathbf{E}\left(V^+\right) < \infty$. Furthermore, let $\kappa = \frac{1}{2}\log\frac{1}{h}$, $G_V(x) = 1 - \mathbf{P}(V < x)$, $x \in \mathbb{R}$. Then for arbitrary $n, N \geq 1$*

$$\mathbf{P}(U_1 + \cdots + U_n < V) < \mathrm{e}^{-n\kappa} + G_V(n\kappa), \tag{11.19}$$

$$\sum_{n \geq N} \mathbf{P}(U_1 + \cdots + U_n < V) < \frac{1}{1 - \mathrm{e}^{-\kappa}} \mathrm{e}^{-N\kappa} + \frac{1}{\kappa} \mathbf{E}\left(V \mathcal{I}_{\{N\kappa \leq V\}}\right) \tag{11.20}$$

is true.

Proof. For arbitrary $x > 0$

$$\begin{aligned}
&\mathbf{P}(U_1 + \cdots + U_n < V) \\
&\quad = \mathbf{P}(U_1 + \cdots + U_n < V,\ V \leq nx) + \mathbf{P}(U_1 + \cdots + U_n < V,\ nx < V) \\
&\quad \leq \mathbf{P}(U_1 + \cdots + U_n < nx) + \mathbf{P}(nx \leq V).
\end{aligned}$$

Using the Markov inequality we obtain

$$\begin{aligned}
\mathbf{P}(U_1 + \cdots + U_n < nx) &\leq \mathbf{E}(\exp\{nx - (U_1 + \cdots + U_n)\}) \\
&= \mathrm{e}^{nx} \prod_{i=1}^{n} \mathbf{E}\left(\mathrm{e}^{-U_i}\right) \\
&= \mathrm{e}^{n(x + \log h)},
\end{aligned}$$

where at $x = \kappa$ Eq. (11.19) follows.

Proof of Eq. (11.20): From inequality (11.19)

$$\sum_{n\geq N}\mathbf{P}\left(U_1+\cdots+U_n<V\right)\leq\sum_{n\geq N}\{\mathrm{e}^{-n\kappa}+G_V(n\kappa)\}$$

$$=\frac{1}{1-\mathrm{e}^{-\kappa}}\mathrm{e}^{-N\kappa}+\sum_{j=0}^{\infty}\mathbf{P}\left((N+j)\kappa\leq V\right).$$

Since

$$\mathbf{E}\left(V\mathcal{I}_{\{N\kappa\leq V\}}\right)\geq\sum_{j=0}^{\infty}(N+j)\kappa\mathbf{P}\left((N+j)\kappa\ \leq\ V\ <\ (N+j+1)\kappa\right)$$

$$=(N-1)\kappa\sum_{j=0}^{\infty}\mathbf{P}\left((N+j)\kappa\ \leq V\ <\ (N+j+1)\kappa\right)$$

$$+\sum_{j=0}^{\infty}(j+1)\kappa\mathbf{P}\left((N+j)\kappa\ \leq\ V\ <\ (N+j+1)\kappa\right)$$

$$=(N-1)\kappa\mathbf{P}\left(N\kappa\leq V\right)+\kappa\sum_{j=0}^{\infty}\mathbf{P}\left((N+j)\kappa\leq V\right),$$

then

$$\sum_{j=0}^{\infty}\mathbf{P}\left((N+j)\kappa\leq V\right)\leq\frac{1}{\kappa}\mathbf{E}\left(V\mathcal{I}_{\{N\kappa\leq V\}}\right)-(N-1)\mathbf{P}\left(N\kappa\leq V\right)$$

$$\leq\frac{1}{\kappa}\mathbf{E}\left(V\mathcal{I}_{\{N\kappa\leq V\}}\right).$$

Using Eq. (11.19) we obtain Eq. (11.20). □

11.6.2 Closed Queueing Network Model of Computers

The queueing network in Fig. 11.27 may be considered the simplest mathematical model for computers.

In a system there are continuously n customers (tasks) and they can move along the routes indicated in the figure. In front of each service unit there is a waiting buffer of corresponding capacity (for at most $n-1$ customers). On the units the service is realized by the FCFS rule; the service times are independent and on the ith unit have distribution function $F_i(x)$, $0\leq i\leq M$. After having completed a service on the 0th unit, the customer moves to the ith unit with probability p_i, $0\leq i\leq M$ $(p_i\geq 0,\ p_0+\cdots+p_M=1)$, which does not depend on the state of the system or the service time. If the service of a customer is completed on the ith

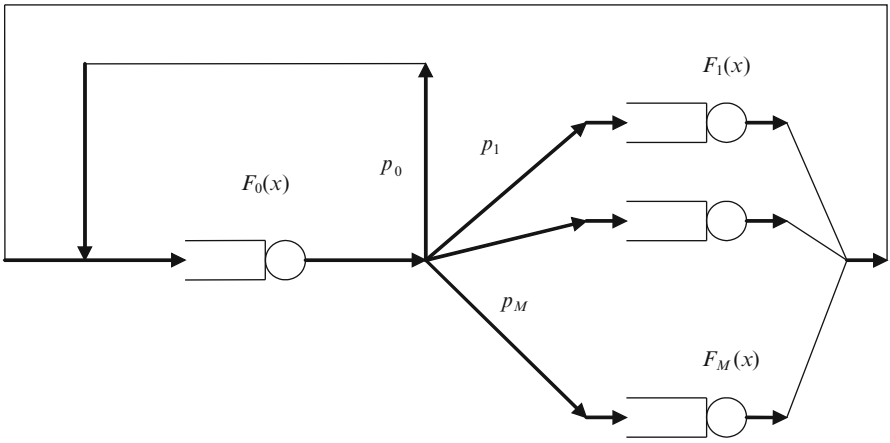


Fig. 11.27 Closed queueing network model

$(1 \leq i \leq M)$ unit, then the customer goes to a unit with index 0 with probability 1. The unit 0 plays a special role in the network and is called a *central unit*.

It may seem too strict a restriction that the number of customers in a computer is fixed, but this model gives accurate results for several important performance parameters. For example, when we are interested in the maximum performance of a computer, we can assume that the load of the computer is maximal, that is, after completion of a task, a new one enters the system immediately.

If we consider the successive moments when all customers stay at the central unit and the service of a customer has just started, these moments are regeneration points for the network. If the service times have finite pth $(p \geq 1)$ order moments, then one can show the pth moment of a regenerative cycle is also finite [86, 87]. It follows that if the mean values of services are finite, then the different characteristics for the system have limiting distributions.

11.7 Retrial Systems

11.7.1 Continuous-Time Retrial System

If in the case of a phone call the subscriber is occupied, one usually repeats the attempt while the conversation is realized. So the system has two types of requests: primary calls and calls generated by occupied lines. Models constructed for systems with losses do not describe this situation, and they do not take into account repetitions. These problems appeared in Erlang's time, but due to a lack of corresponding theoretical results, these repetitions were considered new arrivals.

Retrial queues constitute a special field of queueing systems; their distinguishing feature is that in the case of a busy server, entering customers leave the service area (go to the orbit) and, after a certain (generally random) time, reinitiate their service.

Let us consider some examples of the retrial phenomenon. The first example is connected with the functioning of call centers used by companies to communicate with customers. When a call arrives, it is sent to a call distribution switch. If all agents are busy, then the call center may announce an estimated waiting time. Some customers might decide to wait for a free agent, while some will interrupt the connection immediately or after some time. A portion of these customers will return after some random time.

Random access protocols provide a motivation for the design of communication protocols with retransmission control. If two or more stations transmit packets at the same time, then a collision takes place. Then all packets are destroyed and should be retransmitted. To avoid collisions in the next period, this transmission is realized with a certain random delay for each station. This fact motivates the investigation of the retrial feature of computer networks.

Two textbooks have been published in this field. The book by Falin and Templeton [30] gives analytical solutions in terms of generating functions and Laplace-Stieltjes transforms, and the one by Artalejo and Gómez-Corral [6] focuses on the application of algorithmic methods studying the M/G/1 and M/M/c retrial queues and using matrix-analytic techniques to solve some retrial queues with QBD, GI/M/1, and M/G/1 structures.

We will consider a model connected with the landing process of airplanes in the case of continuous time. The model was introduced in [59], and the results for waiting time are contained in [61].

Let us consider the landing process of airplanes. An airplane appears at an airport ideally positioned for landing. If it is not possible (due to insufficient distance or a waiting queue), it starts circling. The next request for service is possible when it returns to the starting geometrical point on the condition that there are no other airplanes ahead of it.

Similar problems appear at the transmission of optical signals. Signals entering the node must be sent in the order of arrival, but they cannot be stored. They go to delay lines and upon their return can reinitiate their transmission. If all previous signals have already been sent, then the signal is transmitted; in the opposite case they pass through the delay queue again, and the process is repeated.

Let us formulate the queueing problem. We investigate a service system where the service may start at the moment of arrival (if the system is available) or at moments differing from it by multiples of a given cycle time T (in the case of busy server or queue). Service of a customer can be started if all customers who had entered the system earlier have already left (i.e., the FIFO rule works). In such a system the service process is not continuous; during the "busy period" there are idle intervals; these idle intervals are necessary to reach the starting point; for them there is no real service.

