Stochastic Models in Reliability and Maintenance

Springer-Verlag Berlin Heidelberg GmbH

Shunji Osaki
Editor

Stochastic Models in Reliability and Maintenance

With 46 Figures
and 11 Tables

Springer

Professor Shunji Osaki
Nanzan University
Department of Information & Telecommunication Engineering
Faculty of Mathematical Sciences
and Information Engineering
27 Seirei-cho, Seto-shi
Aichi, 489-0863, Japan
e-mail: shunji@nanzan-u.ac.jp

ISBN 978-3-642-07725-8 ISBN 978-3-540-24808-8 (eBook)
DOI 10.1007/978-3-540-24808-8

Library of Congress Cataloging-in-Publication Data applied for
Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Stochastic Models in Reliability and Maintenance: 11 Tables / Ed.: Shunji Osaki. –
Berlin; Heidelberg; New York; Barcelona; Hong Kong; London; Milan; Paris; Tokyo:
Springer, 2002

http://www.springer.de

Originally published by Springer-Verlag Berlin Heidelberg in 2002
Softcover reprint of the hardcover 1st edition 2002

Cover design: Erich Kirchner, Heidelberg

Preface

Our daily lives can be maintained by the high-technology systems. Computer systems are typical examples of such systems. We can enjoy our modern lives by using many computer systems. Much more importantly, we have to maintain such systems without failure, but cannot predict when such systems will fail and how to fix such systems without delay. A stochastic process is a set of outcomes of a random experiment indexed by time, and is one of the key tools needed to analyze the future behavior quantitatively. Reliability and maintainability technologies are of great interest and importance to the maintenance of such systems. Many mathematical models have been and will be proposed to describe reliability and maintainability systems by using the stochastic processes.

The theme of this book is "Stochastic Models in Reliability and Maintainability." This book consists of 12 chapters on the theme above from the different viewpoints of stochastic modeling. Chapter 1 is devoted to "Renewal Processes," under which classical renewal theory is surveyed and computational methods are described. Chapter 2 discusses "Stochastic Orders," and in it some definitions and concepts on stochastic orders are described and aging properties can be characterized by stochastic orders. Chapter 3 is devoted to "Classical Maintenance Models," under which the so-called age, block and other replacement models are surveyed. Chapter 4 discusses "Modeling Plant Maintenance," describing how maintenance practice can be carried out for plant maintenance. Chapter 5 is devoted to "Imperfect Preventive Maintenance," reviewing some imperfect maintenance models. Chapter 6 discusses "Generalized Renewal Processes," and generalized renewal processes and general repair models are described. Chapter 7 is devoted to "Two-Unit Redundant Models," under which several two-unit redundant models are surveyed and their optimization policies are analyzed. Chapter 8 focuses on "Markovian Deteriorating Systems," from which optimal inspection and/or replacement policies are derived. Chapter 9 is devoted to "Semi-Markov Reliability Models," where transient analyses of semi-Markov reliability models are surveyed and numerical solution techniques are described. Chapter 10 discusses "Software Reliability Models," reviewing typical software reliability models. Chapter 11 is devoted to "Data Communication Systems," showing how data communication systems can be formulated by means of reliability models, and

optimal policies are discussed. Chapter 12 discusses "Monte-Carlo Methods": here some Monte-Carlo methods are surveyed in reliability models.

This book will be of interest and importance to such users of "Stochastic Models in Reliability and Maintenance" as operations research analysts, probabilists, statisticians, applied mathematicians, engineers, and graduate students who are interested in reliability and maintenance. References leading further into the subject matter are cited in the end of each chapter. I believe that all the chapters of this book will introduce the readers to the major up-to-date theory and practice in stochastic models in reliability and maintenance.

I would like to express my sincere appreciation to all the contributors to this book. I am also indebted to Professor Tadashi Dohi, Dr. Takashi Satow, and Dr. Hiroyuki Okamura, Hiroshima University, Japan, for their patient support in completion of the editing. Finally, I would like to express my sincere appreciation to Dr. Werner A. Müller, Springer-Verlag, Heidelberg, Germany, for his friendly and patient help.

Shunji Osaki

Contents

1. Renewal Processes and Their Computational Aspects

Tadashi Dohi
Department of Industrial and Systems Engineering,
Hiroshima University
Higashi–Hiroshima 739–8527, Japan,
Naoto Kaio
Department of Economic Informatics,
Hiroshima Shudo University
Hiroshima 731–3195, Japan
and
Shunji Osaki
Department of Information & Telecommunication Engineering,
Nanzan University
Aichi 489–0863, Japan

Summary.

In this chapter, we review classical renewal theory and focus on the computational aspects for the renewal function. As well known, the renewal processes have an important role in understanding the discrete event systems arising in queueing theory, production and inventory control, design of communication systems, performance evaluation in computer science and product warranty estimation and also in reliability and maintenance modeling. On the other hand, from the practical perspective, since the computation of the renewal function is not so easy, the system analyst tends to treat the renewal function via the simplest method for him or her. In fact, a large number of authors have discussed the computation problems of the renewal function. Nevertheless, no articles reporting those results in a systematic way have appeared in the literature. In this chapter, we survey the computational aspects of the renewal function based on the most recent results obtained up to the present stage. A comprehensive bibliography in this research area is also provided.

Keywords: renewal processes, renewal function, approximations, integral equation, series expansion, bounds, interpolation

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

1.1 Introduction

The stochastic processes provide the basis for modeling a variety of uncertain behaviors that we face in our daily life. Especially, the renewal processes play an important role in understanding the discrete event systems arising in queueing theory [1], [2], production and inventory control [3], design of communication systems, performance evaluation in computer science [4] and product warranty estimation [5], [6] as well as reliability and maintenance modeling [7], [8], [9], [10], [11]. In fact, almost standard textbooks on the stochastic processes [12], [13], [14], [15] spend many pages on renewal theory. First, the renewal theory was developed by Feller [16], [17]. After those, Smith [18], [19] and Cox [20] published their excellent papers and monograph, respectively, which include most of the important results on renewal theory established before 1960s. On the other hand, from the practical perspective, since the computation of the renewal function is not so easy, the system analyst tends to treat the renewal function via the simplest method for him or her. In fact, a great number of authors have discussed the computation problems of the renewal function. Nevertheless, no articles reporting those results in a systematic way have appeared in the literature.

In this chapter, we review mainly the computational aspects of the renewal function based on the most recent results obtained up to the present stage. The computation methods discussed here are classified into two categories: analytical approximation method and numerical computation method. The former corresponds to approximating the renewal function based on its mathematical structure, the latter to computing it numerically. In general, the numerical computation methods proposed in the earlier literature can calculate the value of renewal function accurately. However, in many applications, it is often necessary to evaluate any probabilistic quantity based on the analytical form of renewal function. For instance, if one considers the mean inter-departure time in a $GI/G/1$ loss system, then the expected number of renewals in a random interval has to be evaluated. Hence, the analytical approach to approximating the renewal function may be very useful for treating such a complex probabilistic quantity. Consequently, it will be meaningful for the system analyst to master the computation methods of renewal function from both the directions above.

The rest of this chapter proceeds as follows. In Section 1.2, the basic theory of both continuous and discrete renewal processes is outlined. Section 1.3 describes some useful properties of the renewal function. After introduction of a few solvable examples in which the explicit form of the renewal function is available, some specific methods for a special class of inter-arrival (or inter-failure) time distributions are discussed. Further, we consider the asymptotic properties of the renewal function. In Section 1.4, some analytical methods are developed to approximate the renewal function. First, we describe the so-called phase approximation method, which approximates the underlying renewal process by the phase-type renewal processes. Next, two analytical methods with the known theoretical distributions and the equilibrium distribution are explained, respectively. Since these methods are tractable and, at the same time, are developed based on the mathematical structure of the renewal function, it may be easier to evaluate the error bound for the approximation formulae than the (conceptually) rough approximation such as the phase method. Also,

the basic ideas behind these approximations are unique and are applicable to develop any approximation formula for the wider class of stochastic processes.

Section 1.5 is devoted to obtaining some bounds of the renewal function. The evaluation of the renewal function by its upper and lower bounds is interesting as it can disclose the qualitative properties of the renewal function in the shorter time domain, since the asymptotic behaviour of the renewal function is well known. We introduce three kinds of bounds. Their existence depends on the aging property of the underlying inter-arrival time distribution function. In other words, the upper and lower bounds of the renewal function can be evaluated for only a few classes of random variables denoting the inter-arrival time. In Section 1.6, we discuss four typical methods to calculate the renewal function numerically. Since the recent development of computer technology, rather complex numerical calculation problems can be solved within a satisfactory computation time. This fact will support us in executing the numerical calculation of the renewal function and its variations on computer, if we require to seek them with a high degree of accuracy. More specifically, we introduce the rational function approximations as well as the classical methods such as the Laplace inversion techniques, the cubic spline algorithms and the discritization algorithms. Finally, some remarks are presented in Section 1.7. Here, we refer to the other methods of evaluating the renewal function and describe the future view in the renewal approximation theory. A comprehensive bibliography in this research area is also provided.

1.2 Basic Renewal Theory

1.2.1 Continuous renewal theory

First of all, let us define the terminologies used in this chapter. The time to occurrence of an event, X, is called the arrival time or *the failure time*, and can be defined as a non-negative random variable. Suppose that X (≥ 0) is a continuous random variable having probability density function $f(t)$ ($t \geq 0$). Then the probability distribution function or the failure time distribution, $F(t)$, is represented as $F(t) = \Pr\{X \leq t\} = \int_0^t f(x)dx$ ($t \geq 0$), where $F(0) = 0$ and $F(\infty) = 1$. If the event means a failure, *the reliability function* is defined as the probability that the unit does not fail until time t, and is denoted by $\overline{F}(t) = \Pr\{X > t\} = \int_t^\infty f(x)dx$, where $\overline{\phi}(\cdot) = 1 - \phi(\cdot)$ in general, $\overline{F}(0) = 1$ and $\overline{F}(\infty) = 0$. Denoting the mathematical expectation operator by E, then $\mathrm{E}[X] = \int_0^\infty xdF(x) = \int_0^\infty \overline{F}(x)dx$ is said MTTF (mean time to failure) or if not, MTBF (mean time between failures).

Define the probability that the unit fails at $(t, t + \triangle]$, provided that it is operative at time t, as follows.

$$\Pr\{t < X \leq t + \triangle | X > t\} = \frac{f(t)}{\overline{F}(t)} \cdot \triangle + o(\triangle), \tag{1.1}$$

where $\Pr\{A_1 \mid A_2\}$ is the conditional probability for the event A_1 provided that the event A_2 occurs, and the function $o(\triangle)$ satisfies $\lim_{\triangle \to 0} o(\triangle)/\triangle = 0$. Then, the function $r(t) = f(t)/\overline{F}(t)$ is called *the failure rate*

or the hazard rate. Using the failure rate, the reliability function and the probability density function are represented as $\overline{F}(t) = \exp\{-\int_0^t r(x)dx\}$ and $f(t) = r(t)\exp\{-\int_0^t r(x)dx\}$, respectively. In particular, the function $H(t) = \int_0^t r(x)dx$ is called *the cumulative hazard* or simply the mean value function.

Define the convolution between two failure time distributions $F(t)$ and $G(t)$ $(t \geq 0)$ by

$$F * G(t) = \int_0^t F(t-x)g(x)dx = \int_0^t G(t-x)f(x)dx, \tag{1.2}$$

where $g(t) = dG(t)/dt$. For the identical distributions, the n-fold $(n = 0, 1, 2, \cdots)$ convolution of the failure time distribution, $F^{(n)}(t)$, is recursively defined as follows:

$$F^{(0)}(t) = 1(t), \tag{1.3}$$

$$F^{(n)}(t) = F * F^{(n-1)}(t), \qquad n = 1, 2, 3, \cdots, \tag{1.4}$$

where $1(t)$ is the unit step function. Similarly, the n-fold convolution of the probability density function $f^{(n)}(t)$ becomes

$$f^{(0)}(t) = \delta(t), \tag{1.5}$$

$$\begin{aligned} f^{(n)}(t) &= \int_0^t f^{(n-1)}(t-x)f(x)dx \\ &= \frac{dF^{(n)}(t)}{dt}, \qquad n = 1, 2, 3, \cdots, \end{aligned} \tag{1.6}$$

where $\delta(t)$ is the Dirac's delta function. Evidently, the n-fold convolution of the failure time distribution means the probability that a unit fails n times up to time t.

Let $N(t)$ $(t \geq 0)$ be the number of failures (renewals) during the time interval $(0, t]$. More specifically, if the inter-failure times $X_1, X_2, \cdots$ are identically and independently distributed non-negative random variables, the resulting stochastic process $\{N(t), t \geq 0\}$ is called *the renewal process*, where $F(t) = \Pr\{X_k \leq t\}$ $(k = 1, 2, \cdots)$. Denote the time to occurrence of the nth failure by $S_n = X_1 + X_2 + \cdots + X_n$, where $S_0 = 0$ and $n = 1, 2, \cdots$. Since the number of failures up to time t (≥ 0) is $N(t) = \max\{n : S_n \leq t\}$, it is seen that $\Pr\{N(t) \geq n\} = \Pr\{S_n \leq t\}$. Hence, the probability that the number of failures up to time t is just n becomes

$$\begin{aligned} \Pr\{N(t) = n\} &= \Pr\{N(t) \geq n\} - \Pr\{N(t) \geq n+1\} \\ &= \Pr\{S_n \leq n\} - \Pr\{S_{n+1} \leq t\} \\ &= F^{(n)}(t) - F^{(n+1)}(t), \qquad n = 0, 1, \cdots. \end{aligned} \tag{1.7}$$

Then *the renewal function* $M(t)$ is defined as the expectation of the random variable $N(t)$ for fixed t. That is,

$$\begin{aligned} M(t) = \mathrm{E}[N(t)] &= \sum_{n=1}^{\infty} n\Pr\{N(t) = n\} \\ &= \sum_{k=1}^{\infty} \Pr\{N(t) \geq k\} = \sum_{k=1}^{\infty} \Pr\{S_k \leq t\} \\ &= \sum_{k=1}^{\infty} F^{(k)}(t) \end{aligned}$$

$$= F(t) + F * M(t) = \int_0^t m(x)dx, \tag{1.8}$$

where $m(t)$ is *the renewal density* and given by

$$\begin{aligned} m(t) &= \frac{dM(t)}{dt} = \sum_{k=1}^{\infty} f^{(k)}(t) \\ &= f(t) + \int_0^t m(t-x)f(x)dx \\ &= f(t) + \int_0^t f(t-x)m(x)dx. \end{aligned} \tag{1.9}$$

Equation (1.8) is called *the renewal equation* and is a special case of the Volterra integral equations.

The following theorems are fundamental but are the most important results to characterize the renewal function and its variation.

Theorem 1.2.1 (elementary renewal theorem)

$$\lim_{t\to\infty} \frac{M(t)}{t} = \frac{1}{\mathrm{E}[X]}, \tag{1.10}$$

$$\lim_{t\to\infty} m(t) = \frac{1}{\mathrm{E}[X]}. \tag{1.11}$$

Theorem 1.2.2 (Blackwell's theorem) For an arbitrary real number a (≥ 0),

$$\lim_{t\to\infty} \{M(t+a) - M(t)\} = \frac{a}{\mathrm{E}[X]}, \tag{1.12}$$

$$\lim_{t\to\infty} \left\{ M(t) - \frac{t}{\mathrm{E}[X]} \right\} = \frac{\mathrm{Var}[X]}{2\{\mathrm{E}[X]\}^2} - \frac{1}{2}. \tag{1.13}$$

Theorem 1.2.3 (key renewal theorem) Suppose that an arbitrary function $\phi(t)$ is differentiable with respect to $t \in (0, \infty)$ and is non-decreasing. Then,

$$\lim_{t\to\infty} \int_0^t \phi(t-x)dM(x) = \frac{\int_0^\infty \phi(x)dx}{\mathrm{E}[X]}. \tag{1.14}$$

1.2.2 Discrete renewal theory

Next, we summarize the discrete renewal theory. Let D ($D = 0, 1, 2, \cdots$) be the discrete non-negative random variable with period one, and denote the failure time or the number of failures, having probability mass function $f(d)$ ($d = 0, 1, 2, \cdots$, $f(0) = 0$). Then the failure time distribution is $F(d) = \sum_{j=0}^{d} f(j)$, where $F(0) = 0$ and $F(\infty) = 1$. The reliability function is $\overline{F}(d) = \sum_{j=d+1}^{\infty} f(j)$, where $\overline{F}(0) = 1$ and $\overline{F}(\infty) = 0$. Then the MTTF and the failure rate become $E[D] = \sum_{d=0}^{\infty} d f(d) = \sum_{d=0}^{\infty} \overline{F}(d)$ and $r(d) = f(d)/\overline{F}(d-1) = f(d)/\sum_{j=d}^{\infty} f(j)$, respectively. Using the failure rate $r(d)$, the reliability function and the probability mass function can be represented by $\overline{F}(d) = \prod_{j=0}^{d} \overline{r}(j)$ and $f(d) = r(d) \prod_{j=0}^{d-1} \overline{r}(j)$, respectively.

For two discrete failure time distributions $F(d)$ and $G(d)$ ($g(d) = G(d) - G(d-1), d = 0, 1, 2, \cdots$), define the convolution:

$$F * G(d) = \sum_{j=0}^{d} F(d-j) g(j) = \sum_{j=0}^{d} G(d-j) f(j). \tag{1.15}$$

By analogy with (1.3) and (1.4), the n-fold convolution $F^{(n)}(d)$ for identical distributions is recursively defined as

$$F^{(0)}(d) = 1, \tag{1.16}$$

$$F^{(n)}(d) = F * F^{(n-1)}(d), \qquad n = 1, 2, 3, \cdots. \tag{1.17}$$

Then, the corresponding n-fold convolution for identical probability mass functions is

$$f^{(0)}(d) = \begin{cases} 1, & d = 0 \\ 0, & d = 1, 2, 3, \cdots, \end{cases} \tag{1.18}$$

$$\begin{aligned} f^{(n)}(d) &= \sum_{j=0}^{d} f^{(n-1)}(d-j) f(j) \\ &= F^{(n)}(d) - F^{(n)}(d-1), \qquad d = 0, 1, 2, \cdots. \end{aligned} \tag{1.19}$$

Now, we are in a position to develop the discrete renewal theory. Let $N(d)$ ($d = 0, 1, 2, \cdots$) be the random variable representing the number of failures that have occurred during the time period $(0, d]$. Then the discrete renewal function $M(d) = \mathrm{E}[N(d)]$ is defined by

$$\begin{aligned} M(d) = \mathrm{E}[N(d)] &= \sum_{k=1}^{\infty} F^{(k)}(d) \\ &= F(d) + F * M(d) = \sum_{j=0}^{d} m(j), \end{aligned} \tag{1.20}$$

where $m(d)$ ($d = 0, 1, 2, \cdots$, and $m(0) = 0$) is *the renewal probability mass function* and means the probability that the failure occurs at time d, that is,

$$\begin{aligned} m(d) = M(d) - M(d-1) &= \sum_{k=1}^{\infty} f^{(k)}(d) \\ &= f(d) + \sum_{j=0}^{d} m(d-j) f(j) = f(d) + \sum_{j=0}^{d} f(d-j) m(j). \end{aligned} \tag{1.21}$$

Equation (1.20) is called *the discrete renewal equation* and is the analogy of (1.8). In general, the discrete renewal theory has one-to-one correspondence to the continuous version. For instance, we have the following theorems for the discrete renewal processes:

Theorem 1.2.4 (discrete elementary renewal theorem)

$$\lim_{d\to\infty} \frac{M(d)}{d} = \frac{1}{\mathrm{E}[D]}, \tag{1.22}$$

$$\lim_{d\to\infty} m(d) = \frac{1}{\mathrm{E}[D]}. \tag{1.23}$$

Theorem 1.2.5 (discrete Blackwell's theorem) For an arbitrary integer a (≥ 0),

$$\lim_{d\to\infty} \{M(d+a) - M(d)\} = \frac{a}{\mathrm{E}[D]}, \tag{1.24}$$

$$\lim_{d\to\infty} \left\{ M(d) - \frac{t}{\mathrm{E}[D]} \right\} = \frac{\mathrm{Var}[D]}{2\{\mathrm{E}[D]\}^2} - \frac{1}{2}. \tag{1.25}$$

Theorem 1.2.6 (discrete key renewal theorem) Suppose that an arbitrary function $\phi(d)$ can take the difference $\phi(d+1) - \phi(d)$ for $d = 1, 2, \cdots$ and is non-decreasing. Then,

$$\lim_{d\to\infty} \sum_{j=0}^{d} \phi(d-j)m(j) = \frac{\sum_{j=0}^{\infty} \phi(j)}{\mathrm{E}[D]}. \tag{1.26}$$

For more details on the discrete renewal theory, see Karlin and Taylor [14], Munter [21], Allan *et al.* [22] and Kaio and Osaki [23]. In the following sections, we direct our attention to calculating the renewal function for the continuous renewal process.

1.3 Some Useful Properties of the Renewal Function

1.3.1 Specific examples

In this section, we deal with some specific examples arising in reliability and maintenance theory. First, suppose that the inter-failure times $X_1, X_2, \cdots, X_n$ obey the exponential distributions with mean $1/\lambda$ (> 0),

i.e. $F(t) = 1 - \exp(-\lambda t)$. Then, the corresponding renewal process is reduced to the homogeneous Poisson process. That is, since the characteristic function of the partial sum $S_n = X_1 + X_2 + \cdots + X_n$ is

$$\mathrm{E}[e^{iux}] = \int_0^\infty e^{iux} dF(x) = \left(\frac{\lambda}{\lambda - iu}\right)^n \tag{1.27}$$

for $i = \sqrt{-1}$ and is the characteristic function of the gamma distribution, the n-fold convolution $F^{(n)}(t)$ becomes the gamma distribution. Thus, it is easy to obtain

$$\begin{aligned} M(t) &= \sum_{n=1}^{\infty} F^{(n)}(t) \\ &= \sum_{n=1}^{\infty} \int_0^t \frac{e^{-\lambda x}\lambda(\lambda x)^{n-1}}{(n-1)!} dx = \lambda t. \end{aligned} \tag{1.28}$$

Similarly, consider the case in which the inter-failure time distribution is the gamma distribution $F(t) = \int_0^t \exp(-\lambda x)\lambda^\alpha x^{\alpha-1} dx/\Gamma(\alpha)$, where $(\alpha, \lambda) \in (0, \infty) \times (0, \infty)$ are the shape and the scale parameters, respectively, and $\Gamma(\cdot)$ is the standard gamma function. Since the sum of two independent random variables having a gamma (α, λ) distribution and a gamma (β, λ) distribution becomes the gamma $(\alpha + \beta, \lambda)$ distribution, the n-fold convolution is

$$F^{(n)}(t) = \frac{1}{\Gamma(n\alpha)} \int_0^t e^{-\lambda x} \lambda^{n\alpha} x^{n\alpha - 1} dx \tag{1.29}$$

and is reduced to the gamma $(n\alpha, \lambda)$ distribution. For integral values of α, the renewal function can be obtained in closed form [7].

The third example is the case where the inter-failure time is the Weibull distribution $F(t) = 1 - \exp(-t^\alpha \rho^\alpha)$, where $(\alpha, \rho) \in (0, \infty) \times (0, \infty)$ are the shape parameter and the scale one, respectively. Unfortunately, the renewal function in this case cannot be obtained in closed form. Leadbetter [24] and Smith and Leadbetter [25] derived expansions of the renewal function in power series of t^α. Also, Lomnicki [26] proposed an alternative method to expand the Weibull renewal function into finite series of appropriate Poissonian functions of t^α. Soland [27], [28] developed a method involving numerical solution of the integral form. Jaquette [29] improved Soland's method by using asymptotic expansions of the dominating residues of the Laplace transform of the renewal function. Weiss [30] applied the Laguerre expansions to calculate the renewal process with the Weibull inter-failure time distribution. It should be noted that the Laguerre expansion method can be applied to the calculation with the general inter-failure time distribution. For more detail, see *e.g.* Weeks [31], Keilson and Nunn [32], Sumita and Kijima [33], [34] and Abate, Choudhury and Whitt [35]. Although the power series methods have been considered to function well for only the specific distribution cases, it has been found recently that these approaches may be useful to evaluate the complex stochastic models with general distributions. For example, see Blanc [36] and Gong and Hu [37].

1.3.2 Asymptotic properties

Suppose that the inter-failure time distribution $F(t)$ is non-arithmetric with finite second and third moments, $\mathrm{E}[X^2]$ and $\mathrm{E}[X^3]$. Then, from Theorem 1.2.2,

$$\lim_{t\to\infty}\left\{M(t)-\frac{t}{\mathrm{E}[X]}\right\}=\frac{\mathrm{E}[X^2]}{2\{\mathrm{E}[X]\}^2}-1. \tag{1.30}$$

Equation (1.30) tells us that the renewal function becomes, for larger $t \geq 0$,

$$M(t)=\frac{t}{\mathrm{E}[X]}+\frac{c_X^2-1}{2}+o(1), \qquad t\to\infty, \tag{1.31}$$

where $c_X = \sqrt{\mathrm{Var}[X]}/\mathrm{E}[X]$ is the coefficient of variation of the random variable X. Hence, the renewal function approximates to a straight line asymptotically. Leadbetter [38] evaluated the error bounds for the above linear approximation to the renewal function. Sahin [39] provided some conditions for the accuracy of the approximation $M(t) \approx 1/\mathrm{E}[X] + (c_X^2 - 1)/2$ within the context of a class of distributions, and examined how large is enough for a asymptotic approximation to be sufficiently accurate in terms of time scale t. Further, higher order asymptotic approximation has been studied by Carlsson [40]. He obtained

$$\begin{aligned} M(t)=&\frac{t}{\mathrm{E}[X]}+\frac{c_X^2-1}{2}-\frac{\int_t^\infty Q(x)dx}{\{\mathrm{E}[X]\}^2}+\frac{Q*Q(t)}{\{\mathrm{E}[X]\}^3}\\ &+o(t^m\log t), \quad t\to\infty, \end{aligned} \tag{1.32}$$

where $Q(t) = \int_t^\infty \overline{F}(x)dx$, if $F(t)$ is strongly non-lattice with finite moments of integer order m (≥ 2).

1.4 Analytical Approximation Methods

In this section, we introduce some analytical methods to approximate the renewal function. The main advantages of these methods are the analytically easy treatment and the reduction of computation time of the renewal function. If any probabilistic quantity depending on the renewal function needs to be evaluated, then the analytical approaches may be useful. Also, the basic mathematical ideas introduced in this section are quite interesting for developing the approximation formulae for some applications.

1.4.1 Phase renewal processes

The most intuitive argument to approximate the renewal function is to approximate the underlying renewal process by any other stochastic process. First, Neuts [41], [42] considered the renewal processes of phase type. Kao [43] described the computation method of the renewal function applying the phase-type renewal processes. Consider a continuous-time Markov chain with state space $\{1, 2, \cdots, m+1\}$, where each state $k = 1, 2, \cdots, m$ denotes the transient state and $m+1$ is the absorbing one. Then, the infinitesimal generator $\mathbf{Q}$ for the underlying Markov chain is

$$\mathbf{Q}=\begin{bmatrix}\mathbf{T} & \mathbf{T^0}\\ 0 & 0\end{bmatrix}, \tag{1.33}$$

where $\mathbf{T}$ is a nonsingular $m \times m$ matrix having the elements $T_{ij} < 0$ for $1 \leq i \leq m$ and $T_{ij} \geq 0$ for $i \neq j$, and $\mathbf{T^0}$ is an m vector satisfying $\mathbf{T^0} \geq 0$ and $\mathbf{Te} + \mathbf{T^0} = 0$, where $\mathbf{e} = (1, \cdots, 1)$. Define the vector of initial probabilities (α, α_{m+1}), where α is an m vector such that $0 < \alpha \leq 1$. The phase-type distribution is the distribution of the time until absorption in

state $m+1$ given the initial probability vector (α, α_{m+1}), and is defined by

$$F_p(t) = 1 - \alpha \exp(\mathbf{T}t)\mathbf{e}. \tag{1.34}$$

If the inter-failure time distribution is the phase-type distribution, the corresponding stochastic counting process is said *the phase-type renewal process.*

To evaluate the renewal function for the phase-type renewal process, consider the continuous-time Markov chain with state space $\{1, \cdots, m\}$ and infinitesimal generator $\mathbf{Q}^* = \mathbf{T} + \mathbf{T}^0\mathbf{A}^0$, where $\mathbf{T}^0$ is an $m \times m$ matrix with identical columns $\mathbf{T}^0$ and $\mathbf{A}^0 = \mathrm{diag}(\alpha_1, \cdots, \alpha_m)$. Suppose that $\mathbf{Q}^*$ is irreducible with stationary probability vector π and that the stationary probability matrix Π has identical rows π. Define the stochastic process denoting the state of this process by $\{J(t), t \geq 0\}$, where $\nu(t)$ is the probability vector with the element $\nu_j(t) = \Pr\{J(t) = j\}$. When the mean inter-failure time (asymptotic renewal rate) $\mathrm{E}[X]$ is given, Neuts [41], [42] derived the following formula:

$$\begin{aligned} M(t) &= \frac{1}{\mathrm{E}[X]}\Big\{t + \alpha[\mathbf{I} - \exp(\mathbf{Q}^*t)]\mathbf{T}^{-1}\mathbf{e}\Big\} \\ &= \frac{1}{\mathrm{E}[X]}\Big\{t - \nu(t)\mathbf{T}^{-1}\mathbf{e} - 1\Big\}. \end{aligned} \tag{1.35}$$

Since the probability vector $\nu(t)$ is the solution of the following differential equation:

$$\frac{d\nu(t)}{dt} = \nu(t)\mathbf{Q}^*, \tag{1.36}$$

with initial condition $\nu(0) = \alpha$, we can calculate it by some numerical integration methods such as the method of Runge-Kutta. Alternatively, Kao [43] applied the well-known randomization methods [44] for finding the time-dependent state probability vector $\nu(t)$.