Let the service of the nth customer begin at moment t_n, and let us consider the number of customers present at the moment just before service begins. Then the

number of customers present is determined by the recursive formula

$$N_{n+1} = \begin{cases} \Delta_n, & \text{if } N_n = 0, \\ N_n - 1 + \Delta_n, & \text{if } N_n > 0, \end{cases}$$

where Δ_n is the number of customers appearing for $[t_n, t_{n+1})$. We show that these values constitute a Markov chain.

Let us consider the time intervals during which we record the entering customers. Let $\{Z_i\}$ and $\{Y_i\}$ $(i = 1, 2, \ldots)$ be two independent sequences of independent random variables. Z_i means the interarrival time between the ith and $i + 1$th customers (it has an exponential distribution with the parameter λ), Y_i is the service time of the ith customer (in our case it has an exponential distribution with parameter μ).

Let us assume that at the beginning of service there is one customer in the system. If $Z_i \geq Y_i$, then the time till the beginning of service of the next customer is Z_i (the service of the existing customer will be completed, and the server arrives at a free state and the next customer appears later). If $Z_i < Y_i$, then during the service of the first customer a second one appears, and after this moment there will be intervals with length T while we pass the moment of service of the first customer (from the viewpoint of entering customers we are interested in the time from the entry of the second customer till the beginning of its service). The length of this interval is the function of random variables Z_i and Y_i, i.e., a certain $f_1(Z_i, Y_i)$.

If at the beginning of service of the first customer the second one is already present, then the time till the starting moment of its service is determined in the following way. Divide the service time of the first customer into intervals of length T (the last period generally is not full). Since the starting moments for both customers differ by multiples of T from the moments of arrivals, on each interval of length T there is one point where the service of the second customer may start. In reality, this happens at the first moment after the service of the first customer has completed, so the required time period is determined by the service time of the first customer and the interarrival time. Consequently, it will be a certain function of Y_i and Z_i, i.e., $f_2(Y_i, Z_i)$.

Thus, the time intervals for which we consider the number of entering customers are only functions of random variables Y_i and Z_i, consequently they are independent. Taking into account the fact that entering customers form a Poisson process, the quantities Δ_i of these customers are independent random variables, and N_n is a Markov chain.

To describe the functioning of the system we use the embedded Markov chain technique. Our result is formulated in the following theorem.

Theorem 11.8. *Let us consider a service system in which the entering customers form a Poisson process with parameter λ, and the service time is exponentially distributed with parameter μ. The service of a customer may be started at the moment of arrival (in the case of a free system) or at moments differing from it by the multiples of a cycle time T (in the case of a busy server or queue); the service*

discipline is FIFO. Let us define a Markov chain whose states correspond to the number of customers in the system at moments $t_k - 0$ (t_k is the starting moment of service of the kth customer). The matrix of transition probabilities of this Markov chain has the form

$$\begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \dots \\ a_0 & a_1 & a_2 & a_3 & \dots \\ 0 & b_0 & b_1 & b_2 & \dots \\ 0 & 0 & b_0 & b_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{11.21}$$

and its elements are determined by the generating functions

$$A(z) = \sum_{i=0}^{\infty} a_i z^i = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} z \frac{(1-\mathrm{e}^{-\mu T})\mathrm{e}^{-\lambda(1-z)T}}{1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}}, \tag{11.22}$$

$$\begin{aligned} B(z) &= \sum_{i=0}^{\infty} b_i z^i \\ &= \frac{1}{(1-\mathrm{e}^{-\lambda T})(1-\mathrm{e}^{-[\lambda(1-z)+\mu]T})} \\ &\quad \times \left[\frac{1}{2-z} \left(1-\mathrm{e}^{-\lambda(2-z)T}\right) \left(1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}\right) \right. \\ &\quad \left. - \frac{\lambda}{\lambda(2-z)+\mu} \left(1-\mathrm{e}^{-[\lambda(2-z)+\mu]T}\right) \left(1-\mathrm{e}^{-\lambda(1-z)T}\right) \right]. \end{aligned} \tag{11.23}$$

The generating function of the ergodic distribution of this chain is

$$P(z) = p_0 \frac{B(z)(\lambda z+\mu) - zA(z)(\lambda+\mu)}{\mu[B(z)-z]}, \tag{11.24}$$

where

$$p_0 = 1 - \frac{\lambda}{\lambda+\mu} \frac{1-\mathrm{e}^{-(\lambda+\mu)T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}. \tag{11.25}$$

The ergodicity condition is

$$\frac{\lambda}{\mu} < \frac{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}{1-\mathrm{e}^{-\lambda T}}. \tag{11.26}$$

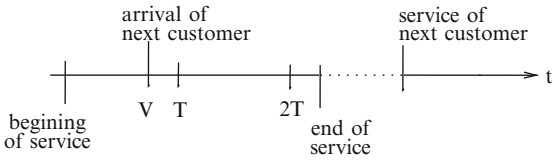


Fig. 11.28 One customer at the beginning of service

Proof. Our original system, where during the busy period there are possible idle intervals, too, is replaced by another one. In it the service process is continuous, and the service time of a customer consists of two parts: the first part is the real service, the second part holds from the end of service till the second one gets to the corresponding position.

For a description of the operation we use an embedded Markov chain; its states are the number of customers in the system at moments $t_k - 0$, i.e., we consider it at moments just before starting service. Let us find the transition probabilities for this chain. We have to distinguish two cases: at the starting moment of service the next customer is present or not. First we will consider the second possibility (Fig. 11.28), which happens for the states 0 and 1. Suppose that the service time of the first customer is U, the second customer enters at V time after the beginning of service. The probability of event $\{U - V < t\}$ is

$$\begin{aligned} P(t) &= \mathbf{P}\,(U - V < t) \\ &= \int_0^t \int_0^U \lambda \mathrm{e}^{-\lambda V} \mu \mathrm{e}^{-\mu U} \mathrm{d}V \mathrm{d}U + \int_t^\infty \int_{U-t}^U \lambda \mathrm{e}^{-\lambda V} \mu \mathrm{e}^{-\mu U} \mathrm{d}V \mathrm{d}U \\ &= \frac{\lambda}{\lambda + \mu} \left(1 - \mathrm{e}^{-\mu t}\right). \end{aligned} \tag{11.27}$$

The time from the entry of the second customer till the beginning of its service equals

$$(I\,(U - V) + 1)\,T,$$

where $I(x)$ denotes an integer part of number x/T. This expression is valid for all points excluding multiples of T, but the probability of an event for this time to equal a multiple of T is equal to zero. To determine the transition probabilities, we need the number of customers entering during this period. According to Eq. (11.27) the time from the entry till the beginning of service is equal to iT with probability

$$\frac{\lambda}{\lambda + \mu} \left(\mathrm{e}^{-\mu(i-1)T} - \mathrm{e}^{-\mu i T}\right),$$

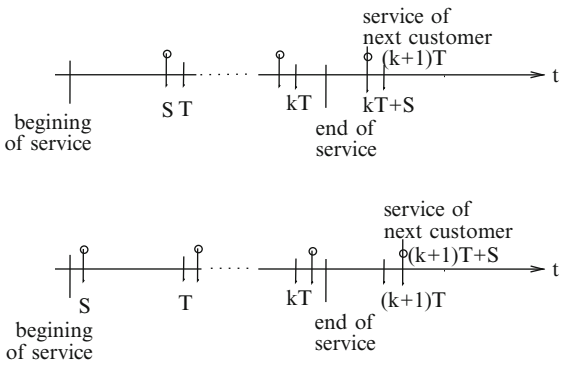


Fig. 11.29 More than one customer at the beginning of service

and the generating function of entering customers equals

$$\frac{\lambda}{\lambda+\mu}\sum_{k=0}^{\infty}\sum_{i=1}^{\infty}\left(\mathrm{e}^{-\mu(i-1)T}-\mathrm{e}^{-\mu iT}\right)\frac{(\lambda iTz)^k}{k!}\mathrm{e}^{-\lambda iT}$$

$$=\frac{\lambda}{\lambda+\mu}\sum_{i=1}^{\infty}\left(\mathrm{e}^{-\mu(i-1)T}-\mathrm{e}^{-\mu iT}\right)\mathrm{e}^{-\lambda iT(1-z)}=\frac{\lambda}{\lambda+\mu}\frac{\mathrm{e}^{-\lambda(1-z)T}(1-\mathrm{e}^{-\mu T})}{1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}}.$$

This formula is valid if for U at least one customer enters the system, so the desired generating function is

$$A(z)=\frac{\mu}{\lambda+\mu}+\frac{\lambda}{\lambda+\mu}z\frac{(1-\mathrm{e}^{-\mu T})\mathrm{e}^{-\lambda T(1-z)}}{1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}},$$

where $\frac{\mu}{\lambda+\mu}=\int_0^{\infty}\mathrm{e}^{-\lambda x}\mu\mathrm{e}^{-\mu x}\mathrm{d}x$ is the probability that for the service time no customer appears.