Of course, the above result is not sufficient to approximate the renewal function with an arbitrary inter-failure distribution function, since the knowledge of the underlying distribution function is not reflected sufficiently. In other words, based on the knowledge on several moments of the inter-failure time, we have to specify completely the approximating renewal process by fitting a convenient distribution to those moments. For this problem, Whitt [45] proposed some methods on the moments matching. The framework to approximate the general distribution by the phase-type distributions was made by Altiok [46]. Also, Van der Heijden [47] discussed a simple three-moment approximation method with a Coxian distribution to approximate the general distribution function. Taking account of the information on finite moments, the phase-type renewal process can approximate a renewal process with an arbitrary accuracy. Recently, Asmussen *et al.* [48] proposed an interesting method based on the EM (expectation-maximization) algorithm. However, the accuracy strongly depends on the number of states for the underlying Markov chain. Though the increasing number of phases leads to the increasing approximation performance, a great deal of computation effort will be required.

1.4.2 Gamma approximations

Tijms [49] presented the gamma approximation for the renewal function, where the underlying inter-failure time distribution function $F(t)$ is replaced by more tractable distribution function having the same first two

moments as $F(t)$. More specifically, define the following approximation scheme:

$$\begin{aligned} M(t) &\equiv \sum_{n=1}^{\infty} F^{(n)}(t) \\ &\approx F(t) + \sum_{n=2}^{\infty} F_G^{(n)}(t) \\ &\approx F(t) + F^{(2)}(t) + \sum_{n=3}^{\infty} F_G^{(n)}(t) \\ &\cdots \\ &\approx \sum_{n=1}^{k} F^{(k)}(t) + \sum_{n=k+1}^{\infty} F_G^{(k)}(t), \end{aligned} \tag{1.37}$$

where $F_G(\cdot)$ is called *the dummy distribution*. Since the accurate approximation depends on the exact computation of the first few terms in the infinite series, it is important to select a suitable dummy distribution $F_G(\cdot)$. From the tractability, Tijms [49] recommended the use of the gamma distribution as a dummy distribution, that is,

$$F_G(t) = \frac{1}{\Gamma(a)} \int_0^t \lambda^a x^{a-1} e^{-\lambda x} dx \tag{1.38}$$

for λ (> 0) and a (> 0). Consequently, from (1.29), $F_G^{(n)}(t)$ is obtained in closed form. The parameters λ and a should be selected by fitting on the first two moments of $F(t)$ and should be the solutions of the following simultaneous equations:

$$\mathrm{E}[X] = \frac{na}{\lambda} \quad \text{and} \quad \mathrm{Var}[X] = \frac{na}{\lambda^2}, \tag{1.39}$$

if they exist. In the computations above, since the infinite series has to be truncated, this can be established using the stopping criterion that the evaluation of the terms in the series is stopped at the first k as soon as $F_G^{(k)}(t)F_G(t)/[1 - F_G(t)] \leq \epsilon$ for some pre-specified $\epsilon > 0$ (*e.g.* $\epsilon = 10^{-5}$).

A similar but somewhat different method from the above was proposed by Smeitink and Dekker [50]. They used the Coxian distribution and its variation instead of the gamma distribution. Since the choice of the dummy distribution depends strongly on the squared coefficient of variation c_X^2, they used the Coxian-2 distribution on the first two moments for $0.5 \leq c_X^2$. From the discussion in 1.4.1, define the Coxian-2 distribution with parameters λ_1 (> 0), λ_2 (> 0) and p ($0 \leq p \leq 1$). Then, the probability density function becomes

$$f_G(t) = \begin{cases} p\lambda e^{-\lambda t} + (1-p)\lambda^2 t e^{-\lambda t}, & \text{if } \lambda_1 = \lambda_2 = \lambda \\ \lambda_1\left(\frac{p\lambda_1 - \lambda_2}{\lambda_1 - \lambda_2}\right)e^{-\lambda_1 t} & \text{if } \lambda_1 \neq \lambda_2. \\ \quad +\lambda_2\left(1 - \frac{p\lambda_1 - \lambda_2}{\lambda_1 - \lambda_2}\right)e^{-\lambda_2 t}, \end{cases} \tag{1.40}$$

The renewal function for the Coxian-2 distribution has the following explicit form:

$$\begin{aligned} M_G(T) &= \sum_{n=1}^{\infty} F_G^{(n)}(t) \\ &= \frac{\lambda_1\lambda_2 t}{\lambda_1(1-p) + \lambda_2} - \frac{\lambda_1(1-p)(\lambda_2 - p\lambda_1)}{[\lambda_1(1-p) + \lambda_2]^2} \end{aligned}$$

$$\times\left\{1-e^{-[\lambda_1(1-p)+\lambda_2]t}\right\}. \tag{1.41}$$

On the other hand, for $0 < c_X^2 \leq 1$, the so-called $E_{k-1,k}$ distribution can fit the underlying inter-failure time distribution with density function

$$f_G(t) = p\lambda^{k-1}\frac{t^{k-2}e^{-\lambda t}}{(k-2)!} + (1-p)\lambda^k\frac{t^{k-1}e^{-\lambda t}}{(k-1)!}, \tag{1.42}$$

where $p \in [0,1]$, $\lambda > 0$ and $k \geq 2$. Note that the $E_{k-1,k}$ distribution is a special case of the Coxian-2 distribution for $0.5 \leq c_X^2 \leq 1$. In [50], the n-fold convolution of an $E_{k-1,k}$ distribution is given in a closed form, that is,

$$F_G^{(n)}(t) = \sum_{j=0}^{n}\binom{n}{j}(1-p)^j p^{n-j}\left\{1-\sum_{l=0}^{n(k-1)+j-1}\frac{(\lambda t)^l e^{-\lambda t}}{l!}\right\}. \tag{1.43}$$

Substituting these dummy distributions into (1.37) yields the approximation form of the renewal function.

An alternative method called the gamma approximation was proposed by Ross [51] independently. Let $Y_1, \cdots, Y_n$ be independent exponentials, each having rate λ (> 0). Suppose that Y_i ($i = 1, 2, \cdots, n$) is also independent of the renewal process $\{N(t), t \geq 0\}$ having inter-failure time distribution $F(t)$. Define $\psi_{r,\lambda} = \mathrm{E}[N(Y_1 + \cdots + Y_r)]$ for an arbitrary r ($= 1, 2, \cdots, n$). Then, Ross [51] proved the following recursive formula:

$$\psi_{r,\lambda} = \frac{\sum_{j=1}^{r}(1+\psi_{r-j,\lambda})\mathrm{E}[X_i^j e^{-\lambda X_i}]\lambda^j/j! + \mathrm{E}[e^{-\lambda X_i}]}{1-\mathrm{E}[e^{-\lambda X_i}]}. \tag{1.44}$$

Since $\psi_{r,\lambda}$ can be recursively solved, then the approximation of the renewal function will be obtained by setting $\lambda = n/t$ in the above formula, *i.e.* $M(t) \approx \psi_{r,n/t}$. For instance, Ross [51] derived the following approximation formulae with three orders:

$$\psi_{1,1/t} = \frac{\mathrm{E}[e^{-X_i/t}]}{1-\mathrm{E}[e^{-X_i/t}]}, \tag{1.45}$$

$$\psi_{2,2/t} = \frac{2\mathrm{E}[X_i e^{-2X_i/t}]}{t\{1-\mathrm{E}[e^{-2X_i/t}]\}^2} + \frac{\mathrm{E}[e^{-2X_i/t}]}{1-\mathrm{E}[e^{-2X_i/t}]}, \tag{1.46}$$

$$\psi_{3,3/t} = \frac{9\mathrm{E}[X_i e^{-3X_i/t}]^2}{t^2\{1-\mathrm{E}[e^{-3X_i/t}]\}^3} + \frac{6t\mathrm{E}[X_i e^{-3X_i/t}] + 9\mathrm{E}[X_i^2 e^{-3X_i/t}]}{2t^2\{1-\mathrm{E}[e^{-3X_i/t}]\}^2} + \frac{\mathrm{E}[e^{-3X_i/t}]}{1-\mathrm{E}[e^{-3X_i/t}]}. \tag{1.47}$$

In general, for an arbitrary r (> 0), it holds that

$$\psi_{r,n/t} = \frac{\sum_{j=1}^{r-1}(1+\psi_{r-1,n/t})\alpha_j(n/t)^j/j! + \alpha_0}{1-\alpha_0}, \tag{1.48}$$

where $\alpha_j = \mathrm{E}[X_i^j e^{-nX_i/t}]$. This means that the renewal function $M(t)$ can be approximated by $\int M(t)dF_G(t)$, where

$$F_G(t) = \frac{1}{\Gamma(n)}\int_0^t \lambda^n x^{n-1}e^{-\lambda x}dx, \tag{1.49}$$

when $r = n$.

In this approximation scheme, it is known that the nth approximation $\psi_{n,n/t}$ converges to $M(t)$ as n goes to infinity, if $M(t)$ is a continuous function of t. More precisely, Angus and Hong [52] proved that the rate of convergence of $\psi_{n,n/t}$ to $M(t)$ is no worse than $O(1/\sqrt{n})$ when $M(t)$ satisfies a local Lipschitz condition at t. Also, Ross [51] showed that the approximation $\psi_{n,n/t}$ is an increasing function of n, each of which is less than $M(t)$, if the inter-failure time distribution is DFR (decreasing failure rate). Similarly, the mean residual life and the mean excess life can also be approximated with this gamma approximation.

1.4.3 Methods based on equilibrium distribution

The methods belonging to this class provide the most traditional approximation schemes. The particular merit of the methods introduced here is that they combine simplicity and accuracy. Of particular interest to us is the derivation procedure of the approximation formulae for the renewal function without specifying the inter-failure time distribution. Actually, the approximations based on the gamma and the Coxian distributions have been developed in terms of tractability. On the other hand, it is known that many approximation formulae in mathematical analysis were made by ignoring unnecessary terms from a complex mathematical expression. Here, we focus on the direct (and analytical) approximation of the renewal equation in (1.8).

Define the equilibrium distribution for the inter-failure time distribution $F(t)$ by

$$F_e(t) = \frac{1}{\mathrm{E}[X]} \int_0^t \overline{F}(\tau) d\tau. \tag{1.50}$$

From the integral equation for the renewal density in (1.9), it is straightforward to derive

$$\overline{F}(t) + \int_0^t m(t-\tau)\overline{F}(\tau) d\tau = 1. \tag{1.51}$$

Since (1.51) simply expresses the fact that the integral of this density between 0 and t has to be unity, we have $F(t)/\int_0^t m(t-\tau)\overline{F}(\tau)d\tau = 1$, and in consequence,

$$m(t) = f(t) + \frac{F(t)\displaystyle\int_0^t m(t-\tau) dF(\tau)}{\displaystyle\int_0^t m(t-\tau)\overline{F}(\tau) d\tau}. \tag{1.52}$$

Bartholomew [53] obtained an approximation form by replacing the second term of the right-hand side in (1.52) by $\int_0^t dF(t)/\int_0^t \overline{F}(\tau)d\tau$. That is,

$$\begin{aligned} m(t) &\approx m_B(t) \\ &= f(t) + \frac{F^2(t)}{F_e(t)\mathrm{E}[X]}, \end{aligned} \tag{1.53}$$

$$\begin{aligned} M(t) &\approx M_B(t) \\ &= F(t) + \frac{1}{\mathrm{E}[X]} \int_0^t \frac{F^2(\tau)}{F_e(\tau)} d\tau. \end{aligned} \tag{1.54}$$

Bartholomew [53] also discussed the approximation of the discrete renewal function along the similar way. Further, he examined the usefulness for the above approximation formula by comparing the most classic methods by Feller [12] and Weiss [30].

Ozbaykal [54] proposed a different approximation form $M_O(t)$ of the renewal function. To explain the derivation procedure for $M_O(t)$, let us start from an alternative expression of the renewal equation (see *e.g.* Karlin and Taylor [14]) as follows.

$$M(t) = 1/\mathrm{E}[X] - F_e(t) + \int_0^t \overline{F}_e(t-\tau)dM(\tau). \tag{1.55}$$

This can be verified by taking the Laplace transform of both sides of the equation in (1.8). Since $M(t)$ satisfies the equation

$$F(t) = \int_0^t m(t-\tau)\overline{F}(\tau)d\tau \tag{1.56}$$

if it also satisfies (1.8), combining (1.9), (1.55) and (1.56), we obtain

$$M(t) = \frac{t}{\mathrm{E}[X]} - F_e(t) + \int_0^t \overline{F}_e(t-\tau) \times \left[\frac{F(\tau)\displaystyle\int_0^\tau m(\tau-t)dF(t)}{\displaystyle\int_0^\tau m(\tau-t)\overline{F}(t)dt} + f(\tau) \right] d\tau. \tag{1.57}$$

If the approximation should be tractable, it is troublesome to evaluate the term within the bracket in (1.57). The idea proposed by Ozbaykal [54] is to drop it from (1.57). Hence, we have

$$\begin{aligned} M(t) &\approx M_O(t) \\ &= \frac{t}{\mathrm{E}[X]} - F_e(t) + \int_0^t \overline{F}_e(t-\tau)d\tau. \end{aligned} \tag{1.58}$$

Deligönul [55] developed a similar approximation formula to Ozbaykal's [54]. By dropping $m(\tau - t)$'s from (1.57), we get the following improved approximation scheme:

$$\begin{aligned} M(t) &\approx M_D(t) \\ &= \frac{t}{\mathrm{E}[X]} - F_e(t) + \int_0^t \overline{F}_e(t-\tau)\left\{ f(\tau) + \frac{F^2(\tau)}{\mu F_e(\tau)} \right\} d\tau. \end{aligned} \tag{1.59}$$

It can be easily verified that the Deligönul approximation is exact if $F(t)$ is the exponential distribution. Although we do not prove it here, the error function $|M(t) - M_D(t)|$ is less than $|M(t) - M_B(t)|$ for all $t \geq 0$, if $F(t)$ is DFR. Also, it is shown in the numerical comparison with the mixed exponential and the gamma distributions that this method outperforms both $M_B(t)$ and $M_O(t)$ [55].

1.5 Bounds

The upper and lower bounds may be useful to evaluate the renewal function if knowledge of the distribution is incomplete and only the information on a few moments is available, since the asymptotic properties for the renewal function are well known. The most famous bounds for the renewal function

were derived by Barlow and Proschan [7], [56] and Barlow [57]. From the asymptotic property in 1.3.2, it is expected that $t/\mathrm{E}[X] - a_l \leq M(t) < t/\mathrm{E}[X] - a_u$, where a_l and a_u ($a_l > a_u$) are constants. It is known that $M(t) \geq t/\mathrm{E}[X] - 1$ when $a_l = 1$ [7]. More specifically, if one restricts the class of distribution functions, a pair of bounds can be derived. If $F(t)$ is NBUE (New Better Than Used in Expectation), then

$$\frac{t}{\mathrm{E}[X]} - 1 \leq M(t) \leq \frac{t}{\mathrm{E}[X]}. \tag{1.60}$$

On the other hand, if $F(t)$ is IFR (Increasing Failure Rate), tighter lower and upper bounds are available. That is,

$$\frac{t}{\int_0^\infty \overline{F}(x)dx} \leq M(t) \leq \frac{tF(t)}{\int_0^\infty \overline{F}(x)dx}. \tag{1.61}$$

Since $t/\mathrm{E}[X] - 1 \leq t/\int_0^\infty \overline{F}(x)dx$ and $tF(t)/\int_0^\infty \overline{F}(x)dx \leq t/\mathrm{E}[X]$, the IFR case can provide tighter bounds of the renewal function [7].

Let us return the general form $t/\mathrm{E}[X] - a_l \leq M(t) < t/\mathrm{E}[X] - a_u$. If we substitute this into the renewal equation in (1.8), then we obtain $M(t) \geq t/\mathrm{E}[X] - F_e(t)$. Since $F_e(t) \leq 1$, after n iterations with the renewal function, we have the following tight lower bound:

$$M(t) \geq \frac{t}{\mathrm{E}[X]} + \sum_{k=1}^{n} F^{(k)}(t) - \sum_{k=1}^{n} F_e * F^{(k-1)}(t) - F^{(n)}(t) \tag{1.62}$$

for all $t \geq 0$. From $t/\mathrm{E}[X] = \sum_{k=1}^{\infty} F_e * F^{(k-1)}(t)$, $\lim_{n\to\infty} F^{(n)}(t) = 0$ and (1.8), the sequence of lower bounds defined by (1.62) is monotone non-decreasing in n for any fixed t and converges to the real value $M(t)$. This fact is quite interesting and is desirable for developing other bounds of the renewal function.

The linear bounds of Marshall [58] are recognized to the best linear bounds of the renewal function. Define the linear function $I_0(a,t) = t/\mathrm{E}[X] + a$. Substituting this into $M(t)$ in the renewal equation yields

$$I_1(a,t) = \frac{t}{\mathrm{E}[X]} + a - a\overline{F}(t) - F_e(t) + F(t). \tag{1.63}$$

By successive iterations, we have

$$I_n(a,t) = \frac{t}{\mathrm{E}[X]} + a - a\overline{F}^{(n)}(t) - \sum_{k=1}^{n} F_e * F^{(k-1)}(t) + \sum_{k=1}^{n} F^{(k)}(t). \tag{1.64}$$

Now, define

$$a_l = \inf_{t\geq 0} \frac{F(t) - F_e(t)}{\overline{F}(t)} \quad \text{and} \quad a_u = \sup_{t\geq 0} \frac{F(t) - F_e(t)}{\overline{F}(t)}. \tag{1.65}$$

Then, Marshall's bounds are given by

$$I_n(a_l,t) \leq M(t) \leq I_n(a_u,t), \tag{1.66}$$

where $\lim_{n\to\infty} I_n(a,t) = M(t)$ for any real a. If we focus on the linear bounds, then

$$\frac{t}{\mathrm{E}[X]} + a_l \leq M(t) \leq \frac{t}{\mathrm{E}[X]} + a_u. \tag{1.67}$$

Also, bounds for the renewal function can be characterized in terms of the aging property for the inter-failure time distribution. Consider the

renewal processes with DFR and IMRL (Increasing Mean Residual Life) inter-failure times. If $F(t)$ is DFR, then the renewal function $M(t)$ is concave and the renewal density $m(t)$ is decreasing [7]. {Although this result was extended such that if $F(t)$ is NBU (NWU), then the renewal function $M(t)$ is convex (concave), where NBU (NWU) is New Better Than Used (New Worse Than Used) [59]} . Based on these results, Brown [60] derived some interesting inequalities on the renewal function. Define the ith moment $\mu_i = \mathrm{E}[X^i] = \int_0^\infty t^i dF(t)$, where $\mu_1 = \mathrm{E}[X]$. Brown [60] proved that if $F(t)$ is IMRL, then, for an arbitrary k (≥ 0),

$$U(t) - \min_{0 \leq i \leq k} c_i t^{-i} \leq M(t) \leq U(t), \tag{1.68}$$

where $U(t) = t/\mu_1 + \mu_2/(2\mu_1^2)$ and c_i is the solution of

$$c_i = \gamma_i - \frac{i! \sum_{j=1}^{i-1} (c_j/j!)(\mu_{i+1-j}/(i+1-j)!)}{\mu_1}, \tag{1.69}$$

$$\gamma_i = \frac{\mu_{i+2}}{(i+1)(i+2)\mu_1^2} - \frac{\mu_2 \mu_{i+1}}{2(i+1)\mu_1^3}. \tag{1.70}$$

Further, Brown [60] obtained

$$U(t) - V(t) \leq M(t) \leq U(t), \tag{1.71}$$

where

$$V(t) = \frac{(\mu_1 a)^{-1} - (\mu_2/2\mu_1^2) - \phi_F(a) - 1)^{-1}}{e^{at} - 1}, \tag{1.72}$$

a is the real number satisfying $(0 < a < a_0)$ and $\phi_F(a_0) = \int_0^\infty \exp(a_0 t) dF(t)$ $< \infty$ for an arbitrary $a_0 > 0$. Since the recursive equation in (1.69) can be solved explicitly for an arbitrary $k > 0$, we can obtain an arbitrary number of lower bounds with IMRL inter-failure time. Further, Brown [60] showed that the above results can be improved for DFR inter-failure time distributions, although we omit to show them. For further monotonicity properties for the renewal process, see [61]. Also, the other upper bounds for the renewal function were found independently by Stone [62] and Daley [63].

1.6 Numerical Methods

The most well-known methods to compute the renewal function are the numerical algorithms introduced in this section. As an advantage of the numerical methods, they can guarantee great accuracy in calculation of the renewal function on computer. As described in Section 1.1, the recent development of computer technology enables us to carry out rather complex numerical calculations within a satisfactory computation time. Hence, if only the numerical value of renewal function at a fixed time is needed, the numerical methods will be superior to the analytical ones in terms of computation accuracy.

1.6.1 Laplace inversion technique

In principle, the most popular method to solve the Voltera integral equation in (1.8) numerically is the Laplace-Stieltjes (LS) inversion transform. Define the LST (Laplace-Stieltjes transform) of the inter-failure time distribution by $F^*(s) = \int_0^\infty \exp(-st)dF(t)$. Then, from a simple manipulation, the LST of the renewal function is

$$M^*(s) = \int_0^\infty \exp(-st)dM(t) = \sum_{n=1}^{\infty} \{F^*(s)\}^n = \frac{F^*(s)}{1 - F^*(s)}. \tag{1.73}$$

Of course, (1.73) means the Laplace transform of the renewal density. Noting that the underlying function $M(t)$ has one-to-one correspondence with $M^*(s)$, the renewal function can be specified completely by $M^*(s)$. The inversion formula for obtaining the renewal density $m(t) = dM(t)/dt$ from $M^*(s)$ is given by

$$m(t) = \lim_{c\to\infty} \frac{1}{2i\pi} \int_{b-ic}^{b+ic} e^{st} M^*(s)ds, \tag{1.74}$$

where $i = \sqrt{-1}$ is an imaginary unit, $b > \max\{\sigma, 0\}$ and σ is a radius of convergence. The inversion formula in (1.74) is known as the Bromwich inversion integral, but may be hard to use directly. A natural way to evaluate the above equation is to apply the Fourier series method. More precisely, since, after some manipulations, we have

$$\begin{aligned} m(t) &= \frac{2e^{bt}}{\pi} \int_0^\infty \mathrm{Re}\Big(M^*(b+iu)\Big) \cos(xt)dx \\ &= \frac{-2e^{bt}}{\pi} \int_0^\infty \mathrm{Im}\Big(M^*(b+iu)\Big) \sin(xt)dx, \end{aligned} \tag{1.75}$$

the Bromwich inversion integral can be calculated by performing a numerical integration. Applying the well-known trapezoidal formula with step size h (> 0), the renewal density can be represented as

$$m(t) \approx \frac{fe^{bt}}{\pi} \mathrm{Re}\Big(M^*(b)\Big) + \frac{2he^{bt}}{\pi} \sum_{k=1}^{\infty} \mathrm{Re}\Big(M^*(b+ikh)\Big) \cos(kht). \tag{1.76}$$

Of course, when the above method is used, it is important to evaluate the discritization error, the truncation error and the roundoff error.

To the best of our knowledge, one of the most famous inversion methods in applications is *the Jagerman's method* [64], [65]. This method is also called *the Post-Widder inversion formula* [66]. For the renewal density $m(t)$ under regularity conditions,

$$m(t) = \lim_{n\to\infty} \frac{(-1)^n}{n!} \Big(\frac{n+1}{t}\Big)^{n+1} \frac{d^n}{ds^n} M^*(s) \mid_{s=(n+1)/t}. \tag{1.77}$$

In the literature, many other inversion techniques have been developed, since this problem plays a significant part in the context of numerical calculation algorithms. For the recent results on the Laplace inversion technique and its variation, see Davies and Martin [67], Abate *et al.* [68] and the references therein.

1.6.2 Cubic spline algorithm

The spline function is the piecewise continuous polynomial function to approximate an arbitrary function so as to satisfy some conditions for continuity. Define the knots of the spline function $\xi_1, \xi_2, \cdots, \xi_n$ $(n > 1)$. The spline function $z(x)$ with knots ξ_j $(j = 1, 2, \cdots, n)$ is called to have the degree m, if

(i) the function $z(x)$ in the interval $[\xi_j, \xi_{j+1}]$ $(j = 0, 1, \cdots, n-1)$ is a polynomial function with less than degree m and

(ii) both the function $z(x)$ and its $(m-1)$th derivative are absolutely continuous in $(-\infty, \infty)$.

The spline function with the above properties is known to approximate an arbitrary function better than the other polynomial functions with the same degree. In addition, if the spline function with the degree $2k-1$ $(k = 1, 2, \cdots)$ can be expressed by a polynomial with degree $k-1$ in two boundary ranges $(-\infty, \xi_1]$ and $[\xi_n, \infty)$, then it is called *the natural spline function*. Usually, the natural cubic spline function is used in the functional approximation problem in terms of the accuracy and the computation effort.

Cléroux and Mcconalogue [69] and Mcconalogue [70] proposed the cubic spline algorithm to calculate the k-fold convolution $F^{(k)}(t)$ for the inter-failure time distribution $F(t)$. Divide the time interval $[0, t]$ by n segments with equal length d (> 0). At each point jd $(j = 1, \cdots, n)$, define a sequence $(0, F^{(k)}(0))$, $(d, F^{(k)}(d))$, $\cdots$, $(nd, F^{(k)}(nd))$, where it is assumed that k-fold convolution of the inter-failure time distribution $F^{(k)}(t)$ $(0 < k < n)$ can be recursively evaluated. Next, interpolate a discrete points $(jd, F^{(k)}(jd)) = (jd, F_j^{(k)})$ in the range $[jd, jd+d]$ by the following piecewise polynomial:

$$z_j(t) = \frac{(jd+d-t)^3}{6d} c_j + \frac{(t-jd)^3}{6d} c_{j+1} + A_j(jd+d-t) + A_{j+1}(t-jd), \qquad jd \le t \le jd+d, \tag{1.78}$$

where

$$A_j = \frac{F_j^{(k)}}{d} - \frac{d}{6} c_j, \qquad j = 0, 1, \cdots, n-1 \tag{1.79}$$

and c_j $(j = 0, 1, \cdots, n)$, called *the spline coefficient*, satisfies

$$\frac{1}{4} c_{j-1} + c_j + \frac{1}{4} c_{j+1} = \frac{3}{2d^2} \left\{ F_{j-1}^{(k)} - 2F_j^{(k)} + F_{(j+1)}^{(k)} \right\}, \tag{1.80}$$

$$c_0 + \frac{1}{2} c_1 = \frac{3}{d} \left\{ \frac{F_1^{(k)} - F_0^{(k)}}{d} - F'^{(k)}_0 \right\}, \tag{1.81}$$

$$c_{n-1} = \frac{F_n^{(k)} - 2F_{n-1}^{(k)} + F_{n-2}^{(k)}}{d^2}. \tag{1.82}$$

In (1.81), $F'^{(k)}_0$ is defined by

$$F'^{(k)}_0 = \begin{cases} f(0), & k = 1 \\ 0, & k > 1, \end{cases} \tag{1.83}$$

if the density $f(t)$ exists. The above simultaneous equations can be solved numerically.

We evaluate the convolution expression by the following discritized one:

$$F_0^{(k+1)} = 0, \tag{1.84}$$

$$F_i^{(k+1)} = \sum_{j=0}^{i-1} \int_{jd}^{jd+d} z_j f(id-t)dt, \qquad i = 1, 2, \cdots, n. \tag{1.85}$$

Since the values of the convoluted distribution function at each point $(0, F_0^{(k)})$, $(d, F_1^{(k)})$, $(2d, F_2^{(k)})$, $\cdots$, $(nd, F_n^{(k)})$ are available recursively, it is possible to compute the renewal function numerically by applying any numerical integral method for the obtained interpolation function $z_j(t)$. Cléroux and Mcconalogue [69] recommended using the five point-robatto formula as a suitable numerical integral method, rather than the usual trapezoidal formula. For an arbitrary continuous function $g(t)$, the five point-robatto formula is given by

$$\int_{jd}^{jd+d} g(t)dt = \frac{d}{180}\Big\{9[g(jd)+g(jd+d)] + 49\Big[g(jd+d_1) + g(jd+d_2)\Big] + 64g(jd+\frac{d}{2})\Big\}, \quad 0 \le j \le n-1 \tag{1.86}$$

where $d_1 = d(1-\sqrt{3/7})/2$ and $d_2 = d(1+\sqrt{3/7})/2$.

Note that success of the cubic spline algorithm depends on the choice of knots. Even though the number of knots is fixed, the position of the knots will influence the shape of the interpolated function. Actually, we often face the problem that the approximated function gives rise to extreme fluctuations for a given position of the knots. Consequently, although the problem of determining the optimal position of knots is essentially important for the function approximation, this is not so easy to solve, since the underlying problem is reduced to the non-linear optimization problem for a multi-modal function with multiple variables. Latterly, Mcconalogue [71] and Baxter [72] have improved the spline algorithm for distributions with densities having singularities at the origin. Since this method can easily be extended for the variance function and the integral of the renewal function, it has been used for a long time as the most effective numerical method to calculate the renewal function. Baxter *et al.* [73] compared various tabulations on the computation results. Further, Deligönül and Bilgen [74] applied the Galerkin technique to the cubic spline algorithm and characterized the solution of the Volterra equation of renewal theory.

1.6.3 Discritization algorithm

It will be intuitive to calculate the renewal function by solving the renewal equation numerically, although the previous numerical methods focused on the convolution expression of the inter-failure time distribution. Xie [75] derived a discritization algorithm or a direct Rieman-Stieltzes integration method, for the renewal equation. Divide the time interval $[0, t]$ by n (> 0) line segments with equal length d (> 0). Define the values of the renewal function and the inter-failure time distribution function at each point $i = 0, d, 2d, \cdots, nd = t$ by $M_i = M(id)$ and $K_i = F(id)$ $(i = 0, 1, \cdots, n)$, respectively, where $F_i = F((i-1/2)d)$. Then, the following recursive formula gives the approximated value of the renewal function M_i:

$$M_i = \frac{K_i + \sum_{j=1}^{i-1}(M_j - M_{j-1})F_{i-j+1} - M_{i-1}F_1}{1-F_1}, \tag{1.87}$$

where $M(0) = 0$ and $i = 1, \cdots, n$. It is not hard to calculate M_i recursively on computer. Also, it is a surprising fact that the above simple method has never been discussed before the paper by Xie [75]. In fact, this method has a nice convergence property, and is applicable to the discrete renewal function. Note in this method that the accuracy depends heavily on the choice of the grid size d. In other words, the suitable grid size d is related to the shape of the underlying distribution $F(t)$ and the time scale t. Though Tijms [49] recommends that d should be selected in the range of $0.05 \sim 0.01$, in general, it is necessary to adjust the grid size d by the trial and error method in advance. Also, it is known that this method results in smaller round errors in the discritized approximation for a sufficiently larger time scale t. Conversely, for a smaller time scale, this method does not always function well. Since the discritization method is based on the simple iteration scheme, it can be extended from a variety of standpoints. For instance, see Boehme *et al.* [76] and Banjevic [77]. Recently, Ayhan, Limón-Robles and Wortman [78] proposed an algorithm based on discritization to compute the bounds of renewal function.