Now we find the transition probabilities for all other states. In this case at the beginning of service the next customer is already present (Fig. 11.29). Let $R=U-I(U)T$ and let S be the *mod* T interarrival time. One can easily see that S has a truncated exponential distribution with distribution function $\frac{1-\mathrm{e}^{-\lambda S}}{1-\mathrm{e}^{-\lambda T}}$. The time between the starting moments of two successive customers is

$$I(U)T+S \quad \text{if} \quad R\le S \qquad \text{and} \qquad (I(U)+1)T+S \quad \text{if} \quad R>S.$$

k customers enter in the two cases with probabilities

$$\frac{(\lambda\{I(U)T+S\})^k}{k}!\exp\left(-\lambda\{I(U)T+S\}\right) \tag{11.28}$$

and

$$\frac{(\lambda\{[I(U)+1]T+S\})^k}{k!}\exp(-\lambda\{[I(U)+1]T+S\}). \tag{11.29}$$

Let us fix S and divide the service time of the customer into intervals of length T. S divides each such interval into two parts (the first has length S, the second $T-S$), and the corresponding probability for the first part is Eq. (11.28), for the second part Eq. (11.29). Let $I(U)=i$. The generating function of the number of entering customers, denoted by N, assuming that the interarrival time *mod* T is equal to S is as follows

$$\begin{aligned}\mathbf{E}\left(z^N|S\right) &= \sum_{k=0}^{\infty}\sum_{i=0}^{\infty}\left(\int_{iT}^{iT+S}\frac{[\lambda(iT+S)z]^k}{k}!\mathrm{e}^{-\lambda(iT+S)}\mu\mathrm{e}^{-\mu U}\mathrm{d}U\right.\\ &\quad\left.+\int_{iT+S}^{(i+1)T}\frac{[\lambda((i+1)T+S)z]^k}{k!}\mathrm{e}^{-\lambda((i+1)T+S)}\mu\mathrm{e}^{-\mu U}\mathrm{d}U\right)\\ &= \frac{1}{1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}}\left(\mathrm{e}^{-\lambda(1-z)S}-\mathrm{e}^{-[\lambda(1-z)+\mu]S}\right.\\ &\quad\left.+\mathrm{e}^{-\lambda(1-z)T}\mathrm{e}^{-[\lambda(1-z)+\mu]S}-\mathrm{e}^{-\lambda(1-z)S}\mathrm{e}^{-[\lambda(1-z)+\mu]T}\right),\end{aligned}$$

Multiplying this expression by $\frac{\lambda\mathrm{e}^{-\lambda S}}{1-\mathrm{e}^{-\lambda T}}$ and integrating by S from 0 to T we obtain the generating function of transition probabilities

$$\begin{aligned}B(z) &= \sum_{i=0}^{\infty}b_i z^i\\ &= \frac{1}{(1-\mathrm{e}^{-\lambda T})(1-\mathrm{e}^{-[\lambda(1-z)+\mu]T})}\\ &\quad\times\left[\frac{1}{2-z}\left(1-\mathrm{e}^{-\lambda(2-z)T}\right)\left(1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}\right)\right.\\ &\quad\left.-\frac{\lambda}{\lambda(2-z)+\mu}\left(1-\mathrm{e}^{-[\lambda(2-z)+\mu]T}\right)\left(1-\mathrm{e}^{-\lambda(1-z)T}\right)\right].\end{aligned}$$

Consider a Markov chain describing the functioning of the system; the matrix of transition probabilities has the form Eq. (11.21). Let us denote the ergodic distribution by p_i $(i=0,1,2,\ldots)$ and introduce the generating function $P(z)=\sum_{i=0}^{\infty}p_i z^i$. Then

$$p_j = p_0 a_j + p_1 a_j + \sum_{i=2}^{j+1} p_i b_{j-i+1},$$

whence

$$\begin{aligned}\sum_{j=0}^{\infty} p_j z^j &= p_0 A(z) + p_1 A(z) + \sum_{j=0}^{\infty} \sum_{i=2}^{j+1} p_i b_{j-i+1} z^j \\ &= \frac{1}{z} P(z) B(z) - \frac{1}{z} p_0 B(z) + p_0 A(z) + p_1 A(z) - p_1 B(z),\end{aligned}$$

i.e.,

$$P(z) = \frac{p_0 [z A(z) - B(z)] + p_1 z [A(z) - B(z)]}{z - B(z)}.$$

This expression contains two unknown probabilities – p_0 and p_1 – but

$$p_0 = p_0 a_0 + p_1 a_0,$$

i.e.,

$$p_1 = \frac{1 - a_0}{a_0} p_0 = \frac{\lambda}{\mu} p_0.$$

p_0 can be found from the condition $P(1) = 1$,

$$p_0 = \frac{1 - B'(1)}{1 + A'(1) - B'(1) + \frac{\lambda}{\mu}[A' - B'(1)]}.$$

The embedded chain is irreducible, so $p_0 > 0$. Using

$$\begin{aligned}A'(1) &= \frac{\lambda}{\lambda + \mu} \left(1 + \frac{\lambda T}{1 - \mathrm{e}^{-\mu T}}\right), \\ B'(1) &= 1 - \frac{\lambda T \mathrm{e}^{-\lambda T}}{1 - \mathrm{e}^{-\lambda T}} + \frac{\lambda}{\lambda + \mu} \lambda T \frac{1 - \mathrm{e}^{-(\lambda+\mu)T}}{(1 - \mathrm{e}^{-\lambda T})(1 - \mathrm{e}^{-\mu T})},\end{aligned}$$

we obtain

$$\left(1 + \frac{\lambda}{\mu}\right) A'(1) - \frac{\lambda}{\mu} B'(1) = \frac{\lambda}{\lambda + \mu} \lambda T \frac{1 - \mathrm{e}^{-(\lambda+\mu)T}}{(1 - \mathrm{e}^{-\lambda T})(1 - \mathrm{e}^{-\mu T})} > 0,$$

so the condition $1 - B'(1) > 0$ must be fulfilled. This leads to the inequality

$$\frac{\lambda T \mathrm{e}^{-\lambda T}}{1 - \mathrm{e}^{-\lambda T}} - \frac{\lambda}{\lambda + \mu} \lambda T \frac{1 - \mathrm{e}^{-(\lambda+\mu)T}}{(1 - \mathrm{e}^{-\lambda T})(1 - \mathrm{e}^{-\mu T})} > 0,$$

i.e.,

$$\frac{\lambda}{\lambda+\mu} < \frac{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}{1-\mathrm{e}^{-(\lambda+\mu)T}}.$$

This is equivalent to Eq. (11.26). Substituting the corresponding values we obtain

$$p_0 = 1 - \frac{\lambda}{\lambda+\mu}\frac{1-\mathrm{e}^{-(\lambda+\mu)T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}.$$

The theorem is proved. □

During the busy period there are idle intervals, which are necessary to get to the starting position, and they alternate between 0 and T. It is clear that if T decreases, then their influence will become increasingly attenuated. In the limit case, the service process becomes continuous, and after having served a customer, we immediately change to the next one.

Theorem 11.9. *The limiting distribution for the system described above as $T \to 0$ is*

$$P^*(z) = \frac{1-\rho}{1-\rho z} \qquad \left(\rho = \frac{\lambda}{\mu}\right),$$

i.e., it is geometrical with parameter ρ.

Proof. We find p_0, $A(z)$ and $B(z)$ as $T \to 0$, and the limiting values are denoted by p_0^*, $A^*(z)$, and $B^*(z)$. On the basis of Eqs. (11.25), (11.22), and (11.23),

$$p_0^* = \lim_{T\to 0} p_0 = \lim_{T\to 0}\left(1 - \frac{\lambda}{\lambda+\mu}\frac{1-\mathrm{e}^{-(\lambda+\mu)T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}\right) = 1 - \frac{\lambda}{\mu} = 1-\rho,$$

$$\begin{aligned} A^*(z) = \lim_{T\to 0} A(z) &= \lim_{T\to 0}\left(\frac{\mu}{\lambda+\mu} + \frac{\lambda z}{\lambda+\mu}\frac{\mathrm{e}^{-\lambda(1-z)T}(1-\mathrm{e}^{-\mu T})}{1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}}\right) \\ &= \frac{\mu}{\lambda(1-z)+\mu}, \end{aligned}$$

$$\begin{aligned} B^*(z) = \lim_{T\to 0} B(z) &= \frac{1}{(1-\mathrm{e}^{-\lambda T})[1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}]} \\ &\quad \times \left\{ \frac{1}{2-z}\left(1-\mathrm{e}^{-\lambda(2-z)T}\right)\left(1-\mathrm{e}^{-[\lambda(1-z)+\mu]T}\right)\right. \\ &\quad \left. -\frac{\lambda}{\lambda(2-z)+\mu}\left(1-\mathrm{e}^{-[\lambda(2-z)+\mu]T}\right)\left(1-\mathrm{e}^{-\lambda(1-z)T}\right)\right\} \\ &= \frac{\mu}{\lambda(1-z)+\mu}. \end{aligned}$$

Using these values

$$P^*(z) = (1-\rho)\frac{(\lambda z+\mu)\frac{\mu}{\lambda(1-z)+\mu} - z(\lambda+\mu)\frac{\mu}{\lambda(1-z)+\mu}}{\mu\left(\frac{\mu}{\lambda(1-z)+\mu} - z\right)} = \frac{1-\rho}{1-\rho z}.$$

The preceding formula is the generating function for an M/M/1 system and coincides with the previous results. □