1.6.4 Approximation by rational functions

Though an approximation based on rational functions is the most classical approach, such an attempt has not been made for a long time. Chaudhry [79] proposed an interesting computation method with rational functions. Suppose that the LST of the inter-failure distribution function $F^*(s) = \int_0^\infty \exp(-st)dF(t)$ can be expressed by a rational function, *i.e.*

$$F^*(s) = \frac{P(s)}{Q(s)}, \tag{1.88}$$

where $Q(s)$ and $P(s)$ are polynomials of degree k (≥ 1) and strictly larger than k, respectively, and $Q(0) = P(0)$ [since $F^*(0) = 1$, we define $0/0 = 1$. That is, the two functions above have no factors in common]. Hence, the LST of the renewal function is, from (1.73),

$$M^*(s) = \frac{P(s)}{Q(s) - P(s)} \tag{1.89}$$

for $\mathrm{Re}(s) > 0$. Since the function $Q(s) - P(s)$ is a polynomial of degree k and $M^*(s)$ can converge for $\mathrm{Re}(s) > 0$, Rouche's theorem [79] provides that the equation $Q(s) - P(s) = 0$ has k roots. Denote the roots by s_i ($i = 1, 2, \cdots, k$), where $s_1 = 0$. Then, the expression in (1.89) can be rewritten as

$$M^*(s) = \sum_{i=1}^{k} \frac{A_i}{s - s_i}, \tag{1.90}$$

where

$$A_i = \lim_{s \to s_i} \left\{ \frac{(s - s_i)P(s)}{Q(s) - P(s)} \right\} = \frac{P(s_i)}{Q^{\{1\}}(s_i) - P^{\{1\}}(s_i)} \tag{1.91}$$

with $\phi^{\{n\}}(s)$ denoting the nth derivative of the function $\phi(s)$. Fortunately, since taking the inverse transform of (1.90) yields $m(t) = \sum_{i=1}^{k} \exp(s_i t)$, a closed-form expression for the renewal function can be obtained as

$$M(t) = A_k t + \sum_{i=1}^{k-1} \frac{A_i e^{s_i t}}{s_i} - \sum_{i=1}^{k-1} \frac{A_i}{s_i}. \tag{1.92}$$

For large t, we have

$$M(t) \approx A_k t - \sum_{i=1}^{k-1} \frac{A_i}{s_i}. \tag{1.93}$$

Actually, the above method can provide a nice performance to calculate the renewal function numerically for some specific distribution functions. Also, Chaudhry [79] showed that the computation time required for this method is much less than that needed for the discritization method. As mentioned above, this method depends heavily on the shape of the inter-failure time distribution function. In other words, if the assumption that the LST of the renewal function can be expressed by a rational function is violated, then this method will lose its validation. However, it is known that the LSTs of the renewal functions for some typical inter-failure time distributions, such as the gamma distribution, become rational functions. Also, since there is one-to-one correspondence between matrix-exponential distributions and distributions having rational Laplace transforms, this method can be applied to a wider class of inter-failure time distributions than the usual phase distributions in 1.4.1 and 1.4.2.

Recently, a more sophisticated method with rational functions was proposed by Garg and Kalagnanam [80]. Their method is completely based on the Padé approximation for general stochastic discrete event systems by Gong, Nananukul and Yan [81]. The fundamental idea is to expand the probability density function and/or its convolution expression into a Maclaurin series. Further, by calculating an appropriate Padé approximation from the Maclaurin coefficients, the renewal function can be approximated consistently. First, we give a brief description of Padé approximation theory [82]. Suppose that the renewal function can be expressed by the following power-series expansion:

$$M(t) = \sum_{j=0}^{\infty} c_j t^j. \tag{1.94}$$

The $[l/m]$ Padé approximation to $M(t)$ is given by the following rational function:

$$[l/m](t) = \frac{P_l(t)}{Q_m(t)}, \tag{1.95}$$

where $P_l(t) = \sum_{j=0}^{l} a_j t^j$ and $Q_m(t) = \sum_{j=0}^{m} b_j t^j$, each of which has a MacLaurin expansion whose first $l + m + 1$ coefficients agree with those of $M(t)$. Since this can be expressed by $M(t) - P_l(t)/Q_m(t) = O(t^{l+m+1})$, multiplying both sides by $Q_m(t)$ and equating the coefficients of $t^0, t^1, \cdots, t^{l+m}$ and let b_0 by convention, we have the matrix equation $\mathbf{Cb} = -\mathbf{c}$, where

$$\mathbf{C} = \begin{bmatrix} c_{l-m+1} & c_{l-m+2} & \cdots & c_l \\ c_{l-m+2} & c_{l-m+3} & \cdots & c_{l+1} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ c_l & c_{l+1} & \cdots & c_{l+m-1} \end{bmatrix}, \tag{1.96}$$

$$\mathbf{b} = \begin{bmatrix} b_m \\ b_{m-1} \\ \cdot \\ \cdot \\ \cdot \\ b_1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_{l+1} \\ c_{l+2} \\ \cdot \\ \cdot \\ \cdot \\ c_{l+m} \end{bmatrix}. \tag{1.97}$$

The values of the coefficients a_j $(j = 0, 1, \cdots, l)$ can be evaluated using the estimates of b_j $(j = 0, 1, \cdots, m)$ as follows.

$$\begin{aligned} a_0 &= c_0 \\ a_1 &= c_1 + b_1 c_0 \\ &\cdot \\ &\cdot \\ &\cdot \\ a_l &= c_l + \sum_{j=1}^{\min(l,m)} b_j c_{l-j}. \end{aligned} \tag{1.98}$$

If the matrix $\mathbf{C}$ is singular, then more precise definition of Padé approximation will be needed. For more detail, see [82].

Consequently, if the renewal function can be expressed by the form in (1.94), the Padé approximation is available. Consider two probability density functions $f(t)$ and $g(t)$, where $f(0) = g(0) = 0$. Then, their MacLaurin series are given by

$$f(t) = \sum_{i=0}^{\infty} \frac{f^{\{i\}}(0)t^i}{i!}, \tag{1.99}$$

$$g(t) = \sum_{i=0}^{\infty} \frac{g^{\{i\}}(0)t^i}{i!}, \tag{1.100}$$

respectively. The MacLaurin series of the convolution $f * g(t)$ is

$$\begin{aligned} f * g(t) &= \int_0^t f(t-x)g(x)dx \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{f^{\{k\}}(0)g^{\{n-k\}}(0)t^{n+1}}{(n+1)!}. \end{aligned} \tag{1.101}$$

Since the nth derivative of the convolution is given by

$$(f * g)^{\{n\}}(0) = \sum_{j=0}^{n-1} f^{\{j\}}(0)g^{\{n-1-j\}}(0), \tag{1.102}$$

replacing $g(t)$ with $f^{(k-1)}(t)$, we have

$$(f^{(k)})^{\{n\}}(0) = \sum_{j=0}^{n-1} f^{\{j\}}(0)(f^{(k-1)})^{\{n-j-1\}}(0). \tag{1.103}$$

Hence, we can calculate the coefficients of the MacLaurin series expansion of the renewal function, *i.e.*

$$\begin{aligned} M(t) &= \sum_{k=1}^{\infty} F^{(k)}(t) \\ &= \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \frac{(F^{(k)})^{\{j\}}(0)t^j}{j!}. \end{aligned} \tag{1.104}$$

This means that the renewal function can be expressed simply by (1.94). Actually, the above algorithm can be applied to more complex computation. For instance, Gong, Nananukul and Yan [81] evaluated the expected number of renewals in a random interval and showed that it can be expressed in a quite simple series expansion.

1.7 Concluding Remarks

In this chapter, we have focused on the computational aspects of the renewal function and have outlined several methods of calculating it from a variety of standpoints. Before closing this chapter, we describe some topics that could not be discussed here because of the limited pages.

The statistical estimation problem for the renewal function is also important from the practical point of view. Frees [83], [84] [85] proposed some non-parametric estimators of the renewal function, provided that a complete sample of the inter-failure time is given. Schneider, Lin and O'Cinneide [86] compared two nonparametric estimators in terms of computation time. Also, Baker [87] and Grübel and Pitts [88] improved the non-parametric estimation method of Frees [83]. From the drastic development of neural computation, it has recently become possible for the neural network architecture to be applied to several kinds of scientific computation and information processing technique. Dohi *et al.* [89], [90] proposed a new computation method to calculate the renewal function applying the radial basis function neural network [91], [92]. In fact, their idea is similar to the cubic spline algorithm in Section 1.6, but it can also be applied to the statistical estimation problem.

The simulation technique for the renewal process also seems to be important, since the classical renewal theory can never provide the probability on the number of renewals. Brown, Solomon and Stephens [93] derived a minimum variance unbiased estimator of the renewal function. Shanthikumar [94] provided a unified framework based on the idea of uniformization to develop hybrid simulation and analytic models of renewal processes. Ross [95] improved the estimator in [93] by combining the techniques of control variates and conditional expectation. A series of results on coupling of the renewal processes were established by Lindvall [96], [97], [98], [99]. In fact, the coupling theory can be used to derive the other bounds of the renewal function as well as the asymptotic results for the renewal process such as Blackwell's theorem.

Although we have introduced several methods to calculate the renewal function in this chapter, of course, a numerical comparison of those methods should be carried out for a set of inter-failure time distributions. In the near future, we will report the comparative results. Also, in forgoing research, the approximation schemes for the general stochastic processes have to be developed. For instance, to the best of our knowledge, no efficient computation algorithm of the renewal function for the two-dimensional renewal process [100], [101], [102] has been reported except the straight line approximation and some bounds. Also, though the superimposed renewal processes [103], [104], [105], [106] are very useful to describe several phenomena, very few methods have been reported in the literature [107].

Acknowledgments

The first author is supported by Telecommunication Advancement Foundation, Tokyo, Japan. This work is also supported by Granted-in-Aid for Scientific Research from the Ministry of Education, Sports, Science and Culture of Japan under Grant No. 10558059, by the Research Program 1999 under the Institute for Advanced Studies of the Hiroshima Shudo University, Hiroshima, Japan, and by Nanzan University Pache Research Subsidy I–A.

References

1. Asmussen, S. (1987), Applied Probability and Queues. John Wiley & Sons, New York
2. Wolff, R. W. (1988), Stochastic Modeling and the Theory of Queues. Englewood Cliffs, Prentice-Hall, New Jersey
3. Heyman, D. P. and Sobel, M. J. (1982), Stochastic Models in Operations Research, Vol. 1. McGraw-Hill, New York
4. Trivedi, K. S. (1982), Probability and Statistics with Reliability, Queueing and Computer Science Applications. Englewood Cliffs. Prentice-Hall, New Jersey
5. Blischke, W. R. and Murthy, D. N. P. (1993), Warranty Cost Analysis. Marcel Dekker, New York
6. Blischke, W. R. and Murthy, D. N. P. (1996), Product Warranty Handbook. Marcel Dekker, New York
7. Barlow, R. E. and Proschan, F. (1965), Mathematical Theory of Reliability. John Wiley & Sons, New York
8. Barlow, R. E. and Proschan, F. (1975), Statistical Theory of Reliability and Life Testing. Holt, Rinehart and Winston, New York
9. Barlow, R. E. (1998), Engineering Reliability. SIAM, Philadelphia
10. Osaki, S. (1985), Stochastic System Reliability Modeling. World Scientific, Singapore
11. Osaki, S. (1992), Applied Stochastic System Modeling. Springer-Verlag, Berlin
12. Feller, W. (1966) An Introduction to Probability Theory and Its Applications II. John Wiley & Sons, New York
13. Çinlar, E. (1975), Introduction to Stochastic Processes. Englewood Cliffs, Prentice-Hall, New Jersey
14. Karlin, S. and Taylor, H. M. (1975), A First Course in Stochastic Processes. Academic Press, New York
15. Ross, S. M. (1983), Stochastic Processes. John Wiley & Sons, New York
16. Feller, W. (1941), "On the integral equation of renewal theory," *Annals of Mathematical Statistics*, **12**, 722–727
17. Feller, W. (1949), "Fluctuation theory of recurrent events," *Transactions on Amarican Mathematical Society*, **67**, 98–119
18. Smith, W. L. (1958), "Renewal theory and its ramifications," *Journal of Royal Statistical Society, Series B*, **20**, 243–302
19. Smith, W. L. (1959), "On the cumulants of renewal processes," *Biometrika*, **46**, 1–29
20. Cox, D. R. (1962), Renewal Theory. Methuen, London
21. Munter, M. (1971), "Discrete renewal processes," *IEEE Transactions on Reliability*, **R-20**, 46–51

22. Allan, R. N., Leite da Silva, A. M., Abu-Nasser, A. A. and Burchett, R. C. (1981), "Discrete convolution in power system reliability," *IEEE Transactions on Reliability*, **R-30**, 452–456
23. Kaio, N. and Osaki, S. (1988), "Review of discrete and continuous distributions in replacement models," *International Journal of Systems Science*, **19**, 171–177
24. Leadbetter, M. R. (1963), "On series expansion for renewal moments," *Biometrika*, **50**, 75–80
25. Smith, W. L. and Leadbetter, M. R. (1963), "On the renewal function for the Weibull distribution," *Technometrics*, **5**, 393–396
26. Lomnicki, Z. A. (1966), "A note on the Weibull renewal process," *Biometrika*, **53**, 375–381
27. Soland, R. M. (1968), A renewal theoretic approach to the estimation of future demand for replacement parts," *Operations Research*, **16**, 36–51
28. Soland, R. M. (1969), Availability of renewal functions for gamma and Weibull distributions with increasing hazard rate," *Operations Research*, **17**, 536–543
29. Jaquette, D. L. (1972), "Approximations to the renewal function $m(t)$," *Operations Research*, **20**, 722–727
30. Weiss, G. H. (1962), "Laguerre expansions for successive generations of a renewal process," *Journal of Research National Bureau of Standards*, **66B**, 165–168
31. Weeks, W. T. (1966), "Numerical inversion of Laplace transforms using Laguerre functions," *Journal of the Association for Computing Machinery*, **13**, 419–426
32. Keilson, J. and Nunn, W. R. (1979), "Laguerre transformation as a tool for the numerical solution of integral equations of convolution type," *Applied Mathematics and Computation*, **5**, 313–359
33. Sumita, U. and Kijima, M. (1988), "Theory and algorithms of the Laguerre transform, part I: theory," *Journal of the Operations Research Society of Japan*, **31**, 467–494
34. Sumita, U. and Kijima, M. (1990), "Theory and algorithms of the Laguerre transform, part II: algorithms," *Journal of the Operations Research Society of Japan*, **34**, 1481-1497
35. Abate, J., Choudhury, G. L. and Whitt, W. (1996), "On the Luguerre method for numerically inverting Laplace transforms," *INFORMS Journal on Computing*, **8**, 413–427
36. Blanc, J. P. C. (1992), "The power-series algorithm applied to the shortest-queue model," *Operations Research*, **40**, 157–167
37. Gong, W. B. and Hu, J. Q. (1992), "The Maclaurin series for the $G/G/1$ queue," *Journal of Applied Probability*, **29**, 176–184
38. Leadbetter, M. R. (1964), "Bounds on the error in the linear approximation to the renewal function," *Technometrics*, **51**, 355–364
39. Sahin, I. (1986), "On approximating the renewal function with its linear asymptot: how large is large enough," *Operations Research Letters*, **4**, 207–211
40. Carlsson, H. (1983), "Remainder term estimates of the renewal function," *Annals of Probability*, **11**, 143–157
41. Neuts, M. F. (1978), "Renewal processes of phase type," *Naval Research Logistics Quarterly*, **25**, 445–454
42. Neuts, M. F. (1981), Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Johns Hopkins University Press, Baltimore

43. Kao, E. P. C. (1988), "Computing the phase-type renewal and related function," *Technometrics*, **30**, 87–93
44. Gross, D. and Miller, D. R. (1984), "The randomization technique as a modeling tool and solution procedure for transient Markov process," *Operations Research*, **32**, 343–361
45. Whitt, W. (1982) "Approximating a point process by a renewal process I: two basic methods," *Operations Research*, **30**, 125–147
46. Altiok, T. (1985), "On the phase-type approximations of general distributions," *IIE Transactions*, **17**, 110–116
47. Van der Heijden, M. C. (1988), "On the three-moment approximation of a general distribution by a Coxian distribution," *Probability in the Engineering and Informational Sciences*, **2**, 257–261
48. Asmussen, S., Nerman, O. and Olsson, M. (1996), "Fitting phase-type distributions via the EM algorithm," *Scandinavian Journal of Statistics*, **23**, 419–441
49. Tijms, H. C. (1986), Stochastic Modelling and Analysis: A Computational Approach. John Wiley & Sons, New York
50. Smeitink, E. and Dekker, R. (1990), "A simple approximation to the renewal function," *IEEE Transactions on Reliability*, **39**, 71–75
51. Ross, S. M. (1987), "Approximations in renewal theory," *Probability in the Engineering and Informational Sciences*, **1**, 163–174
52. Angus, J. E. and Hong, X. (1996), "On the rate of convergence of the Ross approximation to the renewal function," *Probability in the Engineering and Informational Sciences*, **10**, 207–211
53. Bartholomew, D. J. (1963), "An approximate solution of the integral equation of renewal theory," *Journal of Royal Statistical Society, Series B*, **25**, 432–441
54. Ozbaykal, T. (1971), "Bounds and approximations of the renewal function", *M.S. Thesis, Naval Postgraduate School, Department of Operations Research and Administrative Science*, Monterey, California
55. Deligönul, Z. S. (1985), "An approximate solution of the integral equation of renewal theory," *Journal of Applied Probability*, **22**, 926–931
56. Barlow, R. E. and Proschan, F. (1964), "Comparison of replacement policies, and renewal theory implications," *Annals of Mathematical Statistics*, **35**, 577–589
57. Barlow, R. E. (1965), "Bounds on integrals with applications to reliability problems," *Annals of Mathematical Statistics*, **36**, 565–574
58. Marshall, K. T. (1973), "Linear bounds on the renewal function," *SIAM Journal of Applied Mathematics*, **24**, 245–250
59. Chen, Y-H. (1994), "Classes of life distributions and renewal counting process," *Journal of Applied Probability*, **31**, 1110–1115
60. Brown, M. (1980), "Bounds, inequalities, and monotonicity properties for some specialized renewal processes," *Annals of Probability*, **8**, 227–240
61. Brown, M. (1981), "Further monotonicity properties for specialized renewal processes," *Annals of Probability*, **9**, 893–895
62. Stone, C. J. (1972), "An upper bound for the renewal function," *Annals of Mathematical Statistics*, **43**, 2050–2052
63. Daley, D. J. (1978), "Upper bounds for the renewal functions via Fourier methods," *Annals of Probability*, **6**, 876–884
64. Jagerman, D. L. (1978), "An inversion technique for the Laplace transform with application," *Bell System Technical Journal*, **57**, 669–710

65. Jagerman, D. L. (1982), "An inversion technique for the Laplace transform," *Bell System Technical Journal*, **61**, 1995–2002
66. Abate, J. and Whitt, W. (1995), "Numerical inversion of Laplace transforms of probability distributions," *ORSA Journal on Computing*, **7**, 36–43
67. Davies, B. and Martin, B. L. (1979), "Numerical inversion of the Laplace transform: a survey and comparison of methods," *Journal of Computation Physics*, **33**, 1–32
68. Abate, J., Choudhury, G. and Whitt, W. (1999), "An introduction to numerical transform inversion and its application to probability models," in Computational Probability (Grassmann, W. K. ed.). 257–323, Kluwer Academic, The Netherlands
69. Cléroux, R. and Mcconalogue, D. J. (1976), "A numerical algorithm for recursively-defined convolution integrals involving distribution functions," *Management Science*, **22**, 1138–1146
70. Mcconalogue, D. J. (1978), "Convolution integrals involving probability distribution functions (algorithm 102)," *The Computer Journal*, **21**, 270–272
71. Mcconalogue, D. J. (1981), "Numerical treatment of convolution integrals involving distributions with densities having singularities at the origin," *Communications in Statistics B*, **10**, 265–280
72. Baxter, L. A. (1981), "Some remarks on numerical convolution," *Communications in Statistics B*, **10**, 281–288
73. Baxter, L. A., Scheuer, E. M., Mcconalogue, D. J. and Blischke, W. R., (1982), "On the tabulation of the renewal function," *Technometrics*, **24**, 151–156
74. Deligönül, Z. S. and Bilgen, S. (1984), "Solution of the Volterra equation of renewal theory with the Galerkin technique using cubic splines," *Journal of Statistical Computation and Simulation*, **20**, 37–45
75. Xie, M. (1989), "On the solution of renewal-type integral equations," *Communication in Statistics B*, **18**, 281–293
76. Boehme, T. K., Preuss, W. and Van del Wal, V. (1991), "On a simple numerical calculation for computing Stieltjes in reliability theory," *Probability in the Engineering and Informational Sciences*, **5**, 113–128
77. Banjevic, D. (1992), "On the recurrent solution of the renewal-type integral equation," *Probability in the Engineering and Informational Sciences*, **6**, 261–270
78. Ayhan, H., Limón-Robles, J. and Wortman, M. A. (1999), "An approach for computing tight numerical bounds on renewal functions," *IEEE Transactions on Reliability*, **48**, 182–188.
79. Chaudhry, M. L. (1995), "On computations of the mean and variance of the number of renewals: a unified approach," *Journal of the Operational Research Society*, **46**, 1352–1364
80. Garg, A. and Kalagnanam, J. R. (1998), "Approximations for the renewal function," *IEEE Transactions on Reliability*, **47**, 66–72
81. Gong, W-B., Nananukul, S. and Yan, A. (1995), "Padé approximation for stochastic discrete-event systems," *IEEE Transactions on Automatic Control*, **40**, 1349–1358
82. Baker, G. A. and Graves-Morris, P. (1981), Padé Approximations, Part I: Basic Theory. Addison-Wesley, Massachusetts
83. Frees, E. W. (1986a), "Nonparametric renewal function estimation," *Annals of Statistics*, **14**, 1366–1378

84. Frees, E. W. (1986b), "Warranty analysis and renewal function estimation," *Naval Research Logistics Quarterly*, **33**, 361–372
85. Frees, E. W. (1988), "Correction: nonparametric renewal function estimation," *Annals of Statistics*, **16**, 1741
86. Schneider, H., Lin, B. S. and O'Cinneide, C. (1990), "Comparison of nonparametric estimators for the renewal function," *Journal of Royal Statistical Society, Series C; Applied Statistics*, **39**, 56–61
87. Baker, R. D. (1993), "A nonparametric estimator of the renewal function," *Computers and Operations Research*, **20**, 167–178
88. Grübel, R. and Pitts, S. M. (1993), "Nonparametric estimation in renewal theory I: the empirical renewal function," *Annals of Statistics*, **21**, 1431–1451
89. Dohi, T., Nagai, H. and Osaki, S. (1998), "A computational method of the renewal function applying the RBF network," *Transactions of Japan Society for Industrial and Applied Mathematics*, **8**, 13–29
90. Nagai, H., Dohi, T. and Osaki, S. (2000), "The nonparametric estimation of the renewal function applying the radial basis function neural network," to appear in *Transactions of Japan Society for Industrial and Applied Mathematics*
91. Poggio, T. and Girosi, F. (1990), "Regularization algorithms for learning that are equivalent to multilayer networks," *Science*, **247**, 978–982
92. Poggio, T. and Girosi, F. (1990), "Networks for approximation and learning," *Proceedings of the IEEE*, **78**, 978–982
93. Brown, M., Solomon, H. and Stephens, M. A. (1981), "Monte Carlo simulation of the renewal functions," *Journal of Applied Probability*, **18**, 426–434
94. Shanthikumar, J. G. (1986) "Uniformization and hybrid simulation/analytic models of renewal processes," *Operations Research*, **34**, 573–580
95. Ross, S. M. (1989) "Estimating the mean number of renewals by simulation," *Probability in the Engineering and Informational Sciences*, **3**, 319–321
96. Lindvall, T. (1979), "On coupling of discrete renewal processes," *Z. Warsch. Verw. Geb.*, **48**, 57–70
97. Lindvall, T. (1982), "On coupling of continuous time renewal processes," *Journal of Applied Probability*, **19**, 82–89
98. Lindvall, T. (1986), "On coupling of renewal processes with use of failure rates," *Stochastic Processes and Their Applications*, **22**, 1–15
99. Lindvall, T. (1992), "A simple coupling of renewal processes," *Journal of Applied Probability*, **24**, 1010–1011
100. Hunter, J. J. (1974), "Renewal theory in two-dimensions: basic results," *Advances in Applied Probability*, **6**, 376–391
101. Hunter, J. J. (1974), "Renewal theory in two-dimensions: asymptotic results," *Advances in Applied Probability*, **6**, 546–562
102. Hunter, J. J. (1977) "Renewal theory in two-dimensions: bounds on the renewal function," *Advances in Applied Probability*, **9**, 527–541
103. Çinlar, E. (1968), "On the superposition of m-dimensional point processes," *Journal of Applied Probability*, **5**, 169–176
104. Blumenthal, S., Greenwood, J. A. and Herbach, L. (1971), "Superimposed non-stationary renewal processes," *Journal of Applied Probability*, **8**, 184–192

105. Blumenthal, S., Greenwood, J. A. and Herbach, L. (1973), "The transient reliability behavior of series systems or superimposed renewal processes," *Technometrics*, **15**, 255–269
106. Blumenthal, S., Greenwood, J. A. and Herbach, L. (1976), "A comparison of the bad as old and superimposed renewal models," *Management Science*, **23**, 280–285
107. Blumenthal, S. (1993), "New approximations for the event count distribution for superimposed renewal processes at the time origin with application to the reliability of new series systems," *Operations Research*, **41**, 409–418

2. Stochastic Orders in Reliability Theory

Masamitsu Ohnishi
Graduate School of Economics,
Osaka University,
Osaka 560–0043, Japan

Summary.

Stochastic orders and related inequalities are very important in various areas of reliability and maintainability theory. The purpose of this chapter is to provide a brief survey of the useful known results concerning stochastic orders and their applications developed in these areas.

Keywords: stochastic orders, stochastic inequalities, univariate characterization, conditional stochastic orders, bivariate characterization, notions of aging, stochastic comparisons, reliability systems, redundancy improvement, maintenance policies

2.1 Introduction

Stochastic orders (SOs) and inequalities are very important in various areas of applied probability and statistics. In particular, in reliability and maintainability theory, SOs have important roles in, for example, defining notions of positive or negative aging, bounding system reliabilities and availability, and comparing maintenance policies. During the last four decades, vast literatures have been devoted to the developments of the studies on the various SOs and their applications in reliability and maintainability theory. The purpose of this chapter is to provide a brief survey of known results concerning the various SOs and their applications developed in these areas, which are useful to both reliability researchers and practitioners. The criteria adopted for choice of materials are neither generality nor newness, but simplicity, usefulness, and applicability. Accordingly, some results are not new but even classical and well known.

First, in Section 2.2, we define several SOs in a systematic way. These are classified into two categories: (1) SOs generated from univariate function classes (Subsection 2.2.1), and (2) (uniformly) conditional SOs (Subsection 2.2.2). In this section, we give various characterizations of these SOs, and relationships among them. The final subsection (Subsection 2.2.3) presents recent important results which state that both types of

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

these SOs can be characterized by bivariate function classes in a unified way.

In Section 2.3, we gather many applications from various areas in reliability and maintainability theory. The first applications in Subsection 2.3.1 concern the unified treatment of various notions of positive or negative aging. Subsection 2.3.2 gives various and related stochastic inequalities in terms of the SOs, which are useful in the reliability maintainability theory. Subsection 2.3.3 considers the problems of bounding, evaluating, and comparing system reliabilities. Next, Section 2.3.4 discusses the problems of redundancy improvement. In Section 2.3.5, we gather many results on stochastic comparisons of maintenance policies: replacement–upon–failure policy, age replacement policy, block replacement policy, minimal repair policy, and minimal repair policy with block replacement.

2.2 Definitions and Basic Properties

For a given set A, a subset B of the Cartesian product $A \times A$ is called a *binary relation* on A. If $(a, b) \in A \times A$ is in $B \subset A \times A$ then we write aBb.

Definition 2.2.1 (ordering relation) A binary relation B on a given set A is called an *ordering relation* on A if
(P1) (reflexivity): aBa;
(P2) (transitivity): aBb and bBc imply aBc
are satisfied. In addition, if
(P3) (anti–symmetry): aBb and bBa imply $a = b$,
B is called a *partial ordering relation*, where $a, b, c \in A$.

If A is a collection of random elements, such as random variables, random vectors, or stochastic processes, an ordering relation B is called a *stochastic ordering relation* [or a *stochastic order* (SO)], where the equality $=$ in (P3) is interpreted as the equivalence in distribution (or probabilistic law) $=_{\mathrm{d}}$ (or $=_{\mathrm{st}}$). In fact, most of stochastic ordering relations are considered between marginal probability distributions of individual random elements instead of random elements themselves.

In this chapter, we are mainly concerned with univariate SOs, that is, stochastic ordering relations between (real–valued) random variables (RVs). For an RV X, we define

$F_X(x) := P(X \le x)$, $x \in \mathbb{R} := (-\infty, \infty)$: cumulative distribution function (CDF);

$\overline{F}_X(x) := P(X > x)$, $x \in \mathbb{R}$: survival function (SF);

$H_X(x) := -\log \overline{F}_X(x)$, $x \in \mathbb{R}$: cumulative hazard rate function (CHRF);

$m_X(x) := \int_x^\infty \overline{F}_X(u)du \,/\, \overline{F}_X(x)$, $x \in \mathbb{R}$: mean residual life function (MRLF) [provided that the integral $\int_x^\infty \overline{F}_X(u)du$ is well defined, and this is the case when X has a finite mean].

If X is a continuous RV, that is, F_X is absolutely continuous so that it is differentiable almost everywhere, we denote

$f_X(x)$, $x \in \mathbb{R}$: probability density function (PDF);

$h_X(x) := f_X(x)/\overline{F}_X(x)$, $x \in \mathbb{R}$: hazard rate function (HRF) or failure rate function (FRF);

$r_X(x) := f_X(x)/F_X(x)$, $x \in \mathbb{R}$: reversed hazard rate function (RHRF) or reversed failure rate function (RFRF).

In this section, we introduce the definitions of several SOs of interest in reliability theory, and gather basic information needed for subsequent developments. Since the pages are limited, all the results are given with no derivations or proofs. The reader interested in full details of the SOs and/or other SOs including various multivariate SOs should consult Stoyan [61], Marshall and Olkin [45], and Shaked and Shanthikumar [56].

2.2.1 Stochastic orders generated from univariate functions

Most important SOs considered in the reliability literature can be classified into two types: (1) SOs generated from univariate function classes, and (2) SOs based on the conditional distributions.

The first type of SOs consist of those generated from univariate function classes. Let $\mathcal{F}$ be a class of real–valued functions defined on $\mathbb{R}$. For RVs X and Y with CDFs F_X and F_Y, a stochastic ordering relation $X \succeq Y$ (or $F_X \succeq F_Y$) is said to be *generated from* $\mathcal{F}$ if

$$\int_{-\infty}^{\infty} f(x)dF_X(x) \geq \int_{-\infty}^{\infty} f(x)dF_Y(x) \qquad \text{for all } f \in \mathcal{F} \tag{2.1}$$

for which the integrals exist, or equivalently,

$$E[f(X)] \geq E[f(Y)] \qquad \text{for all } f \in \mathcal{F}. \tag{2.2}$$

In this case, we denote it by $X \geq_{\mathcal{F}} Y$ (or $F_X \geq_{\mathcal{F}} F_Y$) to make the dependence of $\mathcal{F}$ explicit. The simplest example of a stochastic order of this type is a generated case when $\mathcal{F}$ is a singleton set consisting of the identity function (*i.e.*, $i_{\mathbb{R}} : x \in \mathbb{R} \mapsto x \in \mathbb{R}$) only. Then, the resulting SO $\geq_{\mathcal{F}}$ becomes the usual ordering by the expected values (or means) of RVs, that is,

$$X \geq_{\mathcal{F}} Y \Longleftrightarrow E[X] \geq E[Y]. \tag{2.3}$$

It is noted that this SO does not satisfy (P3) of Definition 2.2.1; hence it is not a partial order.

SOs of the above type considered in the present chapter, which frequently appear in the reliability literature, are the following.

Definition 2.2.2 Let X and Y be RVs with CDFs F_X and F_Y. Then,

(1) X is said to be greater than Y in the sense of *ordinary stochastic order* (OSO), written as $X \geq_{\mathrm{st}} Y$ (or $F_X \geq_{\mathrm{st}} F_Y$), if it is generated by

$$\mathcal{F}_{\mathrm{st}} := \{f : \mathbb{R} \to \mathbb{R},\ f(x) \text{ is increasing in } x \in \mathbb{R}\}; \tag{2.4}$$

(2) X is said to be greater than Y in the sense of *convex order* (CxO), written as $X \geq_{\mathrm{cx}} Y$ (or $F_X \geq_{\mathrm{cx}} F_Y$), if it is generated by

$$\mathcal{F}_{\mathrm{cx}} := \{f : \mathbb{R} \to \mathbb{R},\ f(x) \text{ is convex in } x \in \mathbb{R}\}; \tag{2.5}$$

(3) X is said to be greater than Y in the sense of *concave order* (CvO), written as $X \geq_{\mathrm{cv}} Y$ (or $F_X \geq_{\mathrm{cv}} F_Y$), if it is generated by

$$\mathcal{F}_{\mathrm{cv}} := \{f : \mathbb{R} \to \mathbb{R},\ f(x) \text{ is concave in } x \in \mathbb{R}\}; \tag{2.6}$$

(4) X is said to be greater than Y in the sense of *increasing convex order* (ICxO), written as $X \geq_{\text{icx}} Y$ (or $F_X \geq_{\text{icx}} F_Y$), if it is generated by

$$\mathcal{F}_{\text{icx}} := \{f : \mathbb{R} \to \mathbb{R},\ f(x) \text{ is increasing and convex in } x \in \mathbb{R}\}; \quad (2.7)$$

(5) X is said to be greater than Y in the sense of *increasing concave order* (ICvO), written as $X \geq_{\text{icv}} Y$ (or $F_X \geq_{\text{icv}} F_Y$), if it is generated by

$$\mathcal{F}_{\text{icv}} := \{f : \mathbb{R} \to \mathbb{R},\ f(x) \text{ is increasing and concave in } x \in \mathbb{R}\}. \quad (2.8)$$

In the above definitions, and throughout this chapter, the terms "increasing" and "decreasing" are used in the weak sense, *i.e.*, to mean 'nondecreasing' and 'nonincreasing,' respectively. Similarly, the terms "concave" and "convex" are used in the weak sense.

Remark 2.2.1

(1) SOs of the above type depend only on the marginal distributions. Hence, for these stochastic ordering relations between X and Y, the dependence between these two RVs does not matter. Furthermore, they are not necessarily defined on the same probability space.
(2) In other areas such as decision theory, economics, and finance, OSO, CvO, and ICvO are usually called the *first order stochastic dominance* (FSD), the *mean preserving contraction* (MPC), and the *second order stochastic dominance* (SSD), respectively.
(3) CvO and ICvO are dual to CxO and ICxO, respectively. That is, $X \geq_{\text{cv}} Y$ if and only if $Y \geq_{\text{cx}} X$, and $X \geq_{\text{icv}} Y$ if and only if $-Y \geq_{\text{icx}} -X$ (note the minus signs for ICvO).
(4) Each of $X \geq_{\text{st}} Y$ and $X \geq_{\text{cx}} Y$ implies $X \geq_{\text{icx}} Y$, and each of $X \geq_{\text{st}} Y$ and $X \geq_{\text{cv}} Y$ implies $X \geq_{\text{icv}} Y$.

Figure 2.1 depicts the conceptual picture of the relations.

Higher degrees of SOs could be generated from the following two families of univariate function classes: for a positive integer $n \in \mathbb{Z}_{++} := \{1, 2, \cdots\}$, we define

$$\mathcal{F}^{n,+} := \left\{f : \mathbb{R} \to \mathbb{R},\ f^{(k)} \geq 0 \text{ for all } k = 1, \cdots, n\right\}; \quad (2.9)$$

$$\mathcal{F}^{n,-} := \left\{f : \mathbb{R} \to \mathbb{R},\ (-1)^{k+1} f^{(k)} \geq 0 \text{ for all } k = 1, \cdots, n\right\}, \quad (2.10)$$

where $f^{(k)}$ is the k–th order derivative of function f. It is noted $\mathcal{F}_{\text{st}}$, $\mathcal{F}_{\text{icx}}$, and $\mathcal{F}_{\text{icv}}$ correspond to $\mathcal{F}^{1,+}$, $\mathcal{F}^{2,+}$, and $\mathcal{F}^{2,-}$, respectively. In order to characterize the SOs generated from function classes $\mathcal{F}^{n,+}$ and $\mathcal{F}^{n,-}$, functions $F_X^{n,+}$ and $F_X^{n,-}$ have important roles, where, for an RV X,

$$F_X^{0,+}(x) = F_X^{0,-}(x) := f_X(x), \qquad x \in \mathbb{R}, \quad (2.11)$$

and, for $n = 1, 2, \cdots$, we recursively define as

$$F_X^{n,+}(x) := \int_x^\infty F_X^{n-1,+}(u)du, \qquad x \in \mathbb{R}; \quad (2.12)$$

$$F_X^{n,-}(x) := \int_{-\infty}^x F_X^{n-1,-}(u)du, \qquad x \in \mathbb{R}, \quad (2.13)$$

provided that these integrals are well defined. Note that

$$F_X^{1,+}(x) = \overline{F}_X(x), \qquad x \in \mathbb{R}; \tag{2.14}$$

$$F_X^{2,+}(x) = \int_x^\infty \overline{F}_X(u)du, \qquad x \in \mathbb{R}; \tag{2.15}$$

$$F_X^{1,-}(x) = F_X(x), \qquad x \in \mathbb{R}; \tag{2.16}$$

$$F_X^{2,-}(x) = \int_{-\infty}^x F_X(u)du, \qquad x \in \mathbb{R}. \tag{2.17}$$

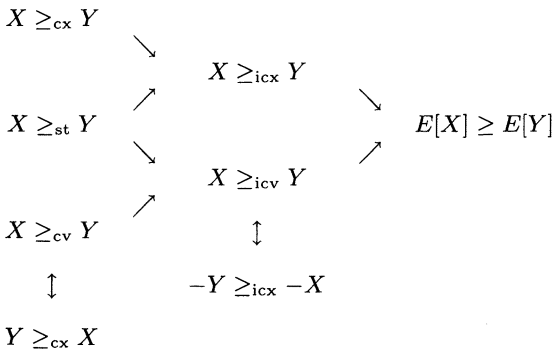


Fig. 2.1. Stochastic orders generated from univariate functions

We start with the well known equivalent conditions of OSO.

Proposition 2.2.1 For RVs X and Y with CDFs F_X and F_Y, the following are equivalent:
(1) $X \geq_{\mathrm{st}} Y$;
(2) $F_X(x) \leq F_Y(x)$ (or $\overline{F}_X(x) \geq \overline{F}_Y(x)$) for all $x \in \mathbb{R}$;
(3) There exist RVs $\widehat{X}$ and $\widehat{Y}$ defined on the same probability space, for which $\widehat{X} \geq \widehat{Y}$ almost certainly, $X =_{\mathrm{d}} \widehat{X}$ and $Y =_{\mathrm{d}} \widehat{Y}$. □

Characterizations such as (3) in Proposition 2.2.1 are called *stochastic coupling* (see also (3) in Propositions 2.2.2–2.2.5 below). Note that, for any RV X with CDF F_X, it holds that

$$\overline{F}_X(x) = P(X > x) = E\left[1_{(x,\infty)}\right] \qquad \text{for all } x \in \mathbb{R}, \tag{2.18}$$

where $1_A(\cdot)$ is the indicator function of set A. This and Proposition 2.2.1 (2) imply that OSO $\geq_{\mathrm{st}}$ can be generated by

$$\{1_{(a,\infty)}(\cdot) : \ a \in \mathbb{R}\} \ (\subset \mathcal{F}_{\mathrm{st}}). \tag{2.19}$$

Proposition 2.2.2 For RVs X and Y with CDFs F_X and F_Y, the following are equivalent:
(1) $X \geq_{\mathrm{cx}} Y$;
(2) $\int_x^\infty \overline{F}_X(u)du \geq \int_x^\infty \overline{F}_Y(u)du$ for all $x \in \mathbb{R}$, and $E[X] = E[Y]$;
(3) There exist RVs $\widehat{X}$ and $\widehat{Y}$ defined on the same probability space for which $E[\widehat{X}|\ \widehat{Y}] = \widehat{Y}$ almost certainly, $X =_{\mathrm{d}} \widehat{X}$, and $Y =_{\mathrm{d}} \widehat{Y}$.
□

Proposition 2.2.3 For RVs X and Y with CDFs F_X and F_Y, the following are equivalent:
(1) $X \geq_{\mathrm{cv}} Y$;
(2) $\int_{-\infty}^{x} F_X(u)du \leq \int_{-\infty}^{x} F_Y(u)du$ for all $x \in \mathbb{R}$, and $E[X] = E[Y]$;
(3) There exist RVs $\widehat{X}$ and $\widehat{Y}$ defined on the same probability space for which $\widehat{X} = E[\widehat{Y} \mid \widehat{X}]$ almost certainly, $X =_{\mathrm{d}} \widehat{X}$, and $Y =_{\mathrm{d}} \widehat{Y}$.

□

Proposition 2.2.4 For RVs X and Y with CDFs F_X and F_Y, the following are equivalent:
(1) $X \geq_{\mathrm{icx}} Y$ (or $F_X \geq_{\mathrm{icx}} F_Y$);
(2) $\int_{x}^{\infty} \overline{F}_X(u)du \geq \int_{x}^{\infty} \overline{F}_Y(u)du$ for all $x \in \mathbb{R}$;
(3) There exist RVs $\widehat{X}$ and $\widehat{Y}$ defined on the same probability space for which $E[\widehat{X} \mid \widehat{Y}] \geq \widehat{Y}$ almost certainly, $X =_{\mathrm{d}} \widehat{X}$ and $Y =_{\mathrm{d}} \widehat{Y}$. □

Note that, for any RV X with CDF F_X, it holds that

$$\int_{x}^{\infty} \overline{F}_X(u)du = E\left[(X-x)_+\right] = E[\max\{x, X\}] - x \quad \text{for all } x \in \mathbb{R}, \tag{2.20}$$

where, for a real number $a \in \mathbb{R}$, $(a)_+ := \max\{a, 0\}$.
Equation (2.20) and Proposition 2.2.4 (2) imply that ICxO $\geq_{\mathrm{icx}}$ can be generated by

$$\{(\cdot - a)_+ : a \in \mathbb{R}\} \ (\subset \mathcal{F}_{\mathrm{icx}}). \tag{2.21}$$

Proposition 2.2.5 For RVs X and Y with CDFs F_X and F_Y, the following are equivalent:
(1) $X \geq_{\mathrm{icv}} Y$ (or $F_X \geq_{\mathrm{icv}} F_Y$);
(2) $\int_{-\infty}^{x} F_X(u)du \leq \int_{-\infty}^{x} F_Y(u)du$ for all $x \in \mathbb{R}$;
(3) There exist RVs $\widehat{X}$ and $\widehat{Y}$ defined on the same probability space for which $\widehat{X} \geq E[\widehat{Y} \mid \widehat{X}]$ almost surely, $X =_{\mathrm{d}} \widehat{X}$, and $Y =_{\mathrm{d}} \widehat{Y}$. □

Note that, for any RV X with CDF F_X, it holds that

$$\int_{-\infty}^{x} F_X(u)du = E\left[(x-X)_+\right] = x - E[\min\{x, X\}] \quad \text{for all } x \in \mathbb{R}. \tag{2.22}$$

Equation (2.22) and Proposition 2.2.5 (2) imply that ICvO $\geq_{\mathrm{icv}}$ can be generated by

$$\{-(a - \cdot)_+ : a \in \mathbb{R}\} \ (\subset \mathcal{F}_{\mathrm{icv}}). \tag{2.23}$$

Stoyan [61] proposed the following properties that SO $\succeq$ is desired to possess: For RVs X and Y,

(C) $X \succeq Y$ implies $X + Z \succeq Y + Z$ for any RV which is independent of X and Y;
(R) For $a, b \in \mathbb{R}$ with $a \geq b$, we have $a \succeq b$;
(M) $X \succeq Y$ implies $aX \succeq aY$ for any $a > 0$;
(E) $X \succeq Y$ implies $E[X] \geq E[Y]$.

Property (C) is called the *convolution property*. We note that any real number can be considered as a degenerate RV, which is independent of any other RV. Hence, property (C) implies that:

(C′) $X \succeq Y$ implies $X + c \succeq Y + c$ for any real number c.

Property (R) is called the *real number property*, (M) the *multiplication property*, and (E) the *expectation property*.

For the SOs given in Definition 2.2.2, we have the following result. The proof of the convolution properties is straightforward from the fact that, for $\mathcal{F} = \mathcal{F}_{\rm st}$, $\mathcal{F}_{\rm cx}$, $\mathcal{F}_{\rm cv}$, $\mathcal{F}_{\rm icx}$, and $\mathcal{F}_{\rm icv}$,

$$\text{if } f(\cdot) \in \mathcal{F} \text{ then } f(\cdot + y) \in \mathcal{F} \qquad \text{for all } y \in \mathbb{R}. \tag{2.24}$$

Proposition 2.2.6
(1) OSO satisfies all the properties;
(2) CxO and CvO satisfy (C) and (M). (E) holds with an equality;
(3) ICxO and ICvO satisfy all the properties. □

2.2.2 Conditional stochastic orders

The second type of SOs are based on the conditional distributions. Let $\mathcal{H}$ be a family of Borel subsets of $\mathbb{R}$. For an RV X and an event E, the RV conditional on the event E is denoted by $[X|\ E]$, provided that $P(E) > 0$. Especially, for an RV X and a Borel subset D of $\mathbb{R}$, the RV conditional on the event $\{X \in D\}$ is denoted by $[X|\ X \in D]$, provided that $P(X \in D) > 0$. Then, for two RVs X and Y, we write $X \succeq_{\mathcal{H}} Y$ if

$$[X|\ X \in D] \geq_{\rm st} [Y|\ Y \in D] \qquad \text{for all } D \in \mathcal{H} \tag{2.25}$$

for which the conditional RVs are well defined. SOs which can be defined in such a way are called (*uniformly*) *conditional stochastic orders* [29], [70], [71]. Among them, we mainly consider the following three SOs of the second type.

Definition 2.2.3 Let X and Y be RVs with CDFs F_X and F_Y. Then,
(1) X is said to be greater than Y in the sense of *likelihood ratio order* (LRO), written as $X \geq_{\rm lr} Y$ (or $F_X \geq_{\rm lr} F_Y$), if (2.25) holds for $\mathcal{H}_{\rm lr} := \{(s,t] : -\infty < s < t < \infty\}$;
(2) X is said to be greater than Y in the sense of *hazard rate order* (HRO), written as $X \geq_{\rm hr} Y$ (or $F_X \geq_{\rm hr} F_Y$), if (2.25) holds for $\mathcal{H}_{\rm hr} := \{(s,\infty) : s \in \mathbb{R}\}$;
(3) X is said to be greater than Y in the sense of *reversed hazard rate order* (RHRO), written as $X \geq_{\rm rh} Y$ (or $F_X \geq_{\rm rh} F_Y$), if (2.25) holds for $\mathcal{H}_{\rm rh} := \{(-\infty, t] : t \in \mathbb{R}\}$.

Remark 2.2.2
(1) By analogy with Remark 2.2.1, these SOs depend only on the marginal distributions. Hence, for a stochastic ordering relation between X and Y, the dependence between the two RVs does not matter. Furthermore, they are not necessarily defined on the same probability space.
(2) RHRO is dual to HRO. That is, $X \geq_{\rm rh} Y$ if and only if $-Y \geq_{\rm hr} -X$ (note the minus signs).
(3) $X \geq_{\rm lr} Y$ implies both $X \geq_{\rm rh} Y$ and $X \geq_{\rm hr} Y$;
(4) Each of $X \geq_{\rm rh} Y$ and $X \geq_{\rm hr} Y$ implies $X \geq_{\rm st} Y$;
(5) For an example, in (2.25), if we replace OSO $\geq_{\rm st}$ with ICxO $\geq_{\rm icx}$, we obtain another rich family of SOs, so-called (*uniformly*) *conditional variability orders* [71].

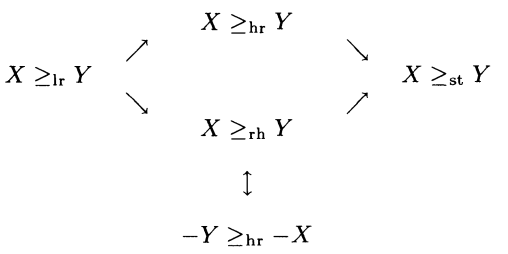


Fig. 2.2. Conditional stochastic orders

Figure 2.2 depicts the conceptual picture of the relations.
Since $\mathcal{H}_{\mathrm{lr}} = \{(s,t] : -\infty < s < t < \infty\}$, (2.25) is rewritten as

$$\frac{P(x < X \leq t)}{P(s < X \leq t)} \geq \frac{P(x < Y \leq t)}{P(s < Y \leq t)} \qquad \text{for all } -\infty < s < x < t < \infty, \tag{2.26}$$

so that $X \geq_{\mathrm{lr}} Y$ if and only if

$$\frac{F_X(t) - F_X(x)}{F_X(t) - F_X(s)} \geq \frac{F_Y(t) - F_Y(x)}{F_Y(t) - F_Y(s)} \qquad \text{for all } -\infty < s < x < t < \infty. \tag{2.27}$$

With the usual limiting arguments, we have the following characterization of LRO.

Proposition 2.2.7 Let X and Y be RVs with CDFs F_X and F_Y, and suppose that they have PDFs f_X and f_Y, respectively. Then, $X \geq_{\mathrm{lr}} Y$ (or $F_X \geq_{\mathrm{lr}} F_Y$) if and only if $f_X(x)/f_Y(x)$ is increasing in x on the union of the supports of X and Y (where $a/0$ is taken to be equal to ∞ whenever $a > 0$), or, equivalently

$$\begin{vmatrix} f_Y(s) & f_Y(t) \\ f_X(s) & f_X(t) \end{vmatrix} \geq 0 \qquad \text{for all } -\infty < s < t < \infty. \tag{2.28}$$

□

Since $\mathcal{H}_{\mathrm{hr}} = \{(s,\infty) : s \in \mathbb{R}\}$, (2.25) is rewritten as

$$\frac{P(X > x,\ X > s)}{P(X > s)} \geq \frac{P(Y > x,\ Y > s)}{P(Y > s)} \qquad \text{for all } -\infty < s < x < \infty, \tag{2.29}$$

so that $X \geq_{\mathrm{rh}} Y$ if and only if

$$\frac{\overline{F}_X(x)}{\overline{F}_X(s)} \geq \frac{\overline{F}_Y(x)}{\overline{F}_Y(s)} \qquad \text{for all } -\infty < s < x < \infty. \tag{2.30}$$

Hence we have the next characterization of HRO, which is similar to Proposition 2.2.7.

Proposition 2.2.8 For RVs X and Y with CDFs F_X and F_Y, $X \geq_{\mathrm{hr}} Y$ (or $F_X \geq_{\mathrm{hr}} F_Y$) if and only if $\overline{F}_X(x)/\overline{F}_Y(x)$ is increasing in $x \in \mathbb{R}$ (where $a/0$ is taken to be equal to ∞ whenever $a > 0$), or, equivalently,

$$\begin{vmatrix} \overline{F}_Y(s) & \overline{F}_Y(x) \\ \overline{F}_X(s) & \overline{F}_X(x) \end{vmatrix} \geq 0 \qquad \text{for all } -\infty < s < x < \infty. \tag{2.31}$$

□

For RV X with PDF f_X, recall that the function $h_X(x) := f_X(x)/\overline{F}_X(x)$, $x \in \mathbb{R}$ is called the *hazard rate function* (HRF) of X (or *failure rate function*). The next proposition justifies the term "hazard rate order."

Proposition 2.2.9 Let X and Y be RVs with CDFs F_X and F_Y. Suppose that they have PDFs f_X and f_Y so that they have HRFs h_X and h_Y, respectively. Then, $X \geq_{\mathrm{hr}} Y$ (or $F_X \geq_{\mathrm{hr}} F_Y$) if and only if $h_X(x) \leq h_Y(x)$ for all $x \in \mathbb{R}$. □

Since $\mathcal{H}_{\mathrm{rh}} = \{(-\infty, t] : t \in \mathbb{R}\}$, (2.25) is rewritten as

$$\frac{P(X \leq x,\ X \leq t)}{P(X \leq t)} \leq \frac{P(Y \leq x,\ Y \leq t)}{P(Y \leq t)} \qquad \text{for all } -\infty < x < t < \infty, \tag{2.32}$$

so that $X \geq_{\mathrm{rh}} Y$ if and only if

$$\frac{F_X(x)}{F_X(t)} \leq \frac{F_Y(x)}{F_Y(t)} \qquad \text{for all } -\infty < x < t < \infty. \tag{2.33}$$

Hence we have the next characterization of RHRO, which is similar to Propositions 2.2.7, 2.2.8.

Proposition 2.2.10 For RVs X and Y with CDFs F_X and F_Y, $X \geq_{\mathrm{rh}} Y$ (or $F_X \geq_{\mathrm{rh}} F_Y$) if and only if $F_X(x)/F_Y(x)$ is increasing in $x \in \mathbb{R}$ (where $a/0$ is taken to be equal to ∞ whenever $a > 0$), or, equivalently,

$$\begin{vmatrix} F_Y(x) & F_Y(t) \\ F_X(x) & F_X(t) \end{vmatrix} \geq 0 \qquad \text{for all } -\infty < x < t < \infty. \tag{2.34}$$

□

For RV X with PDF f_X, recall that the function $r_X(x) := f_X(x)/F_X(x)$, $x \in \mathbb{R}$ is called the *reversed hazard rate function* (RHRF) of X. The next proposition justifies the term "reversed hazard rate order."

Proposition 2.2.11 Let X and Y be RVs with CDFs F_X and F_Y. Suppose that they have PDFs f_X and f_Y so that they have RHRFs r_X and r_Y, respectively. Then, $X \geq_{\mathrm{rh}} Y$ (or $F_X \geq_{\mathrm{rh}} F_Y$) if and only if $r_X(x) \geq r_Y(x)$ for all $x \in \mathbb{R}$.

For the SOs given in Definition 2.2.3, the convolution property (C) needs not hold (see Appendix 2.A for Pólya frequency of order 2 (PF_2) functions).

Proposition 2.2.12

(1) LRO satisfies (C) for Z with a PF_2 PDF. It satisfies (R), (M) and (E);

(2) HRO satisfies (C) for Z with a PF_2 SF. It satisfies (R), (M) and (E).
(3) RHRO satisfies (C) for Z with a PF_2 CDF. It satisfies (R), (M) and (E). □

Let us return to the definition (2.25) of conditional stochastic orders, and replace the ordinary stochastic order ($\geq_{\rm st}$) with ordering by the expected values. Then, with $\mathcal{H} := \{(s, \infty) : s \in \mathbb{R}\}$ $(= \mathcal{H}_{\rm hr})$, we have another stochastic ordering:

$$X \succeq Y \Longleftrightarrow E[X|\ X > s] \geq E[Y|\ Y > s] \qquad \text{for all } s \in \mathbb{R}. \tag{2.35}$$

Since

$$E[X|\ X > s] = E[X - s|\ X > s] + s = m_X(s) + s, \tag{2.36}$$

(2.35) is rewritten as

$$X \succeq Y \Longleftrightarrow m_X(s) \geq m_Y(s) \qquad \text{for all } s \in \mathbb{R}, \tag{2.37}$$

so that, in this case, X is said to be greater than Y in the sense of *mean residual life order* (MRLO), written $X \geq_{\rm mrl} Y$ (or $F_X \geq_{\rm mrl} F_Y$), where, for an RV X, $m_X(x) := \int_x^\infty \overline{F}_X(u)du \,/\, \overline{F}_X(x)$, $x \in \mathbb{R}$ is the *mean residual life function* (MRLF) of X (provided that the integral $\int_x^\infty \overline{F}_X(u)du$ is well defined, and this is the case when X has a finite mean). Furthermore, it is noted that

$$X \geq_{\rm hr} Y \longrightarrow X \geq_{\rm mrl} Y. \tag{2.38}$$

Proposition 2.2.13 For RVs X and Y with CDFs F_X and F_Y, $X \geq_{\rm mrl} Y$ (or $F_X \geq_{\rm mrl} F_Y$) if and only if $\int_x^\infty \overline{F}_X(u)du \,/ \int_x^\infty \overline{F}_Y(u)du$ is increasing in $x \in \mathbb{R}$ (where $a/0$ is taken to be equal to ∞ whenever $a > 0$), or, equivalently,

$$\begin{vmatrix} \int_s^\infty \overline{F}_Y(u)du & \int_x^\infty \overline{F}_Y(u)du \\ \int_s^\infty \overline{F}_X(u)du & \int_x^\infty \overline{F}_X(u)du \end{vmatrix} \geq 0 \qquad \text{for all } -\infty < s < x < \infty. \tag{2.39}$$

□

Propositions 2.2.7, 2.2.8, 2.2.10, and 2.2.13 suggest the introduction of the following two families of stochastic ordering relations:

Definition 2.2.4 For a nonnegative integer $n \in \mathbb{Z}_+ := \{0, 1, \cdots\}$:
(1) $X \geq^{n,+} Y$ (or $F_X \geq^{n,+} F_Y$) if $F_X^{n,+}(x)/F_Y^{n,+}(x)$ is increasing in $x \in \mathbb{R}$ (where $a/0$ is taken to be equal to ∞ whenever $a > 0$), or, equivalently, if

$$\begin{vmatrix} F_Y^{n,+}(s) & F_Y^{n,+}(x) \\ F_X^{n,+}(s) & F_X^{n,+}(x) \end{vmatrix} \geq 0 \qquad \text{for all } -\infty < s < x < \infty, \tag{2.40}$$

or, equivalently for $n \neq 0$, if

$$\frac{F_X^{n-1,+}(x)}{F_X^{n,+}(x)} \leq \frac{F_Y^{n-1,+}(x)}{F_Y^{n,+}(x)} \qquad \text{for all } x \in \mathbb{R}; \tag{2.41}$$

(2) $X \geq^{n,-} Y$ (or $F_X \geq^{n,-} F_Y$) if $F_X^{n,-}(x)/F_Y^{n,-}(x)$ is increasing in $x \in \mathbb{R}$ (where $a/0$ is taken to be equal to ∞ whenever $a > 0$), or, equivalently, if

$$\begin{vmatrix} F_Y^{n,-}(x) & F_Y^{n,-}(t) \\ F_X^{n,-}(x) & F_X^{n,-}(t) \end{vmatrix} \geq 0 \qquad \text{for all } -\infty < x < t < \infty, \tag{2.42}$$

or, equivalently for $n \neq 0$, if

$$\frac{F_X^{n-1,-}(x)}{F_X^{n,-}(x)} \geq \frac{F_Y^{n-1,-}(x)}{F_Y^{n,-}(x)} \qquad \text{for all } x \in \mathbb{R}. \tag{2.43}$$

It is noted that

$$\geq^{0,+} = \geq_{\mathrm{lr}}; \quad \geq^{1,+} = \geq_{\mathrm{hr}}; \quad \geq^{2,+} = \geq_{\mathrm{mrl}}; \tag{2.44}$$

$$\geq^{0,-} = \geq_{\mathrm{lr}}; \quad \geq^{1,-} = \geq_{\mathrm{rh}}. \tag{2.45}$$

2.2.3 Bivariate characterization of stochastic orders

Let $\mathcal{G}$ be a class of real–valued bivariate functions on $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. For two RVs X and Y, a stochastic ordering relation $X \succeq Y$ is said to be generated from $\mathcal{G}$ if

$$E[g(X^*, Y^*)] \geq E[g(Y^*, X^*)] \qquad \text{for all } g \in \mathcal{G} \tag{2.46}$$

for independent RVs X^* and Y^* such that $X^* =_{\mathrm{d}} X$ and $Y^* =_{\mathrm{d}} Y$ [see the univariate characterization (2.1) or (2.2)]. In this case, we denote it by $X \geq_{\mathcal{G}} Y$ (or $F_X \geq_{\mathcal{G}} F_Y$). Note that, as stated in Remarks 2.2.1 (1) and 2.2.2 (2), most stochastic ordering relations between X and Y depend only on their marginal distributions. Therefore, in the bivariate characterization (2.46), we need to consider independent RVs X^* and Y^* such that $X^* =_{\mathrm{d}} X$ and $Y^* =_{\mathrm{d}} Y$.