11.7.2 Waiting Time for Continuous Retrial System

Let us consider the previously described system. Using Koba's results [57] we determine the distribution of the waiting time. Let t_n denote the moment of arrival of the nth customer; then its service may be started at the moment $t_n + T \cdot X_n$, where X_n is a nonnegative integer. Let $Z_n = t_{n+1} - t_n$, and let Y_n be the service time of the nth customer. If $X_n = i$, then between X_n and X_{n+1} the following relation holds. If

$$(k-1)T < iT + Y_n - Z_n \le kT \qquad (k \ge 1),$$

then $X_{n+1} = k$. In this case X_n is a homogeneous Markov chain with transition probabilities p_{ik}, where

$$p_{ik} = \mathbf{P}\left((k-i-1)T < Y_n - Z_n \le (k-i)T\right)$$

if $k \ge 1$, and

$$p_{i0} = \mathbf{P}\left(Y_n - Z_n \le -iT\right).$$

Introduce the notations

$$f_j = \mathbf{P}\left((j-1)T < Y_n - Z_n \le jT\right), \tag{11.30}$$

$$p_{ik} = f_{k-i} \quad \text{ha} \quad k \ge 1, \quad p_{i0} = \sum_{j=-\infty}^{-i} f_j = \hat{f}_i. \tag{11.31}$$

The ergodic distribution of the Markov chain satisfies the system of equations

$$p_j = \sum_{i=0}^{\infty} p_i\, p_{ij} \quad (j \ge 0), \quad \sum_{j=0}^{\infty} p_j = 1.$$

Theorem 11.10. *Let us consider the system described in Theorem 11.8. Define a Markov chain whose states correspond to the waiting times of customers at moments of arrivals. The matrix of transition probabilities has the form*

$$\begin{bmatrix} \sum_{j=-\infty}^{0} f_j & f_1 & f_2 & f_3 & f_4 & \dots \\ \sum_{j=-\infty}^{-1} f_j & f_0 & f_1 & f_2 & f_3 & \dots \\ \sum_{j=-\infty}^{-2} f_j & f_{-1} & f_0 & f_1 & f_2 & \dots \\ \sum_{j=-\infty}^{-3} f_j & f_{-2} & f_{-1} & f_0 & f_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{11.32}$$

and its elements are determined by formulas (11.30)–(11.31). Then the generating function of the waiting time is

$$P(z) = \left[1 - \frac{\lambda}{\mu}\frac{1-\mathrm{e}^{-\lambda T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}\right] \times \frac{\dfrac{\mu}{\lambda+\mu} - \dfrac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu}\dfrac{z}{z-\mathrm{e}^{-\lambda T}}}{1 - \dfrac{\lambda(1-\mathrm{e}^{-\mu T})}{\lambda+\mu}\dfrac{z}{1-z\mathrm{e}^{-\mu T}} - \dfrac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu}\dfrac{z}{z-\mathrm{e}^{-\lambda T}}}, \tag{11.33}$$

and the stability condition is

$$\frac{\lambda}{\mu} < \frac{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}{1-\mathrm{e}^{-\lambda T}}. \tag{11.34}$$

Proof.

$$\mathbf{P}(Z < x) = 1 - \mathrm{e}^{-\lambda x}, \qquad \mathbf{P}(Y < x) = 1 - \mathrm{e}^{-\mu x}.$$

The distribution function of $Y - Z$ is

$$F(x) = \begin{cases} \frac{\mu}{\lambda+\mu}\mathrm{e}^{\lambda x} & \text{if } x \le 0, \\ 1 - \frac{\lambda}{\lambda+\mu}\mathrm{e}^{-\mu x} & \text{if } x > 0. \end{cases}$$

We find the transition probabilities. In the case $j > 0$,

$$f_j = 1 - \frac{\lambda}{\lambda+\mu}\mathrm{e}^{-\mu(j-1)T} - 1 + \frac{\lambda}{\lambda+\mu}\mathrm{e}^{-\mu jT} = \frac{\lambda}{\lambda+\mu}(1-\mathrm{e}^{-\mu T})\mathrm{e}^{-\mu(j-1)T},$$

for the negative values ($j \ge 0$)

$$f_{-j} = \frac{\mu}{\lambda+\mu}\mathrm{e}^{-\lambda jT} - \frac{\mu}{\lambda+\mu}\mathrm{e}^{-\lambda(j+1)T} = \frac{\mu}{\lambda+\mu}(1-\mathrm{e}^{-\lambda T})\mathrm{e}^{-\lambda jT},$$

$$p_{i0} = \hat{f}_i = \sum_{j=-\infty}^{-i} f_j = \sum_{j=i}^{\infty} \frac{\mu}{\lambda+\mu}(1-\mathrm{e}^{-\lambda T})\mathrm{e}^{-\lambda j T} = \frac{\mu}{\lambda+\mu}\mathrm{e}^{-\lambda i T}.$$

Using the matrix of transition probabilities (11.32) we get the system of equations

$$\begin{aligned}
p_0 &= p_0\hat{f}_0 + p_1\hat{f}_1 + p_2\hat{f}_2 + p_3\hat{f}_3 + \dots \\
p_1 &= p_0 f_1 + p_1 f_0 + p_2 f_{-1} + p_3 f_{-2} + \dots \\
p_2 &= p_0 f_2 + p_1 f_1 + p_2 f_0 + p_3 f_{-1} + \dots \\
&\vdots
\end{aligned}$$

Multiplying the jth equation by z^j, summing up by j from 0 to infinity for the generating function $P(z) = \sum_{j=0}^{\infty} p_j z^j$ we obtain

$$P(z) = P(z)F_+(z) + \sum_{j=1}^{\infty} p_j z^j \sum_{i=0}^{j-1} f_{-i} z^{-i} + \sum_{j=0}^{\infty} p_j \hat{f}_j,$$

where

$$\begin{aligned}
F_+(z) &= \sum_{i=1}^{\infty} f_i z^i = \frac{\lambda z}{\lambda+\mu}(1-\mathrm{e}^{-\mu T}) \sum_{i=1}^{\infty} \mathrm{e}^{-\mu(i-1)T} z^{i-1} \\
&= \frac{\lambda(1-\mathrm{e}^{-\mu T})}{\lambda+\mu} \frac{z}{1-z\mathrm{e}^{-\mu T}}, \\
\sum_{i=0}^{j-1} f_{-i} z^{-i} &= \frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \sum_{i=0}^{j-1} \mathrm{e}^{-\lambda i T} z^{-i} = \frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \frac{1-\left(\frac{\mathrm{e}^{-\lambda T}}{z}\right)^j}{1-\frac{\mathrm{e}^{-\lambda T}}{z}}, \\
\sum_{i=0}^{\infty} p_i \hat{f}_i &= \sum_{i=0}^{\infty} p_i \frac{\mu}{\lambda+\mu}\mathrm{e}^{-\lambda i T} = \frac{\mu}{\lambda+\mu} P\left(\mathrm{e}^{-\lambda T}\right).
\end{aligned}$$

Using the preceding equations

$$\begin{aligned}
P(z) &= P(z)F_+(z) + \sum_{j=1}^{\infty} p_j z^j \frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \frac{1-\left(\frac{\mathrm{e}^{-\lambda T}}{z}\right)^j}{1-\frac{\mathrm{e}^{-\lambda T}}{z}} + \frac{\mu}{\lambda+\mu} P\left(\mathrm{e}^{-\lambda T}\right) \\
&= P(z)F_+(z) + \frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu} \frac{z}{z-\mathrm{e}^{-\lambda T}} \left[P(z) - P\left(\mathrm{e}^{-\lambda T}\right)\right]
\end{aligned}$$

$$+\frac{\mu}{\lambda+\mu}P\left(\mathrm{e}^{-\lambda T}\right)$$

or

$$P(z)\left[1-F_+(z)-\frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu}\frac{z}{z-\mathrm{e}^{-\lambda T}}\right]$$
$$=P\left(\mathrm{e}^{-\lambda T}\right)\left[\frac{\mu}{\lambda+\mu}-\frac{\mu(1-\mathrm{e}^{-\lambda T})}{\lambda+\mu}\frac{z}{z-\mathrm{e}^{-\lambda T}}\right].$$

$P(\mathrm{e}^{-\lambda T})$ may be computed from the condition $P(1)=1$,

$$P\left(\mathrm{e}^{-\lambda T}\right)=1-\frac{\lambda}{\mu}\frac{1-\mathrm{e}^{-\lambda T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}.$$

So for the generating function we get Eq. (11.33). From it we get the probability of the event that the waiting time is equal to zero:

$$p_0=\left[1-\frac{\lambda}{\mu}\frac{1-\mathrm{e}^{-\lambda T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}\right]\frac{\mu}{\lambda+\mu}.$$

In order to have $p_0>0$, the inequality

$$\frac{\lambda}{\mu}\frac{1-\mathrm{e}^{-\lambda T}}{\mathrm{e}^{-\lambda T}(1-\mathrm{e}^{-\mu T})}<1$$

must be fulfilled. It leads to condition (11.34) and coincides with the stability condition for the number of customers. □

11.8 Exercises

Exercise 11.1. A transmission link with capacity $C=5\,\mathrm{MB/s}$ serves two kinds of CBR connections. Type i connections arrive according to a Poisson process at a rate λ_i and occupy c_i bandwidth of the link for an exponentially distributed amount of time with the parameter μ_i $(i=1,2)$, where $c_1=1\,\mathrm{MB}$ and $c_2=2\,\mathrm{MB}$.

1. Describe the system behavior with a CTMC and compute the loss probability of type 1 customers if $\lambda_2=0$.
2. Describe the system behavior with a CTMC when both λ_1 and λ_2 are positive, and compute the loss probability of types 1 and 2 connections and the overall loss probability of connections.
3. Which loss probability is higher, that of type 1 or that of type 2 connections? Why?

4. Compute the link utilization factor when both arrival intensities are positive.
5. Compute the link utilization of type 1 and type 2 connections.

Exercise 11.2. There is a transmission link with a capacity of $C = 13\,\text{MB/s}$ that serves adaptive connections. The connections arrive according to a Poisson process at a rate λ, and their length is exponentially distributed with the parameter μ. The minimal and maximal bandwidths of the adaptive connections are $c_{\min} = 2\,\text{MB/s}$ and $c_{\max} = 3\,\text{MB/s}$, respectively. Compute the average bandwidth of an adaptive connection in equilibrium.

Exercise 11.3. There is a transmission link with a capacity of $C = 13\,\text{MB/s}$ that serves elastic connections. The connections arrive according to a Poisson process at a rate λ, and during an elastic connection an exponentially distributed amount of data is transmitted with the parameter γ. The minimal and maximal bandwidths of the elastic connections are $c_{\min} = 2\,\text{MB/s}$ and $c_{\max} = 3\,\text{MB/s}$, respectively. Compute the average bandwidth of an elastic connection in equilibrium. Compute the average time of an elastic connection in equilibrium.

Exercise 11.4. A transmission link with a capacity of $C = 3\,\text{MB/s}$ serves two kinds of elastic connections. Type 1 connections arrive according to a Poisson process at a rate $\lambda_1 = 0.5\,1/\text{s}$ and transmit an exponentially distributed amount of data with the parameter $\gamma_1 = 4\,1/\text{MB}$. The minimal and maximal bandwidths of type 1 connections are $\check{c}_1 = 1\,\text{MB/s}$ and $\hat{c}_1 = 1\,\text{MB/s}$, respectively. Type 2 connections are characterized by $\lambda_2 = 0.1\ 1/\text{s}$, $\gamma_1 = 2\ 1/\text{MB}$, $\check{c}_2 = 1\,\text{MB/s}$, and $\hat{c}_2 = 2\,\text{MB/s}$.

(a) Describe the system behavior with a CTMC.
(b) Compute the mean number of type 1 and type 2 connections.
(c) Compute the mean channel utilization.
(d) Compute the loss probability of type 1 and type 2 connections.
(e) Compute the average bandwidth of type 2 connections.

Exercise 11.5. A transmission link with a capacity of $C = 3\,\text{MB/s}$ serves two kinds of connections, elastic and adaptive. Type 1 elastic connections arrive according to a Poisson process at a rate λ_1 [1/s] and transmit an exponentially distributed amount of data with parameter γ_1 [1/MB]. The minimal and maximal bandwidths of type 1 connections are $\check{c}_1 = 0.75\,\text{MB/s}$ and $\hat{c}_1 = 1.5\,\text{MB/s}$, respectively. Type 2 adaptive connections arrive according to a Poisson process at a rate λ_2 [1/s] and stay in the system for an exponentially distributed amount of time with the parameter μ_2 [1/s]. The minimal and maximal bandwidths of type 2 connections are $\check{c}_2 = 1\,\text{MB/s}$ and $\hat{c}_2 = 2\,\text{MB/s}$, respectively.

(a) Describe the system behavior with a CTMC.
(b) Compute the loss probability of type 1 and type 2 connections.
(c) Compute the average bandwidth of type 1 and type 2 connections.
(d) Compute the mean number of type 1 and type 2 connections on the link.

Exercise 11.6. Compute the mean value of the waiting time in a cyclic waiting system.

Exercise 11.7. Let us consider our cyclic waiting system in the case of discrete time. Divide the cycle time T into n equal parts and suppose that for an interval T/n a new customer enters with probability r (there is no entry with probability $1-r$), and the service in process for such an interval is continued with probability q and completed with probability $1-q$ (i.e., the service time has a geometrical distribution). The service may be started at the moment of arrival or at moments differing from it by multiples of T.

(a) Show that the number of customers in the system at moments $t_k - 0$ constitute a Markov chain, and find its transition probabilities.
(b) Find the generating function of the number of customers in a system in equilibrium and the stability condition.

Appendix: Functions and Transforms

A.1 Nonlinear Transforms

Many theoretical and practical problems can be converted into easier forms if instead of the discrete or continuous distributions their different transforms are applied, which can be solved more readily. In probability theory, numerous transforms are applied. Denote by F the distribution function of a random variable X and by f the density function if it exists. The general form of the most frequently used transform depending on the real or complex parameter w is

$$\mathbf{E}\left(w^X\right) = \int_{-\infty}^{\infty} w^x \mathrm{d}F(x).$$

If the density function exists, then the last Riemann–Stieltjes integral can be rewritten in the form of a Riemann integral as follows:

$$\mathbf{E}\left(w^X\right) = \int_{-\infty}^{\infty} w^x f(x)\mathrm{d}x.$$

1. In the general case, setting $w = \mathrm{e}^{it}$, $t \in \mathbb{R}$, we have the characteristic function (Fourier–Stieltjes tranform)

$$\varphi(t) = \mathbf{E}\left(\mathrm{e}^{itX}\right) = \int_{-\infty}^{\infty} \mathrm{e}^{itx}\mathrm{d}F(x).$$

2. If the random variable X has a discrete distribution with values $0, 1, \ldots$ and probabilities $p_0, p_1, \ldots$ corresponding to them, then setting $z = w$, $|z| < 1$, we get

$$G(z) = \mathbf{E}\left(z^X\right) = \int_{-\infty}^{\infty} z^x \mathrm{d}F(x) = \sum_{k=0}^{\infty} p_k z^k,$$

which is the generating function of X.

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*, 369
DOI 10.1007/978-1-4614-5317-8, © Springer Science+Business Media, LLC 2013

3. The Laplace–Stieltjes transform plays a significant role when considering random variables taking only nonnegative values (usually we consider this type of random variable in queuing theory), which we obtain with $w = \mathrm{e}^{-s},\ s > 0$:

$$F^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx} \mathrm{d}F(x).$$

For the case of continuous distributions it can be rewritten in the form

$$f^*(s) = \int_0^{\infty} \mathrm{e}^{-sx} f(x)\mathrm{d}x = F^{\sim}(s),$$

where f^* denotes the Laplace transform of the density function f.

The identical background of the transformations given above determine some identical properties. When considering various problems, the use of separate transforms may be advantageous. For example, in the general case the use of a characteristic function, in the case of random variables taking nonnegative integer numbers the generating function, and in the case of general nonnegative random variables the Laplace–Stieltjes or Laplace transform is favorable to apply.

Note that we define the transforms given above for more general classes of functions than the distribution functions.

A.2 z-Transform

Let $f_0, f_1, \ldots$ be a sequence of real numbers and define the power series

$$f(z) = \sum_{n=0}^{\infty} f_n z^n = f_0 + f_1 z + f_2 z^2 + \ldots + f_n z^n + \ldots . \tag{A.1}$$

It is known from the theory of power series that if the series (A.1) is not everywhere divergent except the point $z = 0$, then there exists a number $A > 0$ such that the series (A.1) is absolute convergent ($\sum_{n=0}^{K} |f_n z^n| < \infty$) for all $|z| < A$ and divergent for all $|z| > A$. The series (A.1) may be convergent or divergent at the points $z = \pm A$ depending on the values of the parameters $f_i,\ i = 0, 1, \ldots$. The number A is called the **convergence radius** of the power series (A.1). By the Cauchy–Hadamard theorem, A can be given in the form

$$A = 1/a, \quad \text{where} \quad a = \limsup_{n\to\infty} (|f_n|)^{1/n}.$$

In the last formula we set $A = +\infty$ if $a = 0$ and $A = 0$ if $a = +\infty$. The first relation $A = +\infty$ means that the power series (A.1) is convergent in all points of the real line, and the second one means that Eq. (A.1) is convergent only at the point $z = 0$.

A finite power series $f(z) = \sum_{n=0}^{K} f_n z^n$ (K-order polynomial, which corresponds to the case $f_i = 0,\ i \geq K + 1$) is convergent at all points of the real line.

Definition A.1. Let $f_0, f_1, \ldots$ be a sequence of real numbers satisfying the condition $a = \limsup_{n\to\infty} (|f_n|)^{1/n} < \infty$. Then the power series

$$f(z) = \sum_{n=0}^{\infty} f_n z^n, \ |z| < A = 1/a,$$

is called the z-**transform** of the sequence $f_0, f_1, \ldots$

It is clear from this definition that if we use a discrete distribution $f_n, \ k \geq 0, \ \sum_{k=0}^{\infty} f_k = 1$, then the z-transform of the sequence $f_0, f_1, \ldots$ is identical with the generating function $G(z)$, which was introduced earlier.