When a stochastic ordering relation $X \geq_{\mathcal{F}} Y$ is generated from some univariate function class $\mathcal{F}$, that is, it is of the first type, it is also generated by the corresponding bivariate function class [32], [34].

Proposition 2.2.14 Let $\mathcal{F}$ be a class of univariate functions and define

$$\mathcal{G}_{\mathcal{F}} := \{g : \mathbb{R}^2 \to \mathbb{R},\ \Delta g(\cdot, y) \in \mathcal{F} \ \text{ for each } y \in \mathbb{R}\}, \tag{2.47}$$

where we define

$$\Delta g(x, y) = g(x, y) - g(y, x), \qquad (x, y) \in \mathbb{R}^2. \tag{2.48}$$

Then, $X \geq_{\mathcal{F}} Y$ (or $F_X \geq_{\mathcal{F}} F_Y$) if and only if $X \geq_{\mathcal{G}_{\mathcal{F}}} Y$ (or $F_X \geq_{\mathcal{G}} F_Y$). □

Corollary 2.2.1 Let X and Y be RVs with CDFs F_X and F_Y:

(1) $X \geq_{\mathrm{st}} Y$ (or $F_X \geq_{\mathrm{st}} F_Y$) if and only if (2.46) holds for

$$\mathcal{G}_{\mathrm{st}} := \{g : \mathbb{R}^2 \to \mathbb{R},\ \Delta g(x, y) \text{ is increasing in } x \in \mathbb{R} \text{ for each } y \in \mathbb{R}\}; \tag{2.49}$$

(2) $X \geq_{\mathrm{cx}} Y$ (or $F_X \geq_{\mathrm{cx}} F_Y$) if and only if (2.46) holds for

$$\mathcal{G}_{\mathrm{cx}} := \{g : \mathbb{R}^2 \to \mathbb{R},\ \Delta g(x, y) \text{ is convex in } x \in \mathbb{R} \text{ for each } y \in \mathbb{R}\}; \tag{2.50}$$

(3) $X \geq_{\rm cv} Y$ (or $F_X \geq_{\rm cv} F_Y$) if and only if (2.46) holds for
$$\mathcal{G}_{\rm cv} := \{g : \mathbb{R}^2 \to \mathbb{R},\ \Delta g(x,y) \text{ is concave in } x \in \mathbb{R} \text{ for each } y \in \mathbb{R}\}; \tag{2.51}$$
(4) $X \geq_{\rm icx} Y$ (or $F_X \geq_{\rm icx} F_Y$) if and only if (2.46) holds for
$$\mathcal{G}_{\rm icx} := \{g : \mathbb{R}^2 \to \mathbb{R},\ \Delta g(x,y) \text{ is increasing and convex in } x \in \mathbb{R} \text{ for each } y \in \mathbb{R}\}; \tag{2.52}$$
(5) $X \geq_{\rm icv} Y$ (or $F_X \geq_{\rm icv} F_Y$) if and only if (2.46) holds for
$$\mathcal{G}_{\rm icv} := \{g : \mathbb{R}^2 \to \mathbb{R},\ \Delta g(x,y) \text{ is increasing and concave in } x \in \mathbb{R} \text{ for each } y \in \mathbb{R}\}. \tag{2.53}$$

For the SOs of the second type, no univariate characterization is known, while the bivariate characterizations are possible [59].

Proposition 2.2.15 Let X and Y be RVs on $\mathbb{R}$:
(1) $X \geq_{\rm lr} Y$ (or $F_X \geq_{\rm lr} F_Y$) if and only if (2.46) holds for
$$\mathcal{G}_{\rm lr} := \{g : \mathbb{R}^2 \to \mathbb{R},\ \Delta g(x,y) \geq 0 \text{ for all } x \geq y\}; \tag{2.54}$$
(2) $X \geq_{\rm hr} Y$ (or $F_X \geq_{\rm hr} F_Y$) if and only if (2.46) holds for
$$\mathcal{G}_{\rm hr} := \{g : \mathbb{R}^2 \to \mathbb{R},\ \Delta g(x,y) \text{ is increasing in } x \text{ for } x \geq y\}; \tag{2.55}$$
(3) $X \geq_{\rm rh} Y$ (or $F_X \geq_{\rm rh} F_Y$) if and only if (2.46) holds for
$$\mathcal{G}_{\rm rh} := \{g : \mathbb{R}^2 \to \mathbb{R},\ \Delta g(x,y) \text{ is increasing in } x \text{ for } x \leq y\}. \tag{2.56}$$
□

2.3 Applications in Reliability Theory

2.3.1 Notions of aging

Various aging notions in reliability theory are systematically understood by considering SOs. Let X be a nonnegative–valued RV with CDF F_X, representing the lifetime (or failure time) of an item (or a system). F_X is called the *life(time) distribution function* (or *failure time distribution function*), and its complementary function $\overline{F}_X := 1 - F_X$ is the *reliability function* (or *survival function*). Further, for a nonnegative real number $t \in \mathbb{R}_+ := [0, \infty)$, we let $[X - t|\ X > t]$ be the *residual life* (or *remaining life*) (RL) of age t, that is, the RV $X - t$ conditional on the event $\{X > t\}$, provided that $\overline{F}_X(t) = 1 - F_X(t) = P(X > t) > 0$. Let $\succeq$ be some SO defined on a collection of nonnegative RVs. The first kind of *positive* [*negative*] *aging* notions could be defined by comparing the life time of a *new* item with those of *used* ones with various ages: A nonnegative valued RV X (or its CDF F_X) could be said to be (or to have the property of) *new better* [*worse*] *than used* in the sense of SO $\succeq$ if
$$X \succeq [\preceq]\ [X - t|\ X > t] \qquad \text{for all } t \in \mathbb{R}_+. \tag{2.57}$$

Definition 2.3.1 Let X be a nonnegative–valued RV with CDF F_X.

(1) X (or F_X) is said to be (or to have the property of) *new better [worse] than used* (NBU [NWU]) if it is new better [worse] than used in the sense of OSO:

$$X \geq_{\text{st}} [\leq_{\text{st}}] \; [X - t | \; X > t] \qquad \text{for all } t \in \mathbb{R}_+; \tag{2.58}$$

(2) X (or F_X) is said to be (or to have the property of) *new better [worse] than used in expectation* (NBUE [NWUE]) if it is new better [worse] than used in the sense of the usual ordering in expectations:

$$E[X] \geq [\leq] \; E[X - t | \; X > t] \qquad \text{for all } t \in \mathbb{R}_+. \tag{2.59}$$

By definition,

$$\text{NBU [NWU]} \longrightarrow \text{NBUE [NWUE]}. \tag{2.60}$$

Since

$$\begin{aligned} P([X - t | \; X > t] > x) &= P(X - t > x | \; X > t) \\ &= \frac{P(X > t + x)}{P(X > t)} \\ &= \frac{\overline{F}_X(t + x)}{\overline{F}_X(t)}, \qquad t, \; x \in \mathbb{R}_+, \end{aligned} \tag{2.61}$$

and recall that the MRLF of X is given by

$$m_X(t) := \frac{\int_t^\infty \overline{F}_X(u)du}{\overline{F}_X(t)}, \qquad t \in \mathbb{R}_+, \tag{2.62}$$

we have the following characterizations.

Proposition 2.3.1 Let X be a nonnegative–valued RV with CDF F_X.
(1) X (or F_X) is NBU [NWU] if and only if $\log \overline{F}_X(x)$ is *sub–additive [super–additive]* in x on the set $\{x : \overline{F}_X(x) > 0\}$, that is,

$$\overline{F}_X(t + x) \leq [\geq] \; \overline{F}_X(x)\overline{F}_X(t) \qquad \text{for all } x, t \in \mathbb{R}_+; \tag{2.63}$$

(2) X (or F_X) is NBUE [NWUE] if and only if

$$\int_0^\infty \overline{F}_X(u)du \geq [\leq] \; \frac{\int_t^\infty \overline{F}_X(u)du}{\overline{F}_X(t)} \qquad \text{for all } t \in \mathbb{R}_+. \tag{2.64}$$

□

The second kind of positive [negative] aging notions could be defined by comparing the residual life at all ages: A non–negative valued RV X (or its CDF F_X) could be said to be (or to have the property of) *decreasing [increasing] residual life* in the sense of SO $\succeq$ if

$$[X - s | \; X > s] \succeq [\preceq] \; [X - t | \; X > t] \qquad \text{for all } 0 \leq s < t < \infty. \tag{2.65}$$

Definition 2.3.2 Let X be a nonnegative–valued RV with CDF F_X.

(1) X (or F_X) is said to be (or to have the property of) *increasing* [*decreasing*] *likelihood ratio* (ILR [DLR]) if it is decreasing [increasing] residual life in the sense of LRO:

$$[X-s|\ X>s] \geq_{\mathrm{lr}} [\leq_{\mathrm{lr}}]\ [X-t|\ X>t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.66}$$

(2) X (or F_X) is said to be (or to have the property of) *increasing* [*decreasing*] *hazard rate* (IHR [DHR]) if it is decreasing [increasing] residual life in the sense of HRO:

$$[X-s|\ X>s] \geq_{\mathrm{hr}} [\leq_{\mathrm{hr}}]\ [X-t|\ X>t] \quad \text{for all } 0 \leq s < t < \infty \tag{2.67}$$

{in reliability theory, it is more commonly said to be (or to have the property of) *increasing* [*decreasing*] *failure rate* (IFR [DFR])};

(3) X (or F_X) is said to be (or to have the property of) *decreasing* [*increasing*] *reversed hazard rate* (DRHR [IRHR]) if it is decreasing [increasing] residual life in the sense of RHRO:

$$[X-s|\ X>s] \geq_{\mathrm{rh}} [\leq_{\mathrm{rh}}]\ [X-t|\ X>t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.68}$$

(4) X (or F_X) is said to be (or to have the property of) *decreasing* [*increasing*] *mean residual life* (DMRL [IMRL]) if it is decreasing [increasing] residual life in the sense of MRLO:

$$[X-s|\ X>s] \geq_{\mathrm{mrl}} [\leq_{\mathrm{mrl}}]\ [X-t|\ X>t] \quad \text{for all } 0 \leq s < t < \infty. \tag{2.69}$$

Figure 2.3 depicts the conceptual picture of the relations.

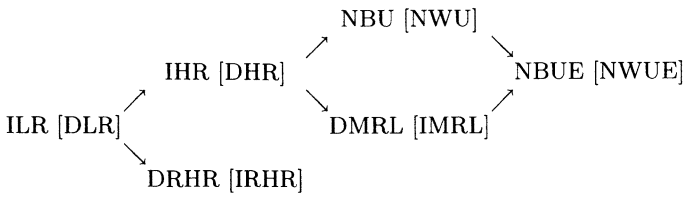


Fig. 2.3. Positive and negative aging

Proposition 2.3.2 (ILR) For a nonnegative–valued RV X with CDF F_X, if X has a PDF f_X, then the following are equivalent:
(1) X (or F_X) is ILR, that is,

$$[X-s|\ X>s] \geq_{\mathrm{lr}} [X-t|\ X>t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.70}$$

(2) f_X is a PF_2 function;
(3) f_X is log–concave, that is, $\log f_X(x)$ is concave in x on the set $\{x \in \mathbb{R}_+ : f_X(x) > 0\}$. □

Proposition 2.3.3 (ILR) For a nonnegative–valued RV X with CDF F_X, the following are equivalent:
(1) X (or F_X) is ILR, that is,

$$[X-s|\ X>s] \geq_{\mathrm{lr}} [X-t|\ X>t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.71}$$

(2) X (or F_X) is new better than used in the sense of LRO, that is,

$$X \geq_{\mathrm{lr}} [X - t|\ X > t] \qquad \text{for all } t \in \mathbb{R}_+; \tag{2.72}$$

(3) It holds that

$$X + s \leq_{\mathrm{lr}} X + t \qquad \text{for all } 0 \leq s < t < \infty. \tag{2.73}$$

□

Proposition 2.3.4 (IHR, DHR) For a nonnegative–valued RV X with CDF F_X, the following are equivalent:
(1) X (or F_X) is IHR [DHR], that is,

$$[X-s|\ X > s] \geq_{\mathrm{hr}} [\leq_{\mathrm{hr}}]\ [X-t|\ X > t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.74}$$

(2) $\overline{F}_X(t+x)/\overline{F}_X(t)$ is decreasing [increasing] in $t \in \mathbb{R}_+$ for each $x \in \mathbb{R}_+$;
(3) $\overline{F}_X$ is a PF_2 function;
(4) $\overline{F}_X$ is log–concave [log–convex], that is, $\log \overline{F}_X(x)$ is concave [convex] in x on the set $\{x \in \mathbb{R}_+ : \overline{F}_X(x) > 0\}$.
Furthermore, if X has a PDF f_X so that it has HRF h_X, the following condition is also equivalent to the above ones:
(5) The HRF $h_X(x)$ is increasing [decreasing] in x on the set $\{x \in \mathbb{R}_+ : \overline{F}_X(x) > 0\}$. □

The characterization (5) of the above proposition justifies the term "increasing [decreasing] hazard rate."

Proposition 2.3.5 (IHR, DHR) For a nonnegative–valued RV X with CDF F_X, the following are equivalent:
(1) X (or F_X) is IHR [DHR], that is,

$$[X-s|\ X > s] \geq_{\mathrm{hr}} [\leq_{\mathrm{hr}}]\ [X-t|\ X > t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.75}$$

(2) X (or F_X) is decreasing [increasing] residual life in the sense of OSO, that is,

$$[X-s|\ X > s] \geq_{\mathrm{st}} [\leq_{\mathrm{st}}]\ [X-t|\ X > t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.76}$$

(3) X (or F_X) is new better [worse] than used in the sense of the hazard rate order, that is,

$$X \geq_{\mathrm{hr}} [\leq_{\mathrm{hr}}]\ [X - t|\ X > t] \qquad \text{for all } t \in \mathbb{R}_+; \tag{2.77}$$

(4) It holds that

$$X + s \leq_{\mathrm{hr}} X + t \qquad \text{for all } 0 \leq s < t < \infty. \tag{2.78}$$

□

Usually, condition (2) of the above proposition is used as the definition of IHR [DHR].

Proposition 2.3.6 (DRHR, IRHR) For a nonnegative–valued RV X with CDF F_X, the following are equivalent:
(1) X (or F_X) is decreasing [increasing] reversed hazard rate, that is,

$$[X-s|\ X > s] \geq_{\mathrm{rh}} [\leq_{\mathrm{rh}}]\ [X-t|\ X > t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.79}$$

(2) F_X is a PF_2 function;
(3) F_X is log–concave [log–convex], that is, $\log F_X(x)$ is concave [convex] in x on the set $\{x \in \mathbb{R}_+ : F_X(x) > 0\}$.
Furthermore, if X has a PDF f_X so that it has RHRF r_X, the following condition is also equivalent to the above ones:
(4) The RHRF $r_X(x)$ is decreasing [increasing] in x on the set $\{x \in \mathbb{R}_+ : F_X(x) > 0\}$. □

The characterization (4) of the above proposition justifies the term "decreasing [increasing] reversed hazard rate."

Proposition 2.3.7 (DRHR, IRHR) For a nonnegative–valued RV X with CDF F_X, the following are equivalent:
(1) X (or F_X) is decreasing [increasing] reversed hazard rate, that is,

$$[X-s|\ X > s] \geq_{\mathrm{rh}} [\leq_{\mathrm{rh}}]\ [X-t|\ X > t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.80}$$

(2) X (or F_X) is new better [worse] than used in the sense of RHRO, that is,

$$X \geq_{\mathrm{rh}} [\leq_{\mathrm{rh}}]\ [X - t|\ X > t] \qquad \text{for all } t \in \mathbb{R}_+; \tag{2.81}$$

(3) It holds that

$$X + s \leq_{\mathrm{rh}} X + t \qquad \text{for all } 0 \leq s < t < \infty. \tag{2.82}$$

□

Proposition 2.3.8 (DMRL, IMRL) For a nonnegative–valued RV X with CDF F_X, X (or F_X) is DMRL [IMRL] if and only if the mean residual life of X at age t is decreasing [increasing] in $t \in \mathbb{R}_+$:

$$E[X-s|\ X > s] \geq [\leq]\ E[X-t|\ X > t] \qquad \text{for all } 0 \leq s < t < \infty, \tag{2.83}$$

or equivalently, $m_X(t) = \int_t^\infty \overline{F}_X(u)du\,/\overline{F}_X(t)$ is decreasing [increasing] in t on the set $\{t \in \mathbb{R}_+ : \overline{F}_X(t) > 0\}$. □

The characterization of the above proposition justifies the term "decreasing [increasing] mean residual life."

Proposition 2.3.9 (DMRL, IMRL) For a nonnegative–valued RV X with CDF F_X, the following are equivalent:
(1) X (or F_X) is DMRL [IMRL], that is,

$$[X-s|\ X > s] \geq_{\mathrm{mrl}} [X-t|\ X > t] \qquad \text{for all } 0 \leq s < t < \infty; \tag{2.84}$$

(2) X (or F_X) is decreasing [increasing] residual life in the sense of ICxO, that is,

$$[X-s|\ X > s] \geq_{\mathrm{icx}} [\leq_{\mathrm{icx}}]\ [X-t|\ X > t] \quad \text{for all } 0 \leq s < t < \infty; \tag{2.85}$$

(3) X (or F_X) is new better than used in the sense of MRLO, that is,

$$X \geq_{\mathrm{mrl}} [X - t|\ X > t] \qquad \text{for all } t \in \mathbb{R}_+; \tag{2.86}$$

(4) It hold that

$$X + s \leq_{\mathrm{mrl}} X + t \qquad \text{for all } 0 \leq s < t < \infty. \tag{2.87}$$

□

2.3.2 Useful stochastic inequalities in reliability theory

The following results are generalizations of the convolution properties (C) of SOs stated in Proposition 2.2.12.

Proposition 2.3.10 (convolution) Let (X_i, Y_i), $i = 1, \cdots, n$ be independent pairs of RVs.

(1) For any SO $\succeq$ of $\geq_{\mathrm{st}}$, $\geq_{\mathrm{cx}}$, $\geq_{\mathrm{cv}}$, $\geq_{\mathrm{icx}}$, and $\geq_{\mathrm{icv}}$,

$$X_i \succeq Y_i,\ i = 1, \cdots, n \Longrightarrow \sum_{i=1}^{n} X_i \succeq \sum_{i=1}^{n} Y_i; \tag{2.88}$$

(2) If all of X_i, $i = 1, \cdots, n$ and Y_i, $i = 1, \cdots, n$ (except possibly one X_k and one Y_l $(k \neq l)$) have PF_2 PDFs (so that these are ILR RVs), then

$$X_i \geq_{\mathrm{lr}} Y_i,\ i = 1, \cdots, n \Longrightarrow \sum_{i=1}^{n} X_i \geq_{\mathrm{lr}} \sum_{i=1}^{n} Y_i; \tag{2.89}$$

(3) If all of X_i, $i = 1, \cdots, n$ and Y_i, $i = 1, \cdots, n$ have PF_2 SFs (so that these are IHR RVs), then

$$X_i \geq_{\mathrm{hr}} Y_i,\ i = 1, \cdots, n \Longrightarrow \sum_{i=1}^{n} X_i \geq_{\mathrm{hr}} \sum_{i=1}^{n} Y_i; \tag{2.90}$$

(4) If all of X_i, $i = 1, \cdots, n$ and Y_i, $i = 1, \cdots, n$ have PF_2 CDFs (so that these RVs are decreasing reversed hazard rate), then

$$X_i \geq_{\mathrm{rh}} Y_i,\ i = 1, \cdots, n \Longrightarrow \sum_{i=1}^{n} X_i \geq_{\mathrm{rh}} \sum_{i=1}^{n} Y_i; \tag{2.91}$$

(5) If all of X_i, $i = 1, \cdots, n$ and Y_i, $i = 1, \cdots, n$ have PF_2 SFs (so that these are IHR RVs), then

$$X_i \geq_{\mathrm{mrl}} Y_i,\ i = 1, \cdots, n \Longrightarrow \sum_{i=1}^{n} X_i \geq_{\mathrm{mrl}} \sum_{i=1}^{n} Y_i. \tag{2.92}$$

□

In the proofs of the next proposition, the following observations are keys: for any $a \in \mathbb{R}$,

(1) $\max\{x, a\}$ is increasing and convex in $x \in \mathbb{R}$;

(2) $\min\{x, a\}$ is increasing and concave in $x \in \mathbb{R}$.

Proposition 2.3.11 (maximum and minimum) Let (X_i, Y_i), $i = 1, \cdots, m$ be independent pairs of RVs.

(1) For any SO $\succeq$ of $\geq_{\mathrm{st}}$, $\geq_{\mathrm{cx}}$, and $\geq_{\mathrm{icx}}$, it holds that

$$X_i \succeq Y_i,\ i = 1, \cdots, n \Longrightarrow \max\{X_1, \cdots, X_n\} \succeq \max\{Y_1, \cdots, Y_n\}; \tag{2.93}$$

(2) For any SO $\succeq$ of $\geq_{\mathrm{st}}$, $\geq_{\mathrm{cv}}$, and $\geq_{\mathrm{icv}}$, it holds that

$$X_i \succeq Y_i,\ i = 1, \cdots, n \Longrightarrow \min\{X_1, \cdots, X_n\} \succeq \min\{Y_1, \cdots, Y_n\}; \tag{2.94}$$

(3) It holds that

$$X_i \geq_{\mathrm{hr}} Y_i,\ i = 1, \cdots, n \Longrightarrow \min\{X_1, \cdots, X_n\} \geq_{\mathrm{hr}} \min\{Y_1, \cdots, Y_n\}; \tag{2.95}$$

(4) It holds that

$$X_i \geq_{\mathrm{rh}} Y_i,\ i = 1, \cdots, n \Longrightarrow \max\{X_1, \cdots, X_n\} \geq_{\mathrm{rh}} \max\{Y_1, \cdots, Y_n\}. \tag{2.96}$$

□

For RVs X_i, $i = 1, \cdots, n$, let

$$X_{(1)} \leq \cdots \leq X_{(n)} \tag{2.97}$$

denote their increasing alignment, that is, $X_{(k)}$, $k = 1, \cdots, n$ are the k–th order statistics of X_i, $i = 1, \cdots, n$. The following results are partial generalizations of Proposition 2.3.11 (3), (4).

Proposition 2.3.12 (order statistics) Let (X_i, Y_i), $i = 1, \cdots, n$ be independent pairs of RVs. Suppose that both of X_i, $i = 1, \cdots, n$ and Y_i, $i = 1, \cdots, n$ are identically distributed. Then:

(1) It holds that

$$X_i \geq_{\mathrm{hr}} Y_i,\ i = 1, \cdots, n \Longrightarrow X_{(k)} \geq_{\mathrm{hr}} Y_{(k)},\ k = 1, \cdots, n; \tag{2.98}$$

(2) It holds that

$$X_i \geq_{\mathrm{rh}} Y_i,\ i = 1, \cdots, n \Longrightarrow X_{(k)} \geq_{\mathrm{rh}} Y_{(k)},\ k = 1, \cdots, n. \tag{2.99}$$

□

2.3.3 Stochastic comparisons of system reliabilities

Consider a *reliability system* consisting of n *components* $C = \{c_1, \cdots, c_n\}$ whose *structure function* is given by

$$\phi : \{0,1\}^n := \underbrace{\{0,1\} \times \cdots \times \{0,1\}}_{n \text{ times}} \to \{0,1\}. \tag{2.100}$$

It is assumed that

(1) $\phi(0, \cdots, 0) = 0$;

(2) $\phi(1, \cdots, 1) = 1$;

(3) (Monotonicity) For $(u_1, \cdots, u_n)$, $(v_1, \cdots, v_n) \in \{0,1\}^n$,

$$u_i \leq v_i,\ i = 1, \cdots, n \Longrightarrow \phi(u_1, \cdots, u_n) \leq \phi(v_1, \cdots, v_n). \tag{2.101}$$

For $i = 1, \cdots, n$, let $\{S_i(t);\ t \in \mathbb{R}_+\}$ be a decreasing, right–continuous, $\{0,1\}$–valued stochastic process representing the state of component c_i at time t, that is,

$$S_i(t) = \begin{cases} 1 \text{ if component } c_i \text{ is functioning at time } t; \\ 0 \text{ if component } c_i \text{ is failed at time } t. \end{cases} \tag{2.102}$$

Similarly, let $\{S(t);\ t \in \mathbb{R}_+\}$ be a decreasing, right–continuous, $\{0,1\}$–valued stochastic process representing the state of the system at time t, that is,

$$S(t) = \begin{cases} 1 \text{ if the system is functioning at time } t; \\ 0 \text{ if the system is failed at time } t. \end{cases} \tag{2.103}$$

By definition, we have

$$S(t) = \phi(S_1(t), \cdots, S_n(t)), \qquad t \in \mathbb{R}_+. \tag{2.104}$$

Let X_i, $i = 1, \cdots, n$ be nonnegative RVs representing the *lifetime* (or *failure time*) of component c_i. Since

$$\{X_i > t\} \Longleftrightarrow \{S_i(t) = 1\}, \tag{2.105}$$

we have

$$\begin{aligned} P(S_i(t) = 1) &= P(X_i > t) = \overline{F}_{X_i}(t) \\ &= 1 - F_{X_i}(t), \qquad t \in \mathbb{R}_+ \end{aligned} \tag{2.106}$$

Similarly, if we let X be nonnegative RVs representing the *lifetime* (or *failure time*) of the system, then, since

$$\{X > t\} \Longleftrightarrow \{S(t) = 1\}, \tag{2.107}$$

we have

$$P(S(t) = 1) = P(X > t) = \overline{F}_X(t) = 1 - F_X(t). \tag{2.108}$$

If we define a function

$$\tau : \mathbb{R}_+^n := \underbrace{\mathbb{R}_+ \times \cdots \times \mathbb{R}_+}_{n \text{ times}} \to \mathbb{R}_+ \tag{2.109}$$

by

$$\begin{aligned} \tau(t_1, \cdots, t_n) &:= \sup\left\{t \in \mathbb{R}_+ : \phi\left(1_{[0,t_i)}(t), \cdots, 1_{[0,t_i)}(t)\right) = 1\right\} \\ &= \inf\left\{t \in \mathbb{R}_+ : \phi\left(1_{[0,t_i)}(t), \cdots, 1_{[0,t_i)}(t)\right) = 0\right\}, \end{aligned} \tag{2.110}$$

then it determines the lifetime of the system by the lifetimes of components, that is,

$$X = \tau(X_1, \cdots, X_n), \tag{2.111}$$

and we call τ the *system lifetime function* of ϕ. It is noted that, since ϕ is an increasing function, τ is also an increasing function.

Furthermore, if the lifetimes X_i, $i = 1, \cdots, n$ of components are mutually independent then the reliability of the system at time $t \in \mathbb{R}_+$ can be given by

$$\begin{aligned} \overline{F}_X(t) &= P(X > t) \\ &= P(\tau(X_1, \cdots, X_n) > t) \\ &= P(\phi(S_1(t), \cdots, S_n(t)) = 1) \\ &= h\left(\overline{F}_{X_1}(t), \cdots, \overline{F}_{X_n}(t)\right), \end{aligned} \tag{2.112}$$

where $\overline{F}_{X_i}$ $(i = 1, \cdots, n)$ is the SF of RV X_i, and

$$h : [0,1]^n := \underbrace{[0,1] \times \cdots \times [0,1]}_{n \text{ times}} \to [0,1] \tag{2.113}$$

is so-called the *system reliability function* of ϕ.

An important example of a (monotone) reliability system is the *k–out–of–n system* ($k \in \{1, \cdots, n\}$), whose structure function $\phi_{k|n} : \{0,1\}^n \to \{0,1\}$ is defined by

$$\phi_{k|n}(s_1, \cdots, s_n) = \begin{cases} 1 \text{ if } \sum_{i=1}^{n} s_i \geq k; \\ 0 \text{ if } \sum_{i=1}^{n} s_i < k, \end{cases} \qquad (s_1, \cdots, s_n) \in \{0,1\}^n. \quad (2.114)$$

In particular, the n–out–of–n and 1–out–of–n systems are called the *series system* and the *parallel system*, and their structure functions $\phi_{n|n}$ and $\phi_{1|n}$ are given by

$$\begin{aligned} \phi_{n|n}(s_1, \cdots, s_n) &= \min\{s_1, \cdots, s_n\} \\ &= \prod_{i=1}^{n} s_i, \qquad (s_1, \cdots, s_n) \in \{0,1\}^n; \end{aligned} \quad (2.115)$$

$$\begin{aligned} \phi_{1|n}(s_1, \cdots, s_n) &= \max\{s_1, \cdots, s_n\} \\ &= 1 - \prod_{i=1}^{n}(1 - s_i), \qquad (s_1, \cdots, s_n) \in \{0,1\}^n. \end{aligned} \quad (2.116)$$

The system lifetime function $\tau_{k|n}$ of k–out–of–n system $\phi_{k|n}$ is given by

$$\tau_{k|n}(t, \cdots, t_n) = t_{(n-k+1)}, \qquad (t, \cdots, t_n) \in \mathbb{R}_+^n, \quad (2.117)$$

where $t_{(1)} \leq \cdots \leq t_{(n)}$ are the increasing alignments of t_i, $i = 1, \cdots, n$. In particular,

$$\tau_{n|n}(t_1, \cdots, t_n) = \min\{t_1, \cdots, t_n\}, \qquad (t_1, \cdots, t_n) \in \mathbb{R}_+^n; \quad (2.118)$$

$$\tau_{1|n}(t_1, \cdots, t_n) = \max\{t_1, \cdots, t_n\}, \qquad (t_1, \cdots, t_n) \in \mathbb{R}_+^n. \quad (2.119)$$

The system reliability function $h_{k|n}$ of the k–out–of–n system $\phi_{k|n}$ is given by

$$\begin{aligned} &h_{k|n}(p_1, \cdots, p_n) \\ &= \sum_{\{(s_1, \cdots, s_n) \in \{0,1\}^n : \sum_{i=1}^{n} s_i \geq k\}} \prod_{i=1}^{n} p_i^{s_i}(1 - p_i)^{1 - s_i}, \\ &\qquad (p_1, \cdots, p_n) \in [0,1]^n. \end{aligned} \quad (2.120)$$

In particular,

$$h_{n|n}(p_1, \cdots, p_n) = \prod_{i=1}^{n} p_i, \qquad (p_1, \cdots, p_n) \in [0,1]^n; \quad (2.121)$$

$$h_{1|n}(p_1, \cdots, p_n) = 1 - \prod_{i=1}^{n}(1 - p_i), \qquad (p_1, \cdots, p_n) \in [0,1]^n. \quad (2.122)$$

Since for any reliability system with a monotone structure function ϕ, the system lifetime function τ is increasing. Hence, we have:

Proposition 2.3.13 Let (X_i, Y_i), $i = 1, \cdots, n$ be independent pairs of nonnegative RVs representing the lifetimes of component c_i. Then, for any (monotone) reliability system with structure function ϕ,

$$X_i \geq_{\text{st}} Y_i, \ i = 1, \cdots, n \Longrightarrow \tau(X_1, \cdots, X_n) \geq_{\text{st}} \tau(Y_1, \cdots, Y_n). \quad (2.123)$$

□

The following proposition is a direct corollary of Proposition 2.3.11.