A.2.1 Main Properties of z-Transform

1. *Derivatives.* If the convergence radius A does not equal 0, then the power series $f(z)$ is an anytime differentiable function for all points $|z| < A$ and

$$\frac{\mathrm{d}^k}{\mathrm{d}z^k} f(z) = \sum_{n=k}^{\infty} n(n-1)\ldots(n-k+1) f_n z^{n-k}, \ k \geq 1.$$

2. *Computing the coefficients of the z-transform.* For all $k = 0, 1, \ldots$ the following relation is true:

$$f_k = \frac{1}{k!} \frac{\mathrm{d}^k}{\mathrm{d}z^k} f(z)\Big|_{z=0}, \ k \geq 0. \tag{A.2}$$

   It is clear from relation (A.1) that if the condition $A > 0$ holds, then the function $f(z)$ defined by the power series (A.1) and the sequence $f_0, f_1, \ldots$ uniquely determine each other, that is, the z-transform realizes a one-to-one correspondence between the function $f(z)$ and the sequence $f_0, f_1, \ldots$ The properties of a z-transform can be analyzed using results that are true for a power series.
3. *Convolutions.* Let $f(z) = \sum_{n=0}^{\infty} f_n z^n$ and $g(z) = \sum_{n=0}^{\infty} g_n z^n$ be two z-transforms determined by the sequences f_n and g_n, respectively. Define the sequence h_n as the convolution of f_n and g_n, that is,

$$h_n = \sum_{k=0}^{n} f_k g_{n-k}, \ n \geq 0.$$

   Then the z-transform $h(z) = \sum_{n=0}^{\infty} h_n z^n$ of the sequence $h_0, h_1, \ldots$ satisfies the equation

$$h(z) = f(z) \cdot g(z).$$

A.3 Laplace–Stieltjes and Laplace Transforms in General Form

Let $H(x),\ 0 \le x < \infty$ be a function of bounded variation. A function H is said to be of bounded variation on the interval $[a,b]$ if its total variation $V_H([a,b])$ is bounded (finite). The total variation is defined as

$$V_H([a,b]) = \sup_P \sum_{k=1}^{K_P} |H(x_{P,k}) - H(_{P,k-1})|,$$

where the supremum is taken over the set of all partitions

$$P = \{x_{P,0} = a < x_{P,1} < \ldots < x_{P,K_P} = b\}$$

of the interval $[a,b]$. The function H is of bounded variation on the interval $[0,\infty)$ if $V_H([0,b])$ is bounded by some number V for all $b > 0$.

The function

$$H^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx}\mathrm{d}H(x) \tag{A.3}$$

is called the **Laplace–Stieltjes transform** of the function H. If the function H can be given in the integral form

$$H(x) = \int_0^x h(u)\mathrm{d}u,\ x \ge 0,$$

where h is an integrable function (this means that H is an absolute continuous function with respect to the Lebesgue measure), then the Laplace transform of the function h satisfies the equation

$$h^*(s) = \int_0^{\infty} \mathrm{e}^{-sx}h(x)\mathrm{d}x = H^{\sim}(s).$$

Theorem A.2. *If the integral (A.3) is convergent for $s > 0$, then $H^{\sim}(s),\ s > 0$ is an analytic function, and for every positive integer k*

$$\frac{\mathrm{d}^k}{\mathrm{d}s^k}H^{\sim}(s) = \int_0^{\infty} \mathrm{e}^{-sx}(-x)^k\mathrm{d}H(x).$$

The transform $H^{\sim}$ satisfies the following asymptotic relation [90]. If the integral (A.3) is convergent for $\mathrm{Re}s > 0$ and there exist constants $\alpha \ge 0$ and A such that

$$\lim_{x\to\infty} \frac{H(x)}{x^{\alpha}} = \frac{A}{\Gamma(\alpha+1)},$$

then the convergence

$$\lim_{s\to 0+} s^{\alpha} H^{\sim}(s) = A \tag{A.4}$$

holds.

Theorem A.3. *Assume that there exists a function $h(x),\ x \geq 0$, and its Laplace transform $h^*(s),\ s > 0$; moreover, the function $h(x)$ is convergent as $x \to \infty$, i.e., $\lim_{x\to\infty} h(x) = h_\infty$. Then*

$$\lim_{s\to 0+} sh^*(s) = h_\infty.$$

Proof. Denote $H(x) = \int_0^x h(s)\mathrm{d}s$, $x \geq 0$. Choosing $\alpha = 1$ we have

$$\lim_{x\to\infty} \frac{H(x)}{x} = \lim \frac{1}{x} \int_0^x h(s)\mathrm{d}s = h_\infty = \frac{h_\infty}{\Gamma(1+1)};$$

thus by relation (A.4) the assertion of the theorem immediately follows.

Theorem A.4. *If there exists a Laplace transform f^* of the nonnegative function $f(t)$, $t \geq 0$, and there exists the finite limit $\lim_{x\to 0+} f(x) = f_0$, then*

$$\lim_{s\to\infty} sf^*(s) = f_0.$$

Proof. It is clear that

$$s \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}x = \int_0^\infty \mathrm{e}^{-x}\mathrm{d}x = 1$$

and

$$s \int_{1/\sqrt{s}}^\infty \mathrm{e}^{-sx}\mathrm{d}x = \int_{\sqrt{s}}^\infty \mathrm{e}^{-y}\mathrm{d}y = \mathrm{e}^{-\sqrt{s}} = o(1),\ s \to \infty;$$

therefore,

$$sf^*(s) - f_0 = s \int_0^{1/\sqrt{s}} \mathrm{e}^{-sx}[f(x) - f_0]\mathrm{d}x + s \int_{1/\sqrt{s}}^\infty \mathrm{e}^{-sx} f(x)\mathrm{d}x + f_0 o(1).$$

Since there exists the finite limit $\lim_{x\to\infty} f(x) = f_0$, with the notation

$$\delta(z) = \sup_{0<x\leq z} |f(x) - f_0| \to 0,\ z \to 0+,$$

we obtain

$$s \int_0^{1/\sqrt{s}} \mathrm{e}^{-sx}\,|f(x) - f_0|\,\mathrm{d}x < \delta(1/\sqrt{s}) \int_0^\infty s\mathrm{e}^{-sx}\mathrm{d}x = \delta(1/\sqrt{s}) \to 0,\ s \to \infty.$$

On the other hand, for all $0 < s \le t$ the relation

$$f^*(s) = \int_0^\infty \mathrm{e}^{-sx} f(x)\mathrm{d}x \le \int_0^\infty \mathrm{e}^{-tx} f(x)\mathrm{d}x = f^*(t),$$

is true, then

$$\begin{aligned} s\int_{1/\sqrt{s}}^\infty \mathrm{e}^{-sx} f(x)\mathrm{d}x &\le s\mathrm{e}^{-(1/2)\sqrt{s}} \int_{1/\sqrt{s}}^\infty \mathrm{e}^{-(s/2)x} f(x)\mathrm{d}x \\ &\le s\mathrm{e}^{-(1/2)\sqrt{s}} \int_0^\infty \mathrm{e}^{-(s/2)x} f(x)\mathrm{d}x = s\mathrm{e}^{-(1/2)\sqrt{s}} f^*(s/2) \\ &\le s\mathrm{e}^{-(1/2)\sqrt{s}} f^*(1) \to 0, \end{aligned}$$

as $s \to \infty$ ($s \ge 2$). Summing up the results obtained above, the assertion of the theorem follows.

A.3.1 Examples of Laplace Transform of Some Distributions

(a) Deterministic distribution ($a > 0$, $\mathbf{P}(X = a) = 1$):

$$F^\sim(s) = \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}F(x) = \mathrm{e}^{-sa}, \quad \mathbf{E}(X) = a.$$

(b) $B(n, p)$ binomial distribution:

$$\begin{aligned} F^\sim(s) &= \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}F(x) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k}\mathrm{e}^{-sk} \\ &= \sum_{k=0}^n \binom{n}{k} (p\mathrm{e}^{-s})^k (1-p)^{n-k} = [1 + p(\mathrm{e}^{-s} - 1)]^n, \\ \mathbf{E}(X) &= np\mathrm{e}^{-s}[1 + p(\mathrm{e}^{-s} - 1)]^{n-1}\big|_{s=0} = np. \end{aligned}$$

(c) Poisson distribution with parameter λ:

$$\begin{aligned} F^\sim(s) &= \int_0^\infty \mathrm{e}^{-sx}\mathrm{d}F(x) = \sum_{k=0}^\infty \mathrm{e}^{-sk}\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} \\ &= \sum_{k=0}^\infty \frac{1}{k!}(\lambda\mathrm{e}^{-s})^k\mathrm{e}^{-\lambda} = \exp\{\lambda(\mathrm{e}^{-s} - 1)\}, \\ \mathbf{E}(X) &= \lambda\mathrm{e}^{-s}\exp\{\lambda(\mathrm{e}^{-s} - 1)\}\big|_{s=0} = \lambda. \end{aligned}$$

(d) Uniform distribution on the interval $[a, b]$:

$$F^{\sim}(s) = \int_a^b \mathrm{e}^{-sx} \frac{1}{b-a} \mathrm{d}x = \begin{cases} \frac{1}{s(b-a)}(\mathrm{e}^{-sa} - \mathrm{e}^{-sb}), & s > 0, \\ 1, & s = 0. \end{cases}$$

and by the use of l'Hospital's rule:

$$\begin{aligned} \mathbf{E}(X) &= \frac{1}{b-a} \lim_{s \to 0+} -\frac{1}{s^2} \left([\mathrm{e}^{-sa} - \mathrm{e}^{-sb}] - [sa\mathrm{e}^{-sa} - sb\mathrm{e}^{-sb}] \right) \\ &= \frac{1}{b-a} \lim_{s \to 0+} \frac{b^2 s\mathrm{e}^{-sb} - a^2 \mathrm{e}^{-sa}}{2s} = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}. \end{aligned}$$

(e) Exponential distribution with parameter λ:

$$\begin{aligned} F^{\sim}(s) &= \int_0^\infty \mathrm{e}^{-sx} \lambda \mathrm{e}^{-\lambda x} \mathrm{d}x = \lambda \int_0^\infty \mathrm{e}^{-(s+\lambda)x} \mathrm{d}x = \frac{\lambda}{s+\lambda}, \\ \mathbf{E}(X) &= \left. \frac{\lambda}{(s+\lambda)^2} \right|_{s=0} = \frac{1}{\lambda}. \end{aligned}$$

A.3.2 Sum of a Random Number of Independent Random Variables

Theorem A.5. *Let K be a random variable with nonnegative integer values, and consider the sum of K random variables $Y = \sum_{n=0}^{K} X_n$, where*

(1) The random variables K and $\{X_n,\ n \geq 0\}$ are independent.
(2) The distributions of the random variables X_n are identical with common distribution function $F(x)$.