Proposition 2.3.14 (series and parallel systems) Let (X_i, Y_i), $i = 1, \cdots, m$ be independent pairs of nonnegative RVs representing the lifetimes of component c_i.

(1) For any SO $\succeq$ of $\geq_{\mathrm{st}}$, $\geq_{\mathrm{cx}}$, and $\geq_{\mathrm{icx}}$, it holds that

$$X_i \succeq Y_i,\ i = 1, \cdots, n \Longrightarrow \tau_{1|n}(X_1, \cdots, X_n) \succeq \tau_{1|n}(Y_1, \cdots, Y_n); \tag{2.124}$$

(2) For any SO $\succeq$ of $\geq_{\mathrm{st}}$, $\geq_{\mathrm{cv}}$, and $\geq_{\mathrm{icv}}$, it holds that

$$X_i \succeq Y_i,\ i = 1, \cdots, n \Longrightarrow \tau_{n|n}(X_1, \cdots, X_n) \succeq \tau_{n|n}(Y_1, \cdots, Y_n); \tag{2.125}$$

(3) It holds that

$$X_i \geq_{\mathrm{hr}} Y_i,\ i = 1, \cdots, n \Longrightarrow \tau_{n|n}(X_1, \cdots, X_n) \geq_{\mathrm{hr}} \tau_{n|n}(Y_1, \cdots, Y_n); \tag{2.126}$$

(4) It holds that

$$X_i \geq_{\mathrm{rh}} Y_i,\ i = 1, \cdots, n \Longrightarrow \tau_{1|n}(X_1, \cdots, X_n) \geq_{\mathrm{rh}} \tau_{1|n}(Y_1, \cdots, Y_n). \tag{2.127}$$

□

Proposition 2.3.15 (k–out–of–n system) Let (X_i, Y_i), $i = 1, \cdots, n$ be independent pairs of nonnegative RVs. Suppose that both of X_i, $i = 1, \cdots, n$ and Y_i, $i = 1, \cdots, n$ are identically distributed. Then:

(1) It holds that, for $k = 1, \cdots, n$,

$$X_i \geq_{\mathrm{hr}} Y_i,\ i = 1, \cdots, n \Longrightarrow \tau_{k|n}(X_1, \cdots, X_n) \geq_{\mathrm{hr}} \tau_{k|n}(Y_1, \cdots, Y_n); \tag{2.128}$$

(2) It holds that, for $k = 1, \cdots, n$,

$$X_i \geq_{\mathrm{rh}} Y_i,\ i = 1, \cdots, n \Longrightarrow \tau_{k|n}(X_1, \cdots, X_n) \geq_{\mathrm{rh}} \tau_{k|n}(Y_1, \cdots, Y_n). \tag{2.129}$$

□

2.3.4 Redundancy improvement

In this section, we consider some redundancy improvement problems of reliability systems. The next proposition implies that component–wise hot standby redundancy is superior to system–wise hot standby redundancy, in the sense of an SO.

Proposition 2.3.16 Let X_i, $i = 1, \cdots, n$ be a collection of mutually independent nonnegative RVs representing the lifetimes of component c_is of a reliability system. Further, let Y_i, $i = 1, \cdots, n$ be another collection of nonnegative RVs representing lifetimes of components which are independent of each other as well as of X_is.

(1) For any reliability with monotone structure function ϕ,

$$\tau(\max\{X_1, Y_1\}, \cdots, \max\{X_n, Y_n\}) \geq_{\mathrm{st}} \max\{\tau(X_1, \cdots, X_n), \tau(Y_1, \cdots, Y_n)\}; \tag{2.130}$$

(2) If X_i, Y_i, $i = 1, \cdots, n$ are identically distributed RVs, then, for any k–out–of–n system ($k \in \{1, \cdots, n\}$),

$$\tau_{k|n}(\max\{X_1, Y_1\}, \cdots, \max\{X_n, Y_n\}) \geq_{\mathrm{lr}} \max\{\tau_{k|n}(X_1, \cdots, X_n), \tau_{k|n}(Y_1, \cdots, Y_n)\}. \tag{2.131}$$

□

Next, we consider a situation where an additional component could be utilized to improve the lifetime of the system. Two cases are considered: In the first case, only one standby component, which is common to all positions of the system, is available, while, in the second case, all standby components, each of which is specific to a single position and stochastically equivalent to the original one, are available. For the proofs of the following results, see Boland and Proschan [9], where the bivariate characterizations of SOs given in Section 2.2.3 play the key parts.

Proposition 2.3.17 Let X_i, $i = 1, \cdots, n$ be mutually independent nonnegative RVs representing the lifetimes of component c_is of a reliability system. Further, let Y be a nonnegative RV representing the lifetime of a common component which could be utilized for a *hot standby* (or *parallel*) *redundancy* in any position of components. If X_is are ordered as $X_1 \geq_{\mathrm{st}} \cdots \geq_{\mathrm{st}} X_n$, then, for any $k = 1, \cdots, n$, the lifetime of the k–out–of–n system with a hot standby redundancy in component c_i

$$\tau_{k|n}(X_1, \cdots, X_{i-1}, \max\{X_i, Y\}, X_{i+1}, \cdots, X_n) \tag{2.132}$$

is increasing in $i \in \{1, \cdots, n\}$ in the sense of OSO $\geq_{\mathrm{st}}$. □

Proposition 2.3.18 Let X_i, $i = 1, \cdots, n$ be a collection of mutually independent nonnegative RVs representing the lifetimes of component c_is of a reliability system. Further, let Y_i, $i = 1, \cdots, n$ be another collection of nonnegative RVs representing lifetimes of components which are independent of each other as well as of X_is. It is assumed that only one of them could be utilized for a hot standby (or parallel) redundancy in the position of corresponding component c_i, respectively. If X_is are ordered as $X_1 \geq_{\mathrm{st}} \cdots \geq_{\mathrm{st}} X_n$ and $X_i =_{\mathrm{d}} Y_i$, $i = 1, \cdots, n$, then:

(1) the lifetime of the *series system* with a hot standby redundancy in component c_i

$$\tau_{n|n}(X_1, \cdots, X_{i-1}, \max\{X_i, Y_i\}, X_{i+1}, \cdots, X_n) \tag{2.133}$$

is increasing in $i \in \{1, \cdots, n\}$ in the sense of OSO $\geq_{\mathrm{st}}$;

(2) the lifetime of the *parallel system* with a hot standby redundancy in component c_i

$$\tau_{1|n}(X_1, \cdots, X_{i-1}, \max\{X_i, Y_i\}, X_{i+1}, \cdots, X_n) \tag{2.134}$$

is decreasing in $i \in \{1, \cdots, n\}$ in the sense of OSO $\geq_{\mathrm{st}}$. □

Proposition 2.3.19 Let X_i, $i = 1, \cdots, n$ be mutually independent nonnegative RVs representing the lifetimes of component c_is of a reliability system. Further, let Y be a nonnegative RV representing the lifetime of a common component which could be utilized for a *cold standby redundancy* in any position of components. If X_is are ordered as $X_1 \geq_{\mathrm{lr}} \cdots \geq_{\mathrm{lr}} X_n$, then:

(1) the lifetime of the *series system* with a cold standby redundancy in component c_i

$$\tau_{n|n}(X_1, \cdots, X_{i-1}, X_i + Y, X_{i+1}, \cdots, X_n) \tag{2.135}$$

is increasing in $i \in \{1, \cdots, n\}$ in the sense of OSO $\geq_{\mathrm{st}}$;

(2) the lifetime of the *parallel system* with a cold standby redundancy in component c_i

$$\tau_{1|n}(X_1, \cdots, X_{i-1}, X_i + Y, X_{i+1}, \cdots, X_n) \tag{2.136}$$

is decreasing in $i \in \{1, \cdots, n\}$ in the sense of OSO $\geq_{\mathrm{st}}$. □

Proposition 2.3.20 Let X_i, $i = 1, \cdots, n$ be a collection of mutually independent nonnegative RVs representing the lifetimes of component c_i of reliability system. Further, let Y_i, $i = 1, \cdots, n$ be another collection of nonnegative RVs representing lifetimes of components which are independent of each other as well as X_is. It is assumed that only one of them could be utilized for a cold standby redundancy in the position of component c_i, respectively. If X_is are ordered as $X_1 \geq_{\mathrm{lr}} \cdots \geq_{\mathrm{lr}} X_n$ and $X_i =_{\mathrm{d}} Y_i$, $i = 1, \cdots, n$, then the lifetime of *parallel system* with a cold standby redundancy in component c_i

$$\tau_{1|n}(X_1, \cdots, X_{i-1}, X_i + Y_i, X_{i+1}, \cdots, X_n) \tag{2.137}$$

is decreasing in $i \in \{1, \cdots, n\}$ in the sense of OSO $\geq_{\mathrm{st}}$. □

2.3.5 Stochastic comparisons of maintenance policies

This section is devoted to the stochastic comparison of maintenance policies. Although many maintenance policies have been proposed in the reliability literatures, we consider only five classical ones: replacement–upon–failure policy, age replacement policy, block replacement policy, minimal repair policy, and minimal repair policy with block replacement.

2.3.5.1 Replacements upon failures. Consider an item (or a system) operating continuously in time. Suppose that it is *replaced by a new item* (or it is *perfectly or maximally repaired* to make the item "as good as new") only when it fails (*failure replacement* or *corrective replacement*). Such a *maintenance policy* is called *replacement–upon–failure policy* (RUFP) We assume that the required time for a replacement is negligible. Let X_i, $i = 1, 2, \cdots$ be the *time to failure* (TTF) after the $(i-1)$–st replacement, and assume that they are independently and identically distributed (IID) nonnegative RVs with common CDF F_X.

Denote the the k–th replacement time by

$$S_0 := 0; \quad S_k := \sum_{i=1}^{k} X_i, \qquad k = 1, 2, \cdots, \tag{2.138}$$

and the number of failures (replacements) during the time interval $[0, t]$ by

$$N_X(t) := \sup\{k \in \mathbb{Z}_+ : S_k \leq t\}, \ t \in \mathbb{R}_+. \tag{2.139}$$

Then, the counting process $\{N_X(t);\ t \in \mathbb{R}_+\}$ is a *renewal process* with inter–renewal times $\{X_i;\ i \in \mathbb{Z}_{++}\}$ and an inter–renewal CDF F_X. For a fixed $t \in \mathbb{R}_+$, the distribution of RV $N_X(t)$ is of great interest in reliability theory. Explicit formulas for its CDF, however, are not available except in some special cases, such as when F_X is an exponential distribution. Therefore, stochastic bounds on the CDF would be practically useful.

Proposition 2.3.21 Assume that F_X has no mass at the origin.

(1) If F_X is NBU [NWU] then

$$P(N_X(t) \leq n) \geq [\leq] \sum_{i=0}^{n} \frac{\{H_X(t)\}^i}{i!} e^{-H_X(t)} \quad \text{for all } t \in \mathbb{R}_+,\ n \in \mathbb{Z}_+; \tag{2.140}$$

(2) If F_X is IFR [DFR] then

$$P(N_X(t) \leq n) \geq [\leq] \sum_{i=0}^{n} \frac{\left\{(n+1)H_X\left(\frac{t}{n+1}\right)\right\}^i}{i!} e^{-(n+1)H_X\left(\frac{t}{n+1}\right)},$$

$$\text{for all } t \in \mathbb{R}_+,\ n \in \mathbb{Z}_+, \tag{2.141}$$

where $H_X(x) := -\log \overline{F}_X(x)$, $x \in \mathbb{R}_+$ is the *cumulative hazard rate function* (CHRF) of X. □

It can be shown that, if F_X is IFR [DFR],

$$\sum_{i=0}^{n} \frac{\{H_X(t)\}^i}{i!} e^{-H_X(t)} \leq [\geq] \sum_{i=0}^{n} \frac{\left\{(n+1)H_X\left(\frac{t}{n+1}\right)\right\}^i}{i!} e^{-(n+1)H_X\left(\frac{t}{n+1}\right)}$$

$$\text{for all } t \in \mathbb{R}_+,\ n \in \mathbb{Z}_+. \tag{2.142}$$

The following proposition is a direct corollary of Proposition 2.3.10 (1) for SO $\geq_{\mathrm{st}}$.

Proposition 2.3.22 If $F_X \geq_{\mathrm{st}} F_Y$, then

$$N_X(t) \leq_{\mathrm{st}} N_Y(t) \qquad \text{for all } t \in \mathbb{R}_+. \tag{2.143}$$

□

2.3.5.2 Age replacement. Suppose an item is replaced when a failure occurs (*failure replacement* or *corrective replacement*) or when it reaches the predetermined age $T \in \mathbb{R}_{++} := (0, \infty)$ (*preventive replacement*). Such a *maintenance policy* is called *age replacement policy* (ARP). For any CDF F_X and *preventive replacement age* $T \in \mathbb{R}_{++}$, define

$N_X^{\mathrm{A}}(t : T)$, $t \in \mathbb{R}_+$: the number of failures during the time interval $[0, t]$ under ARP;

$R_X^{\mathrm{A}}(t : T)$, $t \in \mathbb{R}_+$: the number of (failure or preventive) replacements during the time interval $[0, t]$ under ARP.

By definition

$$N_X^{\mathrm{A}}(t : T) \leq R_X^{\mathrm{A}}(t : T) \text{ a.s.} \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \tag{2.144}$$

Proposition 2.3.23 For any CDF F_X,

$$R_X^{\mathrm{A}}(t : S) \geq_{\mathrm{st}} R_X^{\mathrm{A}}(t : T) \qquad \text{for all } t \in \mathbb{R}_+,\ 0 < S \leq T. \tag{2.145}$$

□

Proposition 2.3.24 The following conditions are equivalent:

(1) F_X is NBU;
(2) It holds that

$$N_X^{\mathrm{A}}(t : kT) \geq_{\mathrm{st}} N_X^{\mathrm{A}}(t : T) \quad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++},\ k \in \mathbb{Z}_{++}; \tag{2.146}$$

(3) It holds that

$$N_X(t) \geq_{\mathrm{st}} N_X^{\mathrm{A}}(t : T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \tag{2.147}$$

□

Proposition 2.3.25 F_X is IFR if and only if

$$N_X^{\mathrm{A}}(t : S) \leq_{\mathrm{st}} N_X^{\mathrm{A}}(t : T) \qquad \text{for all } t \in \mathbb{R}_+,\ 0 < S \leq T. \tag{2.148}$$

□

Proposition 2.3.26 If $F_X \geq_{\mathrm{st}} F_Y$, then
(1)

$$N_X^{\mathrm{A}}(t : T) \leq_{\mathrm{st}} N_Y^{\mathrm{A}}(t : T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}; \tag{2.149}$$

(2)

$$R_X^{\mathrm{A}}(t : T) \leq_{\mathrm{st}} R_Y^{\mathrm{A}}(t : T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \tag{2.150}$$

□

2.3.5.3 Block replacement. Suppose an item is replaced upon failure (*failure replacement* or *corrective replacement*) and at the scheduled (calendar) times kT, $k = 1, 2, \cdots$ (*preventive replacement*), where $T \in \mathbb{R}_{++}$ is a predetermined fixed time interval. Such a *maintenance policy* is called *block replacement policy* (BRP). For any CDF F_X and *preventive replacement interval* $T \in \mathbb{R}_{++}$, define

$N_X^{\mathrm{B}}(t : T)$, $t \in \mathbb{R}_+$: the number of failures during the time interval $[0, t]$ under BRP;

$R_X^{\mathrm{B}}(t : T)$, $t \in \mathbb{R}_+$: the number of (failure or preventive) replacements during the time interval $[0, t]$ under BRP.

By definition

$$N_X^{\mathrm{B}}(t : T) \leq R_X^{\mathrm{B}}(t : T) \text{ a.s.} \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \tag{2.151}$$

Proposition 2.3.27 The following conditions are equivalent:
(1) F_X is NBU;
(2)

$$N_X^{\mathrm{B}}(t : kT) \geq_{\mathrm{st}} N_X^{\mathrm{B}}(t : T) \quad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++},\ k \in \mathbb{Z}_{++}; \tag{2.152}$$

(3)

$$N_X(t) \geq_{\mathrm{st}} N_X^{\mathrm{B}}(t : T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \tag{2.153}$$

□

Proposition 2.3.28 If $F_X \geq_{\mathrm{st}} F_Y$, then
(1)

$$N_X^{\mathrm{B}}(t:T) \leq_{\mathrm{st}} N_Y^{\mathrm{B}}(t:T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}; \tag{2.154}$$

(2)

$$R_X^{\mathrm{B}}(t:T) \leq_{\mathrm{st}} R_Y^{\mathrm{B}}(t:T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \tag{2.155}$$

□

2.3.5.4 Minimal repair. An item (or a system) is *minimally repaired* each time it fails, which means that the *minimal repair* does not make the item as good as new but restores its function and makes it "as good as it was just before it failed." Suppose an item with lifetime X fails at age x and then it is minimally repaired; then the time to the next failure is stochastically equals to $[X - x|\ X > x]$. It is assumed that the time required for a minimal repair is negligible. Such a maintenance policy is called a *minimal repair policy* (MRP).

Define

$N_X^{\mathrm{M}}(t)$ $(= R_X^{\mathrm{M}}(t))$, $t \in \mathbb{R}_+$: the number of failures (minimal repairs) during the time interval $[0, t]$ under MRP,

then it is well known that counting process $\{N_X^{\mathrm{M}}(t);\ t \in \mathbb{R}_+\}$ is a *non-homogeneous Poisson process* with *mean value function* (MVF) $H_X(t)$, $t \in \mathbb{R}_+$ and *intensity function* (IF) $h_X(t)$, $t \in \mathbb{R}_+$ (provided that X has a PDF f_X).

Let T_i, $i \in \mathbb{Z}_{++}$ be the length of time between the $(i-1)$–st and the i–th failure.

Proposition 2.3.29
(1) If F_X is NBU [NWU], then

$$T_1 \geq_{\mathrm{st}} [\leq_{\mathrm{st}}]\ T_i \qquad \text{for all } i \in \mathbb{Z}_{++}; \tag{2.156}$$

(2) If F_X is IFR [DFR], then

$$T_i \geq_{\mathrm{hr}} [\leq_{\mathrm{hr}}]\ T_{i+1} \qquad \text{for all } i \in \mathbb{Z}_{++}. \tag{2.157}$$

□

Proposition 2.3.30 If $F_X \geq_{\mathrm{st}} F_Y$, then

$$N_X^{\mathrm{M}}(t) \leq_{\mathrm{st}} N_Y^{\mathrm{M}}(t) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}; \tag{2.158}$$

□

2.3.5.5 Minimal repair with block replacement. Suppose an item is preventively replaced at the scheduled (calendar) times kT, $k = 1, 2, \cdots$ (*block replacement* or *periodic replacement*), while it is minimally repaired for failures which occur between them (*minimal repair*), where $T \in \mathbb{R}_{++}$ is a predetermined fixed time interval (*block replacement interval*). Such a *maintenance policy* is called a *minimal repair policy with block replacement* (MRPBR). For any CDF F_X and periodic replacement interval $T \in \mathbb{R}_{++}$, define

$N_X^{\mathrm{MB}}(t:T)$, $t \in \mathbb{R}_+$: the number of failures during the time interval $[0,t]$ under MRPBP;
$R_X^{\mathrm{MB}}(t:T)$, $t \in \mathbb{R}_+$: the number of renovations (that is, minimal repairs and (preventive) replacements) during the time interval $[0,t]$ under MRPBP.
By definition

$$N_X^{\mathrm{MB}}(t:T) \le R_X^{\mathrm{MB}}(t:T) \text{ a.s.} \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \quad (2.159)$$

Proposition 2.3.31 If $F_X \ge_{\mathrm{st}} F_Y$, then
(1)

$$N_X^{\mathrm{MB}}(t:T) \le_{\mathrm{st}} N_Y^{\mathrm{MB}}(t:T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}; \quad (2.160)$$

(2)

$$R_X^{\mathrm{MB}}(t:T) \le_{\mathrm{st}} R_Y^{\mathrm{MB}}(t:T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \quad (2.161)$$

□

2.3.5.6 Stochastic comparison of different maintenance policies.

Proposition 2.3.32 For any CDF F_X,

$$R_X^{\mathrm{B}}(t:T) \ge_{\mathrm{st}} R_X^{\mathrm{A}}(t:T) \ge_{\mathrm{st}} N_X(t) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \quad (2.162)$$

□

Proposition 2.3.33 If F_X is NBU [NWU], then
(1)

$$N_X(t) \ge_{\mathrm{st}} [\le_{\mathrm{st}}]\ N_X^{\mathrm{A}}(t:T) \quad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}; \quad (2.163)$$

(2)

$$\begin{aligned} N_X^{\mathrm{M}}(t) &\ge_{\mathrm{st}} [\le_{\mathrm{st}}]\ N_X(t) \\ &\ge_{\mathrm{st}} [\le_{\mathrm{st}}]\ N_X^{\mathrm{B}}(t:T) \quad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}; \end{aligned} \quad (2.164)$$

(3)

$$\begin{aligned} N_X^{\mathrm{M}}(t) &\ge_{\mathrm{st}} [\le_{\mathrm{st}}]\ N_X^{\mathrm{MB}}(t:T) \\ &\ge_{\mathrm{st}} [\le_{\mathrm{st}}]\ N_X^{\mathrm{B}}(t:T) \quad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \end{aligned} \quad (2.165)$$

□

Proposition 2.3.34 If F_X is NBU, then
(B.1)

$$R_X^{\mathrm{MB}}(t:T) \ge_{\mathrm{st}} R_X^{\mathrm{B}}(t:T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \quad (2.166)$$

Conversely, if F_X is NWU, then
(W.1)

$$R_X^{\mathrm{B}}(t:T) \ge_{\mathrm{st}} N_X^{\mathrm{M}}(t) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}; \quad (2.167)$$

(W.2)

$$R_X^{\mathrm{BM}}(t:T) \geq_{\mathrm{st}} N_X^{\mathrm{M}}(t) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}; \tag{2.168}$$

(W.3)

$$R_X^{\mathrm{A}}(t:T) \geq_{\mathrm{st}} N_X^{\mathrm{M}}(t) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \tag{2.169}$$

□

Proposition 2.3.35 If F_X is IFR [DFR], then

$$\begin{aligned} N_X(t) &\geq_{\mathrm{st}} [\leq_{\mathrm{st}}]\ N_X^{\mathrm{A}}(t:T) \\ &\geq_{\mathrm{st}} [\leq_{\mathrm{st}}]\ N_X^{\mathrm{B}}(t:T) \qquad \text{for all } t \in \mathbb{R}_+,\ T \in \mathbb{R}_{++}. \end{aligned} \tag{2.170}$$

□

Although all of the results are concerned with comparisons of marginal distributions of counting processes at an arbitrary single time instant, more generally, stochastic comparisons of processes themselves (that is, comparisons of arbitrary finite dimensional distributions) are possible. But, for the presentation of these results, concepts of multivariate SOs are needed; hence we have avoided them [8].

References

1. Aly, E.–E. A. A. and Kochar, S. C. (1993), "On hazard rate ordering of dependent variables," *Advances in Applied Probability*, **25**, 477–482
2. Ascher, H. and Feingold, H. (1984), Repairable Systems Reliability. Marcel Dekker, New York
3. Barlow, R. E. and Proschan, F. (1965), (with Contributions by Hunter, L. C.), Mathematical Theory of Reliability. John Wiley & Sons, New York
4. Barlow, R. E. and Proschan, F. (1975), Statistical Theory of Reliability and Life Testing: Probability Models. Holt, Rinehart and Winston, New York
5. Basu, A. P., Basu, S. K. and Mukhopadhyay, S. (1998), Frontiers in Reliability. World Scientific, Singapore
6. Bawa, V. S. (1982), "Stochastic dominance: a research bibliography," *Management Science*, **28**, 698–712
7. Block, H. W., Sampson, A. R., and Savits, T. H. (eds.) (1990), Topics in Statistical Dependence. IMS Lecture Notes–Monograph Series **16**, Institute of Mathematical Statistics, Hayward, California
8. Block, H. W. and Savits, T. H. (1994), "Comparison of maintenance policies," in Stochastic Orders and Their Applications (Shaked, M. and Shanthikumar, J. G. eds.). 463–483, Academic Press, San Diego
9. Boland, P. J. and Proschan, F. (1994), "Stochastic order in system reliability theory," in Stochastic Orders and their Applications (Shaked, M. and Shanthikumar, J. G. eds.). 485–508, Academic Press, San Diego
10. Brown, M. and Solomon, H. (1973), "Optimal issuing policies under stochastic fields lives," *Journal of Applied Probability*, **10**, 761–768
11. Dharmadhikari, S. W. and Joag–Dev, K. (1988), Unimodality, Convexity, and Applications. Academic Press, New York
12. van Doorn, E. (1981), Stochastic Monotonicity and Queueing Applications of Birth–Death Processes. Lecture Notes in Statistics **4**, Springer-Verlag, New York
13. Eaton, M. L. (1987), Lectures on Topics in Probability Inequalities. CWI Tracts **35**, Centrum voor Wiskunde en Informatica, Amsterdam
14. Fishburn, P. C. (1970), Utility Theory for Decision Making. John Wiley & Sons, New York
15. Fishburn, P. C. (1982), The Foundations of Expected Utility. D. Reidel, Dordrecht
16. Fishburn, P. C. (1988), Nonlinear Preference and Utility Theory. John Hopkins University Press, Baltimore, Maryland
17. Fishburn, P. C. and Porter, R. B. (1976), "Optimal portfolios with one safe and one risky asset: effects of changes in rate of return and risk," *Management Science*, **22**, 1064–1073

18. Fomby, T. B. and Seo, T. K. (eds.) (1989), Studies in Economics of Uncertainty – In Honor of Josef Hadar. Springer Verlag, New York
19. Gupta, R. C. and Kirmani, S. N. U. A. (1998), "Residual life function in reliability studies," in Frontier in Reliability (Basu, A. P., Basu, S. K., and Mukhopadhyay, S. eds.). 175–190, World Scientific, Singapore
20. Hadar, J. and Russell, W. R. (1978), "Application in economic theory and analysis," in Stochastic Dominances: An Approach to Decision–Making under Risk (Whitmore, G. A. and Findlay, M. C. eds.). Lexington Books, Lexington
21. Hardy, G. H., Littlewood, J. E. and Pólya, G. (1952), Inequalities. 2nd Ed., Cambridge University Press. Cambridge
22. Hirshleifer, J. and Riley, J. G. (1992), The Analytics of Uncertainty and Information. Cambridge University Press, New York
23. Huang, C. F. and Litzenberger, R. H. (1988), Foundations for Financial Economics. Prentice Hall, New Jersey
24. Ingersoll, J. E. Jr. (1987), Theory of Financial Decision Making. Rowman and Littlefield, New York
25. Karlin, S. (1968), Total Positivity. Vol. I, Stanford University Press, Stanford
26. Karlin, S. and Novikov, A. (1963), "Generalized convex inequalities," *Pacific Journal of Mathematics*, **13**, 1251–1279
27. Keeny, R. L. and Raiffa, H. (1976), Decisions with Multiple Objectives: Preferences and Value Tradeoffs. John Wiley & Sons, New York
28. Keilson, J. (1979), Markov Chain Models – Rarity and Exponentiality. Springer–Verlag, New York
29. Keilson, J. and Sumita, U. (1982), "Uniform stochastic ordering and related properties," *Canadian Journal of Statistics*, **10**, 181–198
30. Kijima, M. (1989), "Uniform monotonicity of Markov processes and its related properties," *Journal of Operations Research Society of Japan*, **32**, 475–490
31. Kijima, M. (1997), Markov Chains for Stochastic Modeling. Chapman Hall, London
32. Kijima, M. and Ohnishi, M. (1992), "Addendum to the bivariate characterization of stochastic orders," Technical Report, **11**, Graduate School of Systems Management, University of Tsukuba, Tokyo
33. Kijima, M. and Ohnishi, M. (1993), "Mean–risk analysis for risk aversion and wealth effects on optimal portfolios with many investment opportunities," *Annals of Operations Research*, **45**, 147–163
34. Kijima, M. and Ohnishi, M. (1996), "Portfolio selection problems via the bivariate characterization of stochastic dominance relations," *Mathematical Finance*, **6**, 237–277
35. Kijima, M. and Ohnishi, M. (1996), "Further results on comparative statics for choice under risk," in Stochastic Models in Engineering, Technology and Management (Wilson, R. J., Murthy, D. N. P. and Osaki, S. eds.). *Proceedings of the Second Australia–Japan Workshop*, Gold Coast, Australia, 321–326
36. Kijima, M. and Ohnishi, M. (1999), "Stochastic orders and their applications in financial optimization," *Mathematical Methods of Operations Research*, **50**, 351–372
37. Kochar, S. C. (1998), "Stochastic comparisons and spacing and order statistics," in Frontier in Reliability (Basu, A. P., Basu, S. K. and Mukhopadhyay, S. eds.). 201–216, World Scientific, Singapore

38. Kroll, Y. and Levy, H. (1980), "Stochastic dominance criteria: a review and some new evidence," in Research in Finance (Levy, H. ed.). **2**, 163–227, JAI Press, Greenwich, Connecticut
39. Laffont, J.–J. (1985), Cours de Theorie Microeconomique. **II**, Economie de l'Incerton et de l'Infomation, Économica, Paris
40. Lehman, E. L. (1959), Testing Statistical Hypotheses, John Wiley & Sons, New York
41. Levy, H. (1992), "Stochastic dominance and expected utility: survey and analysis," *Management Science*, **38**, 555–593
42. Lynch, J., Mimmack, G., and Proschan, F. (1987), "Uniform stochastic orderings and total positivity," *The Canadian Journal of Statistics*, **13**, 63–69
43. Makowski, A. R. (1994), "On an elementary characterization of the increasing convex ordering, with an application," *Journal of Applied Probability*, **31**, 834–840
44. Mas–Colell, A., Whinston, M., and Green, J. R. (1995), Microeconomic Theory. Oxford University Press, New York
45. Marshall, A. W. and Olkin, I. (1979), Inequalities: The Theory of Majorization with Applications. Academic Press, New York
46. Mosler, K. and Scarsini, M. (eds.) (1991), Stochastic Orders and Decision under Risk. IMS Lecture Notes–Monograph Series, **19**, Institute of Mathematical Statistics, Hayward, California
47. Mosler, K. and Scarsini, M. (1993), Stochastic Orders and Applications – A Classified Bibliography. Lecture Notes in Economics and Mathematical Systems, **401**, Springer–Verlag, Berlin
48. Ohnishi, M., "Comparative statics for the equilibrium price system in a complete security market," (in preparation).
49. Righter, R. and Shanthikumar, J. G. (1992), "Extension of the bivariate characterization for stochastic orders," *Advances in Applied Probability*, **24**, 506–508
50. Rolski, T., Schmidli, H., Schmidt, V. and Teugels, J. (1999), Stochastic Processes for Insurance and Finance. John Wiley & Sons, Chichester
51. Ross, S. M. (1983), Introduction to Stochastic Dynamic Programming. Academic Press, New York
52. Ross, S. M. (1983), Stochastic Processes. John Wiley & Sons, New York
53. Sakagami, Y. (1997), "The comparative statistics of shifts in risk," *Journal of the Operations Research of Japan*, **40**, 522–535
54. Scarsini, M. (1994), "Comparing risk and risk aversion," in Stochastic Orders and Their Applications. 351–378, Academic Press, San Diego, California
55. Shaked, M. and Shanthikumar, J. G. (1990), "Reliability and maintainability," Handbooks in Operations Research and Management Science (Heyman, D. P. and Sobel, M. J. eds.). 653–713, North–Holland
56. Shaked, M. and Shanthikumar, J. G. (1994), Stochastic Orders and Their Applications. Academic Press, San Diego, California
57. Shaked, M. and Tong, Y. L. (eds.) (1992), Stochastic Inequalities. IMS Lecture Notes–Monograph Series **22**, Institute of Mathematical Statistics, Hayward, California
58. Shanthikumar, J. G., Yamazaki, G., and Sakasegawa, H. (1991), "Characterization of optimal order of servers in a tandem queue with blocking," *Operations Research Letters*, **10**, 17–22

59. Shanthikumar, J. G. and Yao, D. D. (1991), "Bivariate characterization of some stochastic order relations," *Advances in Applied Probability*, **23**, 642–659
60. Singh, H. (1998), "Role of stochastic orderings in spare allocation in systems," in Frontier in Reliability (Basu, A. P., Basu, S. K., and Mukhopadhyay, S. eds.). 353–359, World Scientific, Singapore
61. Stoyan, D. (1983), Comparison Methods for Queues and Other Stochastic Models (Revision by Daley, D. J. ed.). John Wiley & Sons, Chichester
62. Strassen, V. (1965), "The existence of probability measures with given marginals," *Annals of Mathematical Statistics*, **36**, 423–439
63. Tong, Y. L. (1980), Probability Inequalities in Multivariate Distributions. Academic Press, New York
64. Tong, Y. L. (ed.) (1984), Inequalities in Statistics and Probability. IMS Lecture Notes–Monograph Series **5**, Institute of Mathematical Statistics, Hayward, California
65. Tong, Y. L. (1990), The Multivariate Normal Distribution. Springer–Verlag, New York
66. Whitmore, G. A. (1970), "Third degree stochastic dominance," *American Economic Review*, **60**, 457–459
67. Whitmore, G. A. (1989), "Stochastic dominance for the class of completely monotonic utility functions," in Studies in the Economics of Uncertainty (Fomby, T. B. and Seo, T. K. eds.). 77–88, Springer–Verlag
68. Whitmore, G. A. and Findlay, M. C. (eds.) (1978), Stochastic Dominance: An Approach to Decision Making under Risk. Lexington Books, Toronto
69. Whitt, W. (1979), "A note on the influence of the sample on the posterior distribution," *Journal of the American Statistical Association*, **74**, 424–426
70. Whitt, W. (1980), "Uniform conditional stochastic order," *Journal of Applied Probability*, **17**, 112–123
71. Whitt, W. (1985), "Uniform conditional variability ordering on probability distribution," *Journal of Applied Probability*, **22**, 619–633
72. Ziemba, W. T. and Vickson, R. G. (1975), Stochastic Optimization in Finance. Academic Press, New York

2.A TP_2 Functions

In this appendix, we provide the information needed in this chapter about total positivity. The reader interested in more detailed discussions of the theory of total positivity should consult to Karlin (1968).

For a real-valued function $K(x,y)$ on $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, we denote

$$K\begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} := \det\begin{pmatrix} K(x_1,y_1) & K(x_1,y_2) \\ K(x_2,y_1) & K(x_2,y_2) \end{pmatrix}, \quad x_1 < x_2,\ y_1 < y_2. \tag{A.1}$$

Definition 2.A.1 A nonnegative function $K(x,y)$ is said to be *totally positive of order 2* or simply TP_2, denoted by $K \in TP_2$, if

$$K\begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} = K(x_1,y_1)K(x_2,y_2) - K(x_1,y_2)K(x_1,y_2) \geq 0,$$
$$x_1 < x_2,\ y_1 < y_2. \tag{A.2}$$

For nonnegative functions $K(x,z)$ and $L(z,y)$ defined on (a rectangle subset of) $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, let

$$M(x,y) := \int_{-\infty}^{\infty} K(x,z)L(z,y)dz, \qquad x,\ y \in \mathbb{R}. \tag{A.3}$$

The next result is a special case of the well known *composition formula* (see page 17 of Karlin (1968)).

Proposition 2.A.1 We have

$$M\begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} = \int\int_{z_1<z_2} K\begin{pmatrix} x_1 & x_2 \\ z_1 & z_2 \end{pmatrix} L\begin{pmatrix} z_1 & z_2 \\ y_1 & y_2 \end{pmatrix} dz_1 dz_2. \tag{A.4}$$

As a consequence, if $K \in TP_2$ and $L \in TP_2$ then $M \in TP_2$. □

Definition 2.A.2 A nonnegative function $f(x)$ on a subset of $\mathbb{R}$ is said to be *Pólya frequency of order 2* (PF_2) in x if $f(x-y)$ is TP_2 in x and y; *i.e.*,

$$f(x_1-y_1)f(x_2-y_2) - f(x_1-y_2)f(x_2-y_1) \geq 0, \quad x_1 < x_2,\ y_1 < y_2. \tag{A.5}$$

The class of PF_2 functions is important and has many applications in various fields. A key property that every PF_2 function possesses is the characterization that it has the form $f(x) = e^{-\phi(x)}$ where $\phi(x)$ is a convex function.

Proposition 2.A.2 Every PF_2 function on $\mathbb{R}$ is *log-concave.* □

The next result is found in page 128 of Karlin (1968).

Proposition 2.A.3 For CDFs F and G, let

$$H(x) := \int_{-\infty}^{\infty} F(x-y)dG(y) = \int_{-\infty}^{\infty} G(x-y)dF(y), \quad x \in \mathbb{R}. \tag{A.6}$$

If both F and G are PF_2 then so is H (H is a CDF too). □

3. Classical Maintenance Models

Naoto Kaio
Department of Economic Informatics,
Hiroshima Shudo University
Hiroshima 731–3195, Japan
Tadashi Dohi
Department of Industrial and Systems Engineering,
Hiroshima University
Higashi–Hiroshima 739–8527, Japan
and
Shunji Osaki
Department of Information & Telecommunication Engineering,
Nanzan University
Aichi 489–0863, Japan

Summary.

This chapter concerns the basic preventive maintenance policies arising in the context of the mathematical maintenance theory. Simple but practically important preventive maintenance optimization models, which involve age replacement and block replacement, are reviewed in the framework of the well–known renewal reward argument. Some variations to these basic models as well as the corresponding discrete time models are also introduced with the aim of their application of the theory to the practice.

Keywords: Preventive maintenance, age replacement, block replacement, order replacement, inspection strategies, renewal reward policies, continuous and discrete models, cost models, optimization

3.1 Introduction

The mathematical maintenance policies centered on preventive maintenance have been developed mainly in the research area of operations research/ management science, to generate the effective preventive maintenance schedule. The most important problem in mathematical maintenance policies is to design the maintenance plan by two maintenance options, preventive replacement and corrective replacement. In preventive replacement, the system or unit is replaced by a new one before it fails.

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

On the other hand, corrective replacement replaces each failed unit. A huge number of replacement methods have already been proposed in the literature. At the same time, some technical books on this problem have been published. For example, Arrow, Karlin and Scarf [1], Barlow and Proschan [2], [3], Jorgenson, McCall and Radner [4], Gnedenko, Belyayev and Solovyev [5], Gertsbakh [6], Ascher and Feingold [7] and Osaki [8], [9] are the classical but very important works in which to study the mathematical maintenance theory. Many authors have also published monographs on the specific topics. The reader can refer to Osaki and Hatoyama [10], Osaki and Cao [11], Ozekici [12] and Christer, Osaki and Thomas [13]. Recently, Barlow [14] and Aven and Jensen [15] presented excellent textbooks reviewing the mathematical maintenance theory. On the other hand, some survey papers will be useful to review the history of this research context, such as McCall [16], Osaki and Nakagawa [17], Pierskalla and Voelker [18], Sherif and Smith [19] and Valdez–Flores and Feldman [20].

Since the mechanism that will cause failure may be considered to be uncertain in most real complex systems , the mathematical technique used to deal with the maintenance problems should be based on convertional probability theory. If we are interested in the dynamic behavior of system failures depending on time, the problems are essentially reduced so that we study the stochastic processes presenting phenomena of both failure and replacement. In fact, since the theory of stochastic processes depends strongly on the mathematical maintenance theory, many textbooks on the stochastic processes have treated several maintenance problems. See, for example, Feller [21], Karlin and Taylor [22], [23], Taylor and Karlin [24] and Ross [25]. In other words, in order to design the maintenance schedule effectively, both the underlying stochastic process governing the failure mechanism and the role of maintenance options carried out on the process have to be analyzed carefully. In that sense, the mathematical maintenance theory is one of the most important parts in applied probability modelling.

In this tutorial article, we present the basic preventive maintenance policies arising in the context of the maintenance theory. Simple but practically important preventive maintenance optimization models, which involve age replacement and block replacement, are reviewed in the framework of the well–known renewal reward argument. Some variations on these basic models as well as the corresponding discrete time models are also introduced with the aim of applying of the theory to practice. In most textbooks and technical papers, the discrete–time preventive maintenance models have been paid little attention. The main reason is that the discrete–time models are ordinarily considered as trivial analogies of the continuous–time ones. However, we often face some maintenance problems modeled in discrete–time setting in practice. If one considers the situation where the number of take–offs from airports influences the damage to an airplane, the parts making up the airplane should be replaced after a prespecified number of take–offs rather than a specific lapse of time. Also, in the Japanese electric power company under our investigation, the failure time data of electric switching devices are recorded as group data (the number of failures per year) and it is not easy to carry out the preventive replacement schedule at the unit of week or month since the service team is engaged in other works, too. Our questionnaire suggests that it is helpful for practitioners to have the preventive replacement schedule determined roughly for the unit of 1year. This will motivate the discrete–time maintenance models.

3.2 Block Replacement

For block replacement models, the preventive replacement is executed periodically at a prespecified time kt_0 ($t_0 \geq 0$) or kN ($N = 0, 1, 2, \ldots$), ($k = 1, 2, 3, \ldots$). If the unit fails during the time interval $((k-1)t_0, kt_0]$ or $((k-1)N, kN]$, then the corrective maintenance is done at the time of failure. The main property for the block replacement is easier administration in general, since the preventive replacement time is scheduled in advance and one need not observe the age of a unit. In this section, we develop the following three variations of the block replacement model;

(i) A failed unit is replaced instantly at failure (Model I)
(ii) A failed unit remains inoperable until the next scheduled replacement comes (Model II)
(iii) A failed unit undergoes minimal repair (Model III)

The cost components used in this section are the following:

c_p (> 0): unit preventive replacement cost at time kt_0 or kN, ($k = 1, 2, 3, \ldots$).
c_c (> 0): unit corrective replacement cost at failure time.
c_m (> 0): unit minimal repair cost at failure time.

Model I

First, we consider the continuous–time block replacement model [2]. A failed unit is replaced by a new one during the replacement interval t_0, and the scheduled replacement for the non–failed unit is performed at kt_0 ($k = 1, 2, 3, \ldots$). Let $F(t)$ be the continuous lifetime distribution with finite mean $1/\lambda$ (> 0). From the well–known renewal reward theorem, the expected cost per unit of time in the steady–state for the block replacement model can be formulated directly as follows.

$$B_c(t_0) = \frac{c_c M(t_0) + c_p}{t_0}, \qquad t_0 \geq 0, \tag{3.1}$$

where the function $M(t) = \sum_{k=1}^{\infty} F^{(k)}(t)$ denotes the mean number of failures during the time period $(0, t]$ (the renewal function) and $F^{(k)}(t)$ the k–fold convolution of the lifetime distribution. The problem is, of course, to derive the optimal block replacement time t_0^* which minimizes $B_c(t_0)$.

Define the numerator of the derivative of $B_c(t_0)$ as

$$j_c(t_0) = c_c[t_0 m(t_0) - M(t_0)] - c_p, \tag{3.2}$$

where the function $m(t) = dM(t)/dt$ is the renewal density. Then, we have the optimal block replacement time t_0^* which minimizes the expected cost per unit of time in the steady–state $B_c(t_0)$.

Theorem 3.2.1

(1) Suppose that the function $m(t)$ is strictly increasing with respect to t (> 0).
  (i) If $j_c(\infty) > 0$, then there exists one finite optimal block replacement time t_0^* ($0 < t_0^* < \infty$) which satisfies $j_c(t_0^*) = 0$. Then the corresponding minimum expected cost is

$$B_c(t_0^*) = c_c m(t_0^*). \tag{3.3}$$

(ii) If $j_c(\infty) \leq 0$, then $t_0^* \to \infty$, that is, it is optimal to carry out only the corrective replacement, and the corresponding minimum expected cost is

$$B_c(\infty) = \lambda c_c. \tag{3.4}$$

(2) Suppose that the function $m(t)$ is decreasing with respect to t (> 0). Then, the optimal block replacement time is $t_0^* \to \infty$.

Next, we formulate the discrete–time block replacement model [29]. In the discrete–time setting, the expected cost per unit of time in the steady–state is

$$B_d(N) = \frac{c_c M(N) + c_p}{N}, \qquad N = 0, 1, 2, \ldots, \tag{3.5}$$

where the function $M(n) = \sum_{k=1}^{\infty} F^{(k)}(n)$ is the discrete renewal function for the discrete lifetime distribution $F(n)$ $(n = 0, 1, 2, \ldots)$, $F^{(k)}(n)$ the k–fold convolution; for more detail see Munter [26] and Kaio and Osaki [27]. Define the numerator of the difference of $B_d(N)$ as

$$j_d(N) = c_c[Nm(N+1) - M(N)] - c_p, \tag{3.6}$$

where the function $m(n) = M(n) - M(n-1)$ is the renewal probability mass function. Then, we have the optimal block replacement time N^* which minimizes the expected cost per unit of time in the steady–state $B_d(N)$.

Theorem 3.2.2

(1) Suppose that the $m(n)$ is strictly increasing with respect to n (> 0).
(i) If $j_d(\infty) > 0$, then there exists one finite optimal block replacement time N^* $(0 < N^* < \infty)$ which satisfies $j_d(N-1) < 0$ and $j_d(N) \geq 0$. Then the corresponding minimum expected cost satisfies the inequality

$$c_c m(N^*) < B_d(N^*) \leq c_c m(N^*+1). \tag{3.7}$$

(ii) If $j_d(\infty) \leq 0$, then $N^* \to \infty$, that is, it is optimal to carry out only the corrective replacement and the corresponding minimum expected cost is

$$B_d(\infty) = \lambda c_c. \tag{3.8}$$

(2) Suppose that the function $m(n)$ is decreasing with respect to n (> 0). Then, the optimal block replacement time is $N^* \to \infty$.

Remark 3.2.1 A large number of variations on the block replacement model have been studied in the literature. Though we assume in the model above that the cost component is constant, some modifications are possible. Tilquin and Cleroux [30] and Berg and Epstein [31] extended the original model in terms of cost structure.

Model II

For the first model, we have assumed that a failed unit is detected instantly just after failure. This implies that a sensing device monitors the operating unit. Since such a case is not always general, however, we assume

that the failure is detected only at kt_0 $(t_0 \geq 0)$ or kN $(N = 0, 1, 2, \ldots)$, $(k = 1, 2, 3, \ldots)$ [9]. Consequently, in Model II, a unit is always replaced at kt_0 or kN, but is not replaced at the time of failure, and the unit remains inoperable for the time duration from the occurrence of failure until its detection.

In the continuous–time model, since the expected duration from the occurrence of failure until its detection per cycle is given by $\int_0^{t_0}(t_0 - t)dF(t) = \int_0^{t_0} F(t)dt$, we have the expected cost per unit time in the steady–state;

$$C_c(t_0) = \frac{c_c \int_0^{t_0} F(t)dt + c_p}{t_0}, \tag{3.9}$$

where c_c is changed to the cost of failure per unit time, that is, the cost occurs per unit time for system down. Define the numerator of the derivative of $C_c(t_0)$ with respect to t_0 as $k_c(t_0)$, *i.e.*

$$k_c(t_0) = c_c\Big\{F(t_0)t_0 - \int_0^{t_0} F(t)dt\Big\} - c_p. \tag{3.10}$$

Theorem 3.2.3

(i) If $k_c(\infty) > 0$, then there exists one unique optimal block replacement time t_0^* $(0 < t_0^* < \infty)$ which satisfies $k_c(t_0^*) = 0$, and the corresponding minimum expected cost is

$$C_c(t_0^*) = c_c F(t_0^*). \tag{3.11}$$

(ii) If $k_c(\infty) \leq 0$, then $t_0^* \to \infty$ and $C_c(\infty) = c_c$.

On the other hand, in the discrete–time setting, the expected cost per unit of time in the steady–state is

$$C_d(N) = \frac{c_c \sum_{k=1}^{N-1} F(k) + c_p}{N}, \qquad N = 0, 1, 2, \ldots, \tag{3.12}$$

where the function $F(n)$ is the lifetime distribution $(n = 0, 1, 2, \ldots)$. Define the numerator of the difference of $C_d(N)$ as

$$i_d(N) = c_c[NF(N) - \sum_{k=1}^{N-1} F(k)] - c_p. \tag{3.13}$$

Then, we have the optimal block replacement time N^* which minimizes the expected cost per unit time in the steady–state $C_d(N)$.

Theorem 3.2.4

(i) If $j_d(\infty) > 0$, then there exists one finite optimal block replacement time N^* $(0 < N^* < \infty)$ which satisfies $i_d(N-1) < 0$ and $i_d(N) \geq 0$. Then the corresponding minimum expected cost satisfies the inequality

$$c_c F(N^* - 1) < C_d(N^*) \leq c_c F(N^*). \tag{3.14}$$

(ii) If $i_d(\infty) \leq 0$, then $N^* \to \infty$.

Remark 3.2.2 It is noticed that Model II has not been studied in many works in the literature, since this cannot detect the failure instantly and is not invariably superior to Model I in terms of cost minimization. However, as described previously, one can see that continuous monitoring of the operating unit is not always possible for all practical applications.

Model III

In the final model, we assume that minimal repair is performed when a unit fails and the failure rate is not disturbed by each repair. If we consider a stochastic process $\{N(t), t \geq 0\}$ in which $N(t)$ represents the number of minimal repairs up to time t, the process $\{N(t), t \geq 0\}$ is governed by a non–homogeneous Poisson process with mean value function

$$\Lambda(t) = \int_0^t r(x)dx, \tag{3.15}$$

which is also called the hazard function, where the function $r(t) = f(t)/\bar{F}(t)$ is called the failure rate or the hazard rate, in general $\bar{\psi}(\cdot) = 1 - \psi(\cdot)$. Noting this fact, Barlow and Hunter [28] gave the expected cost per unit time in the steady–state for the continuous–time model;

$$V_c(t_0) = \frac{c_m \Lambda(t_0) + c_p}{t_0}. \tag{3.16}$$

Define the numerator of the derivative of $V_c(t_0)$ as

$$l_c(t_0) = c_m[t_0 r(t_0) - \Lambda(t_0)] - c_p. \tag{3.17}$$

Then, we have the optimal block replacement time (with minimal repair) t_0^* which minimizes the expected cost per unit time in the steady–state $V_c(t_0)$.

Theorem 3.2.5

(1) Suppose that $F(t)$ is strictly IFR (Increasing Failure Rate), *i.e.* the failure rate is strictly increasing.

(i) If $l_c(\infty) > 0$, then there exists one finite optimal block replacement time with minimal repair t_0^* $(0 < t_0^* < \infty)$ which satisfies $l_c(t_0^*) = 0$. Then the corresponding minimum expected cost is

$$V_c(t_0^*) = c_m r(t_0^*). \tag{3.18}$$

(ii) If $l_c(\infty) \leq 0$, then $t_0^* \to \infty$ and the corresponding minimum expected cost is

$$V_c(\infty) = c_m r(\infty). \tag{3.19}$$

(2) Suppose that $F(t)$ is DFR (Decreasing Failure Rate), *i.e.* the failure rate is decreasing. Then, the optimal block replacement time with minimal repair is $t_0^* \to \infty$.

Next, we formulate the discrete–time block replacement model with minimal repair [29]. In the discrete–time setting, the expected cost per unit time in the steady–state is

$$V_d(N) = \frac{c_m \Lambda(N) + c_p}{N}, \qquad N = 0, 1, 2, \ldots, \tag{3.20}$$

where the function $\Lambda(n)$ is the mean value function of the discrete non–homogeneous Poisson process. Define the numerator of the difference of $V_d(N)$ as

$$l_d(N) = c_m[Nr(N+1) - \Lambda(N)] - c_p, \tag{3.21}$$

where the function $r(n) = \Lambda(n) - \Lambda(n-1) = f(n)/\bar{F}(n-1)$ is the failure rate function. Then, we have the optimal block replacement time with minimal repair N^* which minimizes the expected cost per unit time in the steady–state $V_d(N)$.

Theorem 3.2.6

(1) Suppose that $F(n)$ is strictly IFR.
   (i) If $l_d(\infty) > 0$, then there exists one finite optimal block replacement time with minimal repair N^* $(0 < N^* < \infty)$ which satisfies $l_d(N-1) < 0$ and $l_d(N) \geq 0$. Then the corresponding minimum expected cost satisfies the inequality

$$c_m r(N^*) < V_d(N^*) \leq c_m r(N^*+1). \tag{3.22}$$

   (ii) If $l_d(\infty) \leq 0$, then $N^* \to \infty$.

(2) Suppose that $F(n)$ is DFR. Then, the optimal block replacement time with minimal repair is $N^* \to \infty$.

Remark 3.2.3 So far as we know, a large number of papers on minimal repair models have been published. Morimura [32], Tilquin and Cleroux [33] and Cleroux, Dubuc and Tilquin [34] deal with several interesting modifications. Latter, Park [35], Nakagawa [36]–[40], Nakagawa and Kowada [41], Phelps [42], Berg and Cleroux [43], Boland [44], Boland and Proschan [45], Block, Borges and Savits [46] and Beichelt [47] propose extended minimal repair models from the standpoint of generalization. Among the most interesting models with minimal repair is the (t,T)–policy. The (t,T)–policy is a combined policy with three kinds of maintenance options; minimal repair, failure replacement and preventive replacement. That is, minimal repair is executed for failures during the first period $[0,t)$, but failure replacement is done for $[t,T]$, where T is the preventive replacement time. Tahara and Nishida [48] were the first to formulate this model. After investigations by Phelps [49] and Segawa, Ohnishi and Ibaraki [50], Ohnishi [51] recently proved the optimality of the (t,T)–policy under average cost criterion, via the dynamic programming approach. This tells us that the (t,T)–policy is optimal if we have only three kinds of maintenance options.

3.3 Age Replacement

As well known, in the age replacement model, if the unit does not fail before a prespecified time t_0 $(t_0 \geq 0)$ or N $(N = 0,1,2,\ldots)$, then it is replaced by a new one preventively; otherwise, it is replaced at the failure time. Denote the corrective and the preventive replacement costs by c_c and c_p, respectively, where, without loss of generality, $c_c > c_p$. This model plays a central role in all replacement models, since the optimality of the age

replacement model has been proved by Bergman [52] if the replacement by a new unit is the only maintenance option (*i.e.*, if no repair is considered as an alternative option). In the rest of this section, we introduce three kinds of age replacement models.

Basic Age Replacement Model (Barlow and Proschan [2], Barlow and Hunter [53] and Osaki and Nakagawa [54])

From the renewal reward theorem, it can be seen that the expected cost per unit time in the steady-state for the age replacement model is

$$A_c(t_0) = \frac{c_c F(t_0) + c_p \bar{F}(t_0)}{\int_0^{t_0} \bar{F}(t)dt}, \qquad t_0 \geq 0. \tag{3.23}$$

If one can assume that the density is $f(t)$ for the lifetime distribution $F(t)$ $(t \geq 0)$, the failure rate $r(t) = f(t)/\bar{F}(t)$ necessarily exists. Define the numerator of the derivative of $A_c(t_0)$ with respect to t_0, divided by $\bar{F}(t_0)$ as $h_c(t_0)$, *i.e.*

$$h_c(t_0) = r(t_0) \int_0^{t_0} \bar{F}(t)dt - F(t_0) - \frac{c_p}{c_c - c_p}. \tag{3.24}$$

Then, we have the optimal age replacement time t_0^* which minimizes the expected cost per unit time in the steady-state $A_c(t_0)$.

Theorem 3.3.1

(1) Suppose that the lifetime distribution $F(t)$ is strictly IFR.

(i) If $r(\infty) > K = \lambda c_c/(c_c - c_p)$, then there exists one finite optimal age replacement time t_0^* $(0 < t_0^* < \infty)$ which satisfies $h_c(t_0^*) = 0$. Then the corresponding minimum expected cost is

$$A_c(t_0^*) = (c_c - c_p)r(t_0^*). \tag{3.25}$$

(ii) If $r(\infty) \leq K$, then $t_0^* \to \infty$ and $A_c(\infty) = B_c(\infty) = \lambda c_c$.

(2) If $F(t)$ is DFR, then $t_0^* \to \infty$.

Next, let us consider the case where the cost is discounted by the discount factor α $(\alpha > 0)$ [55]. The present value of a unit cost after t $(t \geq 0)$ period is $\exp(-\alpha t)$. In the continuous-time age replacement problem, the expected total discounted cost over an infinite time horizon is

$$A_c(t_0;\alpha) = \frac{c_c \int_0^{t_0} e^{-\alpha t} f(t)dt + c_p e^{-\alpha t_0} \bar{F}(t_0)}{\alpha \int_0^{t_0} e^{-\alpha t} \bar{F}(t)dt}, \qquad t_0 \geq 0. \tag{3.26}$$

Define the numerator of the derivative of $A_c(t_0;\alpha)$ with respect to t_0, divided by $\bar{F}(t_0)\exp(-\alpha t_0)$ as $h_c(t_0;\alpha)$,

$$h_c(t_0;\alpha) = r(t_0) \int_0^{t_0} e^{-\alpha t} \bar{F}(t)dt - \int_0^{t_0} e^{-\alpha t} f(t)dt - \frac{c_p}{c_c - c_p}. \tag{3.27}$$

Then, we have the optimal age replacement time t_0^* which minimizes the expected total discounted cost over an infinite time horizon $A_c(t_0;\alpha)$.

Theorem 3.3.2

(1) Suppose that the lifetime distribution $F(t)$ is strictly IFR.

(i) If $r(\infty) > K(\alpha)$, then there exists one finite optimal age replacement time t_0^* $(0 < t_0^* < \infty)$ which satisfies $h_c(t_0^*; \alpha) = 0$, where

$$K(\alpha) = \frac{c_c F^*(\alpha) + c_p \bar{F}^*(\alpha)}{(c_c - c_p)\bar{F}^*(\alpha)/\alpha}, \tag{3.28}$$

$$F^*(\alpha) = \int_0^\infty e^{-\alpha t} f(t) dt. \tag{3.29}$$

Then the corresponding minimum expected cost is

$$A_c(t_0^*; \alpha) = \frac{(c_c - c_p) r(t_0^*)}{\alpha} - c_p. \tag{3.30}$$

(ii) If $r(\infty) \leq K(\alpha)$, then $t_0^* \to \infty$ and

$$A_c(\infty; \alpha) = c_c F^*(\alpha)/\bar{F}^*(\alpha). \tag{3.31}$$

(2) If $F(t)$ is DFR, then $t_0^* \to \infty$.

Following Nakagawa and Osaki [56], let us consider the discrete age replacement model. Define the discrete lifetime distribution $F(n)$ $(n = 0, 1, 2, \dots)$, the probability mass function $f(n)$ and the failure rate $r(n) = f(n)/\bar{F}(n-1)$. From the renewal reward theorem, it can be seen that the expected cost per unit time in the steady–state for the age replacement model is

$$A_d(N) = \frac{c_c F(N) + c_p \bar{F}(N)}{\sum_{i=1}^{N} \bar{F}(i-1)}, \qquad N = 0, 1, 2, \dots. \tag{3.32}$$

Define the numerator of the difference of $A_d(N)$ as

$$h_d(N) = r(N+1) \sum_{i=1}^{N} \bar{F}(i-1) - F(N) - \frac{c_p}{c_c - c_p}. \tag{3.33}$$

Then, we have the optimal age replacement time N^* which minimizes the expected cost per unit time in the steady–state $A_d(N)$.