Denote by $F_X^{\sim}(s)$ the Laplace–Stieltjes transform of X_n and by $G_K(z)$ the generating function of K. Then the Laplace–Stieltjes transform of random variable Y has the form

$$\mathbf{E}\left(\mathrm{e}^{-sY}\right) = G_K(F_X^{\sim}(s)).$$

Proof. Since

$$\mathbf{E}\left(\exp\left\{ -s \sum_{n=0}^{K} X_n \right\} \;\middle|\; K = k \right) = F_X^{\sim}(s)^k,$$

then we obtain by the use of the formula of total expected value

$$\begin{aligned}
&\mathbf{E}\left(\exp\left\{-s\sum_{n=0}^{K}X_n\right\}\right)\\
&\quad=\sum_{k=0}^{\infty}\left[\mathbf{E}\left(\exp\left\{-s\sum_{n=0}^{K}X_n\right\}\;\middle|\;K=k\right)\mathbf{P}(K=k)\right]\\
&\quad=\sum_{k=0}^{\infty}F_X^{\sim}(s)^k\,\mathbf{P}(K=k)=\mathbf{E}\left(F_X^{\sim}(s)^K\right)=G_K(F_X^{\sim}(s)).
\end{aligned}$$

A.4 Bessel and Modified Bessel Functions of the First Kind

Definition A.6. The nonzero solutions of Bessel's differential equation

$$x^2u''+xu'+(x^2-\nu^2)u=0 \tag{A.5}$$

are called ν**-order Bessel functions**, where ν is a real number.

Definition A.7. The solutions of Bessel's differential equation are called **Bessel functions of the first kind** and denoted by $J_\nu(x)$, which are nonsingular at the origin $x=0$.

The ν-order Bessel functions of the first kind $J_\nu(x)$ ($\nu\geq 0$) can be defined by their Taylor series expansion around $x=0$ as follows:

$$J_\nu(x)=\sum_{k=0}^{\infty}\frac{(-1)^k}{\Gamma(k+\nu+1)\Gamma(k+1)}\left(\frac{x}{2}\right)^{2k+\nu}, \tag{A.6}$$

where $\Gamma(x)=\int_0^\infty \mathrm{e}^{-t}t^{x-1}\mathrm{d}t$ is the gamma function. This formula is valid, providing $\nu\neq -1,-2,\ldots$. The Bessel function

$$J_{-\nu}(x)=\sum_{k=0}^{\infty}\frac{(-1)^k}{\Gamma(k+\nu+1)\Gamma(k+1)}\left(\frac{x}{2}\right)^{2k-\nu}$$

is given by replacing ν in Eq. (A.6) with a $-\nu$.

An important special case of a Bessel function of the first kind is that of a purely imaginary argument.

Definition A.8. The function

$$I_\nu(x)=i^{-\nu}J_\nu(ix)=\sum_{k=0}^{\infty}\frac{1}{\Gamma(k+\nu+1)\Gamma(k+1)}\left(\frac{x}{2}\right)^{2k+\nu}$$

is calleda **modified** ν**-order Bessel function of the first kind**.

Both the Bessel functions $J_\nu(x)$ and $I_\nu(x)$ can be expressed in terms of the generalized hypergeometric function ${}_0F_1(\nu;x)$ as follows [76]:

$$J_\nu(x) = \frac{1}{\Gamma(\nu+1)}\left(\frac{x}{2}\right)^\nu {}_0F_1(\nu+1;-\frac{x^2}{4}),$$
$$I_\nu(x) = \frac{1}{\Gamma(\nu+1)}\left(\frac{x}{2}\right)^\nu {}_0F_1(\nu+1;\frac{x^2}{4}),$$

where

$${}_0F_1(\nu;x) = \sum_{k=0}^{\infty} \frac{\Gamma(\nu)}{\Gamma(k+\nu)\Gamma(k+1)} x^k.$$

A.5 Notations

$\mathbb{N}^+$	Set of nonnegative integer numbers
$\mathbb{R}$	Set of real numbers ($\mathbb{R} = (-\infty, \infty)$)
δ_{ij}	Kronecker delta function, that is, $\delta_{ij} = 1$, if $i = j$; otherwise it equals 0
a^+	Positive part of a real number a, i.e., $a^+ = \max(a, 0)$
$\overline{A}$	Complementary event of A
$\mathcal{I}_{\{A\}}$	Indicator function of an event A, that is, it equals 1 if the event A occurs, and otherwise it equals 0
$\mathbf{P}(A)$	Probability of an event A
$\mathbf{E}(X)$	Expected value of a random variable X
$\mathbf{D}(X)$	Variation of a random variable X
S	State space of a Markov chain
$\boldsymbol{P}$	(One-step) transition probability matrix of a discrete-time Markov chain
$\boldsymbol{Q}$	Rata matrix of a continuous-time Markov chain
$\boldsymbol{I}$	Unit matrix

Bibliography

1. 802.11. IEEE standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements - part 11: Wireless LAN medium access control (mac) and physical layer (phy) specifications. http://ieeexplore.ieee.org/servlet/opac?punumber=4248376, 2007.
2. N. Abramson. The aloha system: another alternative for computer communications. In: *Proceedings Fall Joint Computer Conference*. AFIPS Press, 1970.
3. D. Aldous, L. Shepp. The least variable phase type distribution is Erlang. *Stoch. Models*, 3:467–473, 1987.
4. T. Apostol. *Calculus I.* Wiley, New York, 1967.
5. T. Apostol. *Calculus II.* Wiley, New York, 1969.
6. J. R. Artalejo, A. Gómez-Corral. *Retrial Queueing Systems: A Computational Approach*. Springer, Berlin Heidelberg New York, 2008.
7. S. Asmussen. *Applied Probability and Queues*. Springer, Berlin Heidelberg New York, 2003.
8. F. Baccelli, P. Brémaud. *Elements of Queueing Theory, Applications of Mathematics*. Springer, Berlin Heidelberg New York, 2002.
9. F. Baskett, K. Mani Chandy, R. R. Muntz, F. G. Palacios. Open, closed and mixed networks of queues with different classes of customers. *J. ACM*, 22:248–260, 1975.
10. S. N. Bernstein. *Theory of Probabilities*. Moskva, Leningrad, 1946. (in Russian).
11. G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE J. Select. Areas Commun.*, 18:535–547, 2000.
12. D. Bini, G. Latouche, B. Meini. *Numerical methods for structured Markov chains*. Oxford University Press, Oxford, 2005.
13. A. Bobbio, M. Telek. A benchmark for PH estimation algorithms: results for Acyclic-PH. *Stoch. Models*, 10:661–677, 1994.
14. A. A. Borovkov. *Stochastic processes in queueing theory*. Applications of Mathematics. Springer, Berlin Heidelberg New York, 1976.
15. A. A. Borovkov. *Asymptotic Methods in Queueing Theory*. Wiley, New York, 1984.
16. L. Breuer, D. Baum. *An Introduction to Queueing Theory and Matrix-Analytic Methods*. Springer, Berlin Heidelberg New York, 2005.
17. P. J. Burke. The output of a queuing system. *Oper. Res.*, 4:699–704, 1956.
18. J. Buzen. Computational algorithms for closed queueing networks with exponential servers. *Commun. ACM*, 16:527–531, 1973.
19. V. Ceric, L. Lakatos. Measurement and analysis of input data for queueing system models used in system design. *Syst. Anal. Modell. Simul.*, 11:227–233, 1993.
20. Hong Chen, David D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, Berlin Heidelberg New York, 2001.
21. Y. Chow, H. Teicher. *Probability Theory*. Springer, Berlin Heidelberg New York, 1978.

DOI 10.1007/978-1-4614-5317-8

22. K. Chung. *Markov chains with stationary transition probabilities*. Springer, Berlin Heidelberg New York, 1960.
23. E. Cinlar. *Introduction to Stochastic Processes*. Prentice-Hall, Englewood Cliffs, NJ, 1975.
24. D. R. Cox. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proc. Cambridge Philos. Soc.*, 51:433–440, 1955.
25. A. Cumani. On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectron. Reliab.*, 22:583–602, 1982.
26. D.J. Daley, D. Vere-Jones. *An Introduction to the Theory of Point Process*. Springer, Berlin Heidelberg New York, 2008. 2nd edn.
27. Gy. Dallos, Cs. Szabó. *Random access methods of communication channels*. Akadémiai Kiadó, Budapest, 1984 (in Hungarian).
28. M. De Prycker. *Asynchronous Transfer Mode, Solutions for Broadband ISDN*. Prentice Hall, Englewood Cliffs, NJ, 1993.
29. P. Erdős, W. Feller, H. Pollard. A theorem on power series. *Bull. Am. Math. Soc.*, 55:201–203, 1949.
30. G. I. Falin, J. G. C. Templeton. *Retrial queues*. Chapman and Hall, London, 1997.
31. W. Feller. *An Introduction to Probability Theory and its Applications*, vol. I. Wiley, New York, 1968.
32. Chuan Heng Foh, M. Zukerman. Performance analysis of the IEEE 802.11 MAC protocol. In *Proceedings of European wireless conference*, Florence, February 2002.
33. F. G. Foster. On the stochastic matrices associated with certain queuing processes. *Ann. Math. Stat.*, 24:355–360, 1953.
34. G. Giambene. *Queuing Theory and Telecommunications: Networks and Applications*. Springer, Berlin Heidelberg New York, 2005.
35. I.I. Gihman, A.V. Skorohod. *The Theory of Stochastic Processes*, vol. I. Springer, Berlin Heidelberg New York, 1974.
36. I. I. Gihman, A. V. Skorohod. *The Theory of Stochastic Processes*, vol. II. Springer, Berlin Heidelberg New York, 1975.
37. B. Gnedenko, E. Danielyan, B. Dimitrov, G. Klimov, V. Matveev. *Priority Queues*. Moscow State University, Moscow, 1973 (in Russian).
38. B. V. Gnedenko. *Theory of Probability*. Gordon and Breach, Amsterdam, 1997. 6th edn.
39. B. V. Gnedenko, I. N. Kovalenko. *Introduction to Queueing Theory*, 2nd edn. Birkhauser, Boston 1989.
40. W. J. Gordon, G. F. Newell. Closed queueing systems with exponential servers. *Oper. Res.*, 15:254–265, 1967.
41. D. Gross, J. F. Shortle, J. M. Thompson, C. M. Harris. *Fundamentals of Queueing Theory*, 4th edn. Wiley, New York, 2008.
42. W. Henderson. Alternative approaches to the analysis of the M/G/1 and G/M/1 queues. *J. Oper. Res. Soc. Jpn.*, 15:92–101, 1972.
43. A. Horváth, M. Telek. PhFit: A general purpose phase type fitting tool. In *Tools 2002*, pages 82–91, London, April 2002. Lecture Notes in Computer Science, vol. 2324. Springer, Berlin Heidelberg New York.
44. J. R. Jackson. Jobshop-like queueing systems. *Manage. Sci.*, 10:131–142, 1963.
45. N. K. Jaiswal. *Priority Queues*. Academic, New York, 1968.
46. N. L. Johnson, S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions*. Applied Probability and Statistics. Wiley, New York, 1972.
47. V.V. Kalashnikov. *Mathematical Methods in Queueing Theory*. Kluwer, Dordrecht, 1994.
48. S. Karlin, H. M. Taylor. *A First Course in Stochastic Processes*. Academic, New York, 1975.
49. S. Karlin, H. M. Taylor. *A Second Course in Stochastic Processes*. Academic, New York, 1981.
50. J. Kaufman. Blocking in a shared resource environment. *IEEE Trans. Commun.*, 29: 1474–1481, 1981.
51. F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, New York, 1979.
52. D. G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann. Math. Stat.*, 24:338–354, 1953.

53. A. Khinchin. Mathematisches über die Erwartung vor einem öffentlichen Schalter. *Rec. Math.*, 39:72–84, 1932 (in Russian with German summary).
54. J. F. C. Kingman. *Poisson Processes*. Clarendon, Oxford, 1993.
55. L. Kleinrock. *Queuing Systems. Volume 1: Theory*. Wiley-Interscience, New York, 1975.
56. G. P. Klimov. *Extremal Problems in Queueing Theory*. Energia, Moskva, 1964 (in Russian).
57. E. V. Koba. On a retrial queueing system with a FIFO queueing discipline. *Theory Stoch. Proc.*, 8:201–207, 2002.
58. V. G. Kulkarni. *Modeling and Analysis of Stochastic Systems*. Chapman & Hall, London, 1995.
59. L. Lakatos. On a simple continuous cyclic waiting problem. *Annal. Univ. Sci. Budapest Sect. Comp.*, 14:105–113, 1994.
60. L. Lakatos. A note on the Pollaczek-Khinchin formula. *Annal. Univ. Sci. Budapest Sect. Comp.*, 29:83–91, 2008.
61. L. Lakatos. Cyclic waiting systems. *Cybern. Syst. Anal.*, 46:477–484, 2010.
62. G. Latouche, V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*. SIAM, 1999.
63. A. Lewandowski. Statistical tables. http://www.alewand.de. Nov. 13., 2012.
64. D. V. Lindley. The theory of queues with a single server. *Math. Proc. Cambridge Philos. Soc.*, 48:277–289, 1952.
65. T. Lindwall. *Lectures on the Coupling Method*. Wiley, New York, 1992.
66. J. D. C. Little. A proof of the queuing formula: L =AW. *Oper. Res.*, 9:383–387, 1961.
67. A. A. Markov. Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 15:135–156, 1906 (in Russian).
68. L. Massoulie, J. Roberts. Bandwidth sharing: Objectives and Algorithms. In *Infocom*, 1999.
69. V. F. Matveev, V. G. Ushakov. *Queueing systems*. Moscow State University, Moskva, 1984 (in Russian).
70. P. Medgyessy, L. Takács. *Probability Theory*. Tankönyvkiadó, Budapest, 1973 (in Hungarian).
71. S. Meyn, R. Tweedie. *Markov chains and stochastic stability*. Springer, Berlin Heidelberg New York, 1993.
72. NIST: National Institute of Standards and Technology. Digital library of mathematical functions. http://dlmf.nist.gov. Nov. 13., 2012.
73. M. Neuts. Probability distributions of phase type. In *Liber Amicorum Prof. Emeritus H. Florin*, pp. 173–206. University of Louvain, Louvain, Belgium, 1975.
74. M.F. Neuts. *Matrix Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore, 1981.
75. C. Palm. Methods of judging the annoyance caused by congestion. *Telegrafstyrelsen*, 4:189–208, 1953.
76. A. P. Prudnikov, Y. A. Brychkov, O. I. Marichev. *Integrals and series*, vol. 2. Gordon and Breach, New York, 1986. Special functions.
77. S. Rácz, M. Telek, G. Fodor. Call level performance analysis of 3rd generation mobile core network. In *IEEE International Conference on Communications, ICC 2001*, 2:456–461, Helsinki, Finland, June 2001.
78. S. Rácz, M. Telek, G. Fodor. Link capacity sharing between guaranteed- and best effort services on an atm transmission link under GoS constraints. *Telecommun. Syst.*, 17(1–2):93–114, 2001.
79. M. Reiser, S. S. Lavenberg. Mean value analysis of closed multi-chain queueing networks. *J. ACM*, 27:313–322, 1980.
80. J. Roberts. A service system with heterogeneous user requirements - application to multi-service telecommunications systems. In *Proceedings of Performance of Data Communications Systems and Their Applications*, pp. 423–431, Paris, 1981.
81. K. W. Ross. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer, Berlin Heidelberg New York, 1995.
82. T. Saaty. *Elements of Queueing Theory*. McGraw-Hill, New York, 1961.
83. R. Serfozo. *Introduction to Stochastic Networks*. Springer, Berlin Heidelberg New York, 1999.
84. A. N. Shiryaev. *Probability*. Springer, Berlin Heidelberg New York, 1994.

85. D. L. Snyder. *Random Point Processes*. Wiley, New York, 1975.
86. L. Szeidl. Estimation of the moment of the regeneration period in a closed central-server queueing network. *Theory Probab. Appl.*, 31:309–313, 1986.
87. L. Szeidl. On the estimation of moment of regenerative cycles in a general closed central-server queueing network. *Lect. Notes Math.*, 1233:182–189, 1987.
88. L. Takács. Investigation of waiting time problems by reduction to Markov processes. *Acta Math. Acad. Sci. Hung.*, 6:101–129, 1955.
89. L. Takács. The distribution of the virtual waiting time for a single-server queue with Poisson input and general service times. *Oper. Res.*, 11:261–264, 1963.
90. L. Takács. *Combinatorial Methods in the Theory of Stochastic Processes*. Wiley, New York, 1967.
91. H. Takagi. *Queueing Analysis*. North Holland, Amsterdam, 1991.
92. M. Telek. Minimal coefficient of variation of discrete phase type distributions. In G. Latouche, P. Taylor, eds., *Advances in algorithmic methods for stochastic models, MAM3*, pp. 391–400. Notable Publications, 2000.
93. A. Thümmler, P. Buchholz, M. Telek. A novel approach for fitting probability distributions to trace data with the em algorithm. *IEEE Trans. Depend. Secure Comput.*, 3(3):245–258, 2006. Extended version of DSN 2005 paper.
94. H. Tijms. *Stochastic Models: An Algorithmic Approach*. Wiley, New York, 1994.
95. W. Whitt. A review of $l = \lambda w$ and extensions. *Queue. Syst.*, 9:235–268, 1991.
96. V. M. Zolotarev. *Modern Theory of Summation of Random Variables*. VSP, Utrecht, 1997.

Index

L. Lakatos et al., *Introduction to Queueing Systems with Telecommunication Applications*,
DOI 10.1007/978-1-4614-5317-8, © Springer Science+Business Media, LLC 2013