Theorem 3.3.3

(1) Suppose that the lifetime distribution $F(N)$ is strictly IFR.

(i) If $r(\infty) > K$, then there exists one finite optimal age replacement time N^* $(0 < N^* < \infty)$ which satisfies $h_d(N^* - 1) < 0$ and $h_d(N^*) \geq 0$. Then the corresponding minimum expected cost satisfies the following inequality

$$(c_c - c_p) r(N^*) < A_d(N^*) \leq (c_c - c_p) r(N^* + 1). \tag{3.34}$$

(ii) If $r(\infty) \leq K$, then $N^* \to \infty$ and $A_d(\infty) = B_d(\infty) = \lambda c_c$.

(2) If $F(N)$ is DFR, then $N^* \to \infty$.

We introduce the discount factor β $(0 < \beta < 1)$ in the discrete-time age replacement problem. The present value of a unit cost after n $(n = 0, 1, 2, \ldots)$ periods is β^n. In the discrete-time age replacement problem, the expected total discounted cost over an infinite time horizon is

$$A_d(N;\beta) = \frac{c_c \sum_{j=0}^{N} \beta^j f(j) + c_p \beta^N \bar{F}(N)}{\frac{1-\beta}{\beta} \sum_{i=1}^{N} \beta^i \bar{F}(i-1)}, \qquad N = 0, 1, 2, \ldots. \tag{3.35}$$

Define the numerator of the difference of $A_d(N;\beta)$ as

$$h_d(N;\beta) = r(N+1) \sum_{i=1}^{N} \beta^i \bar{F}(i-1) - \sum_{j=0}^{N} \beta^j f(j) - \frac{c_p}{c_c - c_p}. \tag{3.36}$$

Then, we have the optimal age replacement time N^* which minimizes the expected total discounted cost over an infinite time horizon $A_d(N;\beta)$.

Theorem 3.3.4

(1) Suppose that the lifetime distribution $F(N)$ is strictly IFR.

(i) If $r(\infty) > K(\beta)$, then there exists one finite optimal age replacement time N^* $(0 < N^* < \infty)$ which satisfies $h_d(N^* - 1;\beta) < 0$ and $h_d(N^*;\beta) \geq 0$. Then the corresponding minimum expected cost satisfies the following inequalities.

$$\frac{(c_c - c_p) r(N^*)}{(1-\beta)/\beta} - c_p < A_d(N^*;\beta) \tag{3.37}$$

and

$$A_d(N^*;\beta) \leq \frac{(c_c - c_p) r(N^* + 1)}{(1-\beta)/\beta} - c_p, \tag{3.38}$$

where

$$K(\beta) = \frac{c_c \sum_{j=0}^{\infty} \beta^j f(j) + c_p \sum_{j=0}^{\infty} (1 - \beta^j) f(j)}{(c_c - c_p) \sum_{i=1}^{\infty} \beta^i \bar{F}(i-1)}. \tag{3.39}$$

(ii) If $r(\infty) \leq K(\beta)$, then $N^* \to \infty$ and

$$A_d(\infty;\beta) = \frac{c_c \sum_{j=0}^{\infty} \beta^j f(j)}{\sum_{j=0}^{\infty} (1 - \beta^j) f(j)}. \tag{3.40}$$

(2) If $F(N)$ is DFR, then $N^* \to \infty$.

Theorem 3.3.5

(1) For the continuous–time age replacement problems, the following relationships hold.

$$A_c(t_0) = \lim_{\alpha \to 0} \alpha A_c(t_0; \alpha), \tag{3.41}$$

$$h_c(t_0) = \lim_{\alpha \to 0} h_c(t_0; \alpha), \tag{3.42}$$

$$K = \lim_{\alpha \to 0} K(\alpha). \tag{3.43}$$

(2) For the discrete–time age replacement problems, the following relationships hold.

$$A_d(N) = \lim_{\beta \to 1} (1 - \beta) A_d(N; \beta), \tag{3.44}$$

$$h_d(N) = \lim_{\beta \to 1} h_d(N; \beta), \tag{3.45}$$

$$K = \lim_{\beta \to 1} K(\beta). \tag{3.46}$$

Remark 3.3.1 Glasser [57], Scheaffer [58], Cleroux and Hanscom [59], Osaki and Yamada [60] and Nakagawa and Osaki [61] extended the basic age replacement model mentioned above. Here, we introduce an interesting topic on the age replacement policy under a different cost criterion from the expected cost per unit of time in the steady–state. Based on the seminal idea by Derman and Sacks [62], Ansell, Bendell and Humble [63] analyzed the age replacement model under an alternative cost criterion. In the continuous–time model with no discount, let Y_i and S_i denote the total cost and the time length for i–th cycle ($i = 1, 2, \ldots$), respectively, where $Y_i = c_c I_{\{X_i < t_0\}} + c_p I_{\{X_i \geq t_0\}}$, $S_i = \min\{X_i, t_0\}$, X_i is the lifetime for the i–th cycle and $I_{\{\cdot\}}$ is the indicator function.

In (3.23), we find that

$$\lim_{t \to \infty} \frac{\mathrm{E}[\text{total cost on } (0, t]]}{t} = \mathrm{E}[Y_i]/\mathrm{E}[S_i] = A_c(t_0). \tag{3.47}$$

On the other hand, let $NU(t)$ denote the number of cycles up to time t. Then, we define

$$\eta(t) \equiv \frac{1}{NU(t)} \sum_{i=1}^{NU(t)} \mathrm{E}[Y_i/S_i], \tag{3.48}$$

where $\eta(t)$ is the mean of the ratio $\mathrm{E}[Y_i/S_i]$ during $NU(t)$ cycles. From the independence of each cycle, we have

$$\begin{aligned} A_c^*(t_0) &= \lim_{t \to \infty} \eta(t) = \mathrm{E}[Y_i/S_i] \\ &= \int_0^{t_0} (c_c/t) dF(t) + \int_{t_0}^{\infty} (c_p/t_0) dF(t). \end{aligned} \tag{3.49}$$

This interesting cost criterion is called the expected cost ratio and is of course different from $\mathrm{E}[Y_i]/\mathrm{E}[S_i]$. Ansell, Bendell and Humble [63] compared this model with an approximated age replacement policy with finite time horizon by Christer [64], [65] and Christer and Jack [66].

3.4 Order Replacement

In both block and age replacement problems, a spare unit is available whenever the original unit fails. However, it should be noted that this assumption is questionable in most practical cases. In fact, if a sufficiently large number of spare units is always kept on hand, a large inventory holding cost will be needed. Hence, if system failure can be considered as a rare event for the operating system, the spare unit will be ordered when it is required. There were seminal contributions by Wiggins [67], Allen and D'Esopo [68], [69], Simon and D'Esopo [70], Nakagawa and Osaki [71] and Osaki [72]. A large number of order replacement models have been analyzed by many authors. For instance, the reader should refer to Thomas and Osaki [73], [74], Kaio and Osaki [75]–[78] and Osaki, Kaio and Yamada [79]. A comprehensive bibliography in this research area is listed in Dohi, Kaio and Osaki [80].

Continuous-Time Model

Let us consider a replacement problem for one–unit system where each failed unit is scrapped and each spare is provided, after a lead time, in response to an order. The original unit begins operating at time $t = 0$, and the planning horizon is infinite. If the original unit does not fail up to a prespecified time $t_0 \in [0, \infty)$, the regular order for a spare is made at the time t_0 and after a lead time L_2 (> 0) the spare is delivered. Then if the original unit has already failed by $t = t_0 + L_2$, the deliver spare is put into operation immediately. But even if the original unit is still operating, the unit is replaced by the spare preventively. On the other hand, if the original unit fails before the time t_0, an expedited order is made immediately at the failure time and the spare is put into operation just after it is delivered after a lead time L_1 (> 0). In this situation, it should be noted that the regular order is not made. The same cycle repeats itself continually.

Under this model, define the interval from one replacement to the following replacement as one cycle. Let c_1 (> 0), c_2 (> 0), k_1 (> 0), w (> 0) and s (< 0) be the expedited ordering cost, the regular ordering cost, the system down (shortage) cost per unit of time, the operation cost per unit of time and the salvage cost per unit of time, respectively. Then, the expected cost per unit time in the steady–state is

$$O_c(t_0) = V_c(t_0)/T_c(t_0), \tag{3.50}$$

where

$$\begin{aligned} V_c(t_0) &= c_1 \int_0^{t_0} dF(t) + c_2 \int_{t_0}^{\infty} dF(t) + k_1 \Big\{ \int_0^{t_0} L_1 dF(t) \\ &\quad + \int_{t_0}^{t_0+L_2} (t_0 + L_2 - t) dF(t) \Big\} + w \Big\{ \int_0^{t_0+L_2} t dF(t) \\ &\quad + \int_{t_0+L_2}^{\infty} (t_0 + L_2) dF(t) \Big\} + s \int_{t_0+L_2}^{\infty} (t - t_0 - L_2) dF(t) \\ &= c_1 F(t_0) + c_2 \bar{F}(t_0) + k_1 \Big\{ (L_1 - L_2) F(t_0) + \int_{t_0}^{t_0+L_2} F(t) dt \Big\} \\ &\quad + w \int_0^{t_0+L_2} \bar{F}(t) dt + s \int_{t_0+L_2}^{\infty} \bar{F}(t) dt, \qquad t_0 \geq 0 \end{aligned} \tag{3.51}$$

and

$$T_c(t_0) = \int_0^{t_0} (t + L_1)dF(t) + \int_{t_0}^{\infty} (t_0 + L_2)dF(t)$$
$$= (L_1 - L_2)F(t_0) + L_2 + \int_0^{t_0} \bar{F}(t)dt. \tag{3.52}$$

Define the numerator of the derivative of $O_c(t_0)$ with respect to t_0, divided by $\bar{F}(t_0)$ as $q_c(t_0)$, *i.e.*

$$q_c(t_0) = \Big\{(k_1 - w + s)R(t_0) + (w - s) + \Big[k_1(L_1 - L_2) + (c_1 - c_2)\Big]r(t_0)\Big\}\Big\{(L_1 - L_2)F(t_0) + L_2 + \int_0^{t_0} \bar{F}(t)dt\Big\}$$
$$-\Big\{w\int_0^{t_0+L_2} \bar{F}(t)dt + c_1F(t_0) + c_2\bar{F}(t_0)$$
$$+k_1\Big[(L_1 - L_2)F(t_0) + \int_{t_0}^{t_0+L_2} F(t)dt\Big]$$
$$+s\int_{t_0+L_2}^{\infty} \bar{F}(t)dt\Big\}\Big\{(L_1 - L_2)r(t_0) + 1\Big\}, \tag{3.53}$$

where the function

$$R(t_0) = \frac{F(t_0 + L_2) - F(t_0)}{\bar{F}(t_0)} \tag{3.54}$$

has the same monotone properties as the failure rate $r(t_0)$, that is, $R(t)$ is increasing (decreasing) if and only if $r(t)$ is increasing (decreasing). Then, we have the optimal order replacement time t_0^* which minimizes the expected cost per unit time in the steady-state $O_c(t_0)$.

Theorem 3.4.1

(1) Suppose that the lifetime distribution $F(t)$ is strictly IFR.

(i) If $q_c(0) < 0$ and $q_c(\infty) > 0$, then there exists one finite optimal order replacement time t_0^* $(0 < t_0^* < \infty)$ which satisfies $q_c(t_0^*) = 0$. Then the corresponding minimum expected cost is

$$O_c(t_0^*) = \frac{(k_1 - w + s)R(t_0^*) + (w - s) + \mu(t_0^*)}{(L_1 - L_2)r(t_0^*) + 1}, \tag{3.55}$$

where

$$\mu(t_0) = \{k_1(L_1 - L_2) + (c_1 - c_2)\}r(t_0). \tag{3.56}$$

(ii) If $q_c(\infty) \leq 0$, then $t_0^* \to \infty$ and

$$O_c(\infty) = \frac{w/\lambda + c_1 + k_1L_1}{L_1 + 1/\lambda}. \tag{3.57}$$

(iii) If $q_c(0) \geq 0$, then $t_0^* = 0$ and

$$O_c(0) = \frac{1}{L_2}\Big\{w\int_0^{L_2} \bar{F}(t)dt + c_2 + k_1\int_0^{L_2} F(t)dt + s\int_{L_2}^{\infty} \bar{F}(t)dt\Big\}. \tag{3.58}$$

(2) Suppose that $F(t)$ is DFR. If the inequality

$$\Big\{w\int_0^{L_2} \bar{F}(t)dt + c_2 + k_1\int_0^{L_2} F(t)dt$$

$$+s\int_{L_2}^{\infty}\bar{F}(t)dt\Big\}(L_1+1/\lambda)$$
$$<(w/\lambda+c_1+k_1L_1)L_2 \tag{3.59}$$

holds, then $t_0^*=0$, otherwise, $t_0^*\to\infty$.

Discrete-Time Model

In the discrete order–replacement model, the function in (3.54) is given by

$$R(N)=\frac{\sum_{n=1}^{N+L_2}f(n)-\sum_{n=1}^{N}f(n)}{\sum_{n=N+1}^{\infty}f(n)}. \tag{3.60}$$

Then, the expected cost per unit time in the steady–state is

$$O_d(N)=V_d(N)/T_d(N), \tag{3.61}$$

where

$$\begin{aligned}V_d(N)=w&\sum_{i=1}^{N+L_2}\sum_{n=i}^{\infty}f(n)+c_1\sum_{n=1}^{N-1}f(n)+c_2\sum_{n=N}^{\infty}f(n)\\&+k_1\Big\{(L_1-L_2)\sum_{n=1}^{N-1}f(n)+\sum_{i=N+1}^{N+L_2}\sum_{n=1}^{i-1}f(n)\\&+s\sum_{i=N+L_2+1}^{\infty}\sum_{n=i}^{\infty}f(n)\Big\}\end{aligned} \tag{3.62}$$

and

$$T_d(N)=(L_1-L_2)\sum_{n=1}^{N-1}f(n)+L_2+\sum_{i=1}^{N}\sum_{n=i}^{\infty}f(n). \tag{3.63}$$

Note in the equations above that L_1 and L_2 are positive integers.

Similar to (3.53), define the numerator of the difference of $O_d(N)$ with respect to N, divided by $\bar{F}(N)$ as $q_d(N)$, *i.e.*

$$\begin{aligned}q_d(N)=&\Big\{(k_1-w+s)R(N)+(w-s)+[k_1(L_1-L_2)\\&+(c_1-c_2)]r(N)\Big\}\Big\{(L_1-L_2)\sum_{n=1}^{N-1}f(n)+L_2\\&+\sum_{i=1}^{N}\sum_{n=i}^{\infty}f(n)\Big\}-\Big\{w\sum_{i=1}^{N+L_2}\sum_{n=i}^{\infty}f(n)+c_1\sum_{n=1}^{N-1}f(n)\\&+c_2\sum_{n=N}^{\infty}f(n)+k_1\Big[(L_1-L_2)\sum_{n=1}^{N-1}f(n)\\&+\sum_{i=N+1}^{N+L_2}\sum_{n=1}^{i-1}f(n)\Big]+s\sum_{i=N+L_2+1}^{\infty}\sum_{n=i}^{\infty}f(n)\Big\}\\&\times[(L_1-L_2)r(N)+1].\end{aligned} \tag{3.64}$$

Then, we have the optimal order replacement time N^* which minimizes the expected cost per unit time in the steady–state $O_d(N)$.

Theorem 3.4.2
(1) Suppose that the lifetime distribution $F(N)$ is strictly IFR.
  (i) If $q_d(0) < 0$ and $q_d(\infty) > 0$, then there exists one finite optimal order replacement time N^* $(0 < N^* < \infty)$ which satisfies $q_d(N^* - 1) < 0$ and $q_d(N^*) \geq 0$.
  (ii) If $q_d(\infty) \leq 0$, then $N^* \to \infty$.
  (iii) If $q_d(0) \geq 0$, then $N^* = 0$.
(2) Suppose that $F(N)$ is DFR. Then $N^* = 0$, otherwise, $N^* \to \infty$.

Remark 3.4.1 This section has dealt with typical order–replacement models in both continuous and discrete–time settings. These models can be extended from various viewpoints. Thomas and Osaki [74] and Dohi, Kaio and Osaki [80] presented continuous models with stochastic lead times. Osaki, Kaio and Yamada [79] proposed a combined model with minimal repair and introduced the alternative concept of negative ordering time. Recently, Dohi, Kaio and Osaki [81], Dohi, Shibuya and Osaki [82] and Shibuya, Dohi and Osaki [83] applied the order–replacement model to analysis of the special problems of continuous review cyclic inventory control.

3.5 Inspection Strategies

There are several systems where failures are not detected immediately they occur, usually those in which failure is not catastrophic and an inspection is needed to reveal the fault. If we execute too many inspections, then system failure is detected more rapidly, but we incur a high inspection cost. Conversely, if we execute few inspections, the interval between the failure and its detection increases and we incur a high cost of failure. The optimal inspection policy minimizes the total expected cost composed of costs for inspection and system failure. From this viewpoint, many authors have discussed optimal and/or near–optimal inspection policies [86], [2], [87]–[95].

Among those policies, the inspection policy discussed by Barlow *et al.* [86], [2] is the best known. They have discussed the optimal inspection policy in the following model. A one–unit system is considered, which obeys an arbitrary lifetime distribution $F(t)$ with a p.d.f. (probability density function) $f(t)$. The system is inspected at prespecified times t_k $(k = 1, 2, 3, \ldots)$, where each inspection is executed perfectly and instantaneously. The policy terminates with an inspection which can detect the system failure. The costs considered are $c_i(> 0)$, the cost of an inspection, and $k_f(> 0)$, the cost of failure per unit of time. Then the total expected cost is

$$C_B = \sum_{k=0}^{\infty} \int_{t_k}^{t_{k+1}} [c_i(k+1) + k_f(t_{k+1} - t)] dF(t) . \tag{3.65}$$

They have obtained an algorithm to seek the optimal inspection–time sequence which minimizes the total expected cost in (3.65) by using the recurrence formula

$$t_{k+1} - t_k = \frac{F(t_k) - F(t_{k-1})}{f(t_k)} - \frac{c_i}{k_f} , \qquad k = 1, 2, 3, \ldots , \tag{3.66}$$

where $f(t)$ is a PF_2 (Pólya frequency function of order 2) with $f(t+\triangle)/f(t)$ strictly decreasing for $t \geq 0$, $\triangle > 0$, and with $f(t) > 0$ for $t > 0$, and $t_0 = 0$. The algorithm is as follows:

begin:
choose t_1 to satisfy $c_i = k_f \int_0^{t_1} F(t)dt$;
repeat
  compute $t_2, t_3, \ldots$, recursively using (3.66);
  if any $t_{k+1} - t_k > t_k - t_{k-1}$,
  then reduce t_1 ;
  if any $t_{k+1} - t_k < 0$,
  then increase t_1 ;
until $t_1 < t_2 < \cdots$ are determined to the degree of accuracy required;
end.

However, this algorithm by Barlow *et al.* is complicated to execute, because one must apply trial and error to decide the first inspection time t_1, and the assumption on $f(t)$ is restrictive. To overcome these difficulties, some improved procedures for obtaining the near–optimal inspection policy have been proposed [87], [89]–[95].

We review the near–optimal inspection policies proposed by Kaio and Osaki [87], [94], [95], Munford and Shahani [89], and Nakagawa and Yasui [92]. We follow the inspection model and notation introduced by Barlow *et al.* For more detail, see each of the cited papers.

Near-optimal inspection policy of Kaio and Osaki (K&O policy)

Introduce the inspection density at time t, $n(t)$, which is a smooth function and denotes the approximate number of inspections per unit of time at time t. Then the total expected cost up to the detection of the failure is approximately

$$C_n(n(t)) = c_i \int_0^\infty n(t)\bar{F}(t)dt + k_f \int_0^\infty \frac{1}{2n(t)} dF(t), \tag{3.67}$$

where $\bar{\psi} = 1 - \psi$, in general. The density $n(t)$ which minimizes the functional $C_n(n(t))$ in (3.67) is

$$n(t) = [k_c r(t)]^{1/2}, \tag{3.68}$$

where $k_c = k_f/(2c_i)$, and $r(t) = f(t)/\bar{F}(t)$, a failure rate. The inspection times t_k $(k = 1, 2, 3, \ldots)$ satisfy

$$k = \int_0^{t_k} n(t)dt, \qquad k = 1, 2, 3, \ldots . \tag{3.69}$$

Substituting $n(t)$ in (3.68) into (3.69) yields the near–optimal inspection–time sequence.

Kaio and Osaki obtained this procedure by developing that of Keller [93]. For details, see Kaio and Osaki [87]. Note that the procedure does not depend on assumptions about the p.d.f. $f(t)$.

Near-optimal inspection policy of Munford and Shahani (M&S policy)
Let

$$\frac{F(t_k) - F(t_{k-1})}{\bar{F}(t_{k-1})} = p, \qquad k = 1, 2, 3, \ldots, \qquad 0 < p < 1. \tag{3.70}$$

Then the inspection times t_k $(k = 1, 2, 3, \ldots)$ are

$$t_k = F^{-1}(1 - \bar{p}^k), \qquad k = 1, 2, 3, \ldots, \tag{3.71}$$

where the probability p is chosen such that the near–total expected cost up to the detection of the failure, $C_p(p)$, is minimized:

$$C_p(p) = \frac{c_i}{p} + k_f \left(\sum_{k=1}^{\infty} t_k \bar{p}^{k-1} p - \int_0^{\infty} t f(t) dt \right). \tag{3.72}$$

This procedure does not depend on assumptions about the p.d.f. $f(t)$. For details, see Munford and Shahani [89], and additionally Munford and Shahani [90] for the case of the Weibull distribution and Tadikamalla [91] for the gamma distribution.

Near-optimal inspection policy of Nakagawa and Yasui (N&Y policy)
This procedure is based on one of Barlow *et al.* [86], [2] (abbreviated to *B policy* below). If the p.d.f. $f(t)$ is a PF_2, the following algorithm is obtained:

begin:
  choose d appropriately for $0 < d < \frac{c_i}{k_f}$;
  determine t_n after sufficient time has elapsed to give the degree of accuracy required;
  compute t_{n-1} to satisfy

$$t_n - t_{n-1} - d = \frac{F(t_n) - F(t_{n-1})}{f(t_n)} - \frac{c_i}{k_f} \ ;$$

  repeat
    compute $t_{n-2} > t_{n-3} > \cdots$ recursively using (3.66);
  until $t_i < 0$ or $t_{i+1} - t_i > t_i$;
end.

For details, see Nakagawa and Yasui [92].

Remark 3.5.1 From numerical comparisons with Weibull and gamma distributions, we conclude that there are no significant differences between the optimal and near–optimal inspection policies [95], and consequently we should adopt the that is policy simplest to compute, that of Kaio and Osaki [87]. There are the following advantages when we use the *K&O policy*:
(i) We can obtain the nearly optimal inspection policy uniquely, immediately and easily from (3.68) and (3.69) for any distributions, while the *B* and *N&Y policies* cannot treat non–PF_2 distributions.
(ii) We can analyze more complicated models and easily obtain their near–optimal inspection policies; for example, see Kaio and Osaki [87].

3.6 Conclusions

This chapter has been concerned with basic preventive maintenance policies and their variations, in terms of both continuous and discrete–time modeling. For further detail on the discrete models, see Nakagawa [29], [84] and Nakagawa and Osaki [85]. Since the mathematical maintenance models are applicable to a variety of real problems, such a modeling technique will be useful for practitioners and researchers. Though we have reviewed only the most basic maintenance models in this limited space, a number of earlier models should be re–formulated in a discrete–time setting, because, in most cases, the continuous–time models can be regarded as approximated models for actual maintenance problems and the maintenance schedule is often desired in discretized circumstance. These motivations for discrete–time setting will be evident from the recent development of computer technologies and their related computation abilities.

Acknowledgments

This work was partially supported by a Grant–in–Aid for Scientific Research from the Ministry of Education, Sports, Science and Culture of Japan Grants (No. 09780411 and No. 09680426), by the Research Program 1999 of the Institute for Advanced Studies of the Hiroshima Shudo University, Hiroshima, Japan, and by Nanzan University Pache Research Subsidy I–A.

References

1. Arrow, K. J., Karlin, S. and Scarf, H. (eds.) (1962), Studies in Applied Probability and Management Science. Stanford University Press, California
2. Barlow, R. E. and Proschan, F. (1965), Mathematical Theory of Reliability. John Wiley & Sons, New York
3. Barlow, R. E. and Proschan, F. (1975), Statistical Theory of Reliability and Life Testing: Probability Models. Holt, Rinehart and Winston, New York
4. Jorgenson, D. W., McCall, J. J. and Radner, R. (1967), Optimal Replacement Policy. North–Holland, Amsterdam
5. Gnedenko, B. V., Belyayev, Y. K. and Solovyev, A. D. (1969), Mathematical Methods of Reliability Theory. Academic Press, New York
6. Gertsbakh, I. B. (1977), Models of Preventive Maintenance. North–Holland, Amsterdam
7. Ascher, H. and Feingold, H. (1984), Repairable Systems Reliability. Marcel Dekker, New York
8. Osaki, S. (1985), Stochastic System Reliability Modeling. World Scientific, Singapore
9. Osaki, S. (1992), Applied Stochastic System Modeling. Springer–Verlag, Berlin
10. Osaki, S. and Hatoyama, Y. (eds.) (1984), Stochastic Models in Reliability Theory, Lecture Notes in Economics and Mathematical Systems, **235**. Springer–Verlag, Berlin
11. Osaki, S. and Cao, J. (eds.) (1987), Reliability Theory and Applications. World Scientific, Singapore
12. Ozekici, S. (ed.) (1996), Reliability and Maintenance of Complex Systems. NATO ASI Series, Springer, Berlin
13. Christer, A. H., Osaki, S. and Thomas, L. C. (eds.) (1997), Stochastic Modelling in Innovative Manufacturing. Lecture Notes in Economics and Mathematical Systems, **445**, Springer–Verlag, Berlin
14. Barlow, R. E. (1998), Engineering Reliability. SIAM, Philadelphia
15. Aven, T. and Jensen, U. (1999), Stochastic Models in Reliability. Springer–Verlag, New York
16. McCall, J. J. (1965), "Maintenance policies for stochastically failing equipment: a survey," *Management Science*, **11**, 493–521
17. Osaki, S. and Nakagawa, T. (1976), "Bibliography for reliability and availability of stochastic systems," *IEEE Transactions on Reliability*, **R–25**, 284–287
18. Pierskalla, W. P. and Voelker, J. A. (1976), "A survey of maintenance models: the control and surveillance of deteriorating systems," *Naval Research Logistics Quarterly*, **23**, 353–388

19. Sherif, Y. S. and Smith, M. L. (1981), "Optimal maintenance models for systems subject to failure – a review," *Naval Research Logistics Quarterly*, **28**, 47–74
20. Valdez–Flores, C. and Feldman, R. M. (1989), "A survey of preventive maintenance models for stochastically deteriorating single–unit systems," *Naval Research Logistics Quarterly*, **36**, 419–446
21. Feller, W. (1957), An Introduction to Probability Theory and Its Applications. John Wiley & Sons, New York
22. Karlin, S. and Taylor, H. M. (1975), A First Course in Stochastic Processes. Academic Press, New York
23. Karlin, S. and Taylor, H. M. (1981), A Second Course in Stochastic Processes. Academic Press, New York
24. Taylor, H. M. and Karlin, S. (1984), An Introduction to Stochastic Modeling. Academic Press, New York
25. Ross, S. M. (1970), Applied Probability Models with Optimization Applications. Holden–Day, San Francisco
26. Munter, M. (1971), "Discrete renewal processes," *IEEE Transactions on Reliability*, **R–20**, 46–51
27. Kaio, N. and Osaki, S. (1988), "Review of discrete and continuous distributions in replacement models," *International Journal of Systems Science*, **19**, 171–177
28. Barlow, R. E. and Hunter, L. C. (1960), "Optimum preventive maintenance policies," *Operations Research*, **8**, 90–100
29. Nakagawa, T. (1979), "A summary of discrete replacement policies," *European Journal of Operational Research*, **17**, 382–392
30. Tilquin, C. and Cleroux, R. (1975), "Block replacement with general cost structure," *Technometrics*, **17**, 291–298
31. Berg, M. and Epstein, B. (1976), "A modified block replacement policy," *Naval Research Logistics Quarterly*, **23**, 15–24
32. Morimura, H. (1970), "On some preventive maintenance policies for IFR," *Journal of the Operations Research Society of Japan*, **12**, 94–124
33. Tilquin, C. and Cleroux, R. (1975), "Periodic replacement with minimal repair at failure and adjustment costs," *Naval Research Logistics Quarterly*, **22**, 243–254
34. Cleroux, R., Dubuc, S. and Tilquin, C. (1979), "The age replacement problem with minimal repair and random repair costs," *Operations Research*, **27**, 1158–1167
35. Park, K. S. (1979), "Optimal number of minimal repairs before replacement," *IEEE Transactions on Reliability*, **R–28**, 137–140
36. Nakagawa, T. (1979), "A summary of periodic replacement with minimal repair at failure," *Journal of the Operations Research Society of Japan*, **24**, 213–228
37. Nakagawa, T. (1981a), "Generalized models for determining optimal number of minimal repairs before replacement," *Journal of the Operations Research Society of Japan*, **24**, 325–357
38. Nakagawa, T. (1981b), "Modified periodic replacement with minimal repair at failure," *IEEE Transactions on Reliability*, **R–30**, 165–168
39. Nakagawa, T. (1984), "Optimal policy of continuous and discrete replacement with minimal repair at failure," *Naval Research Logistics Quarterly*, **31**, 543–550
40. Nakagawa, T. (1986), "Periodic and sequential preventive maintenance policies," *Journal of Applied Probability*, **23**, 536–542

41. Nakagawa, T. and Kowada, M. (1983), "Analysis of a system with minimal repair and its application to replacement policy," *European Journal of Operational Research*, **12**, 176–182
42. Phelps, R. I. (1981), "Replacement policies under minimal repair," *Journal of the Operational Research Society*, **32**, 549–554
43. Berg, M. and Cleroux, R. (1982), "The block replacement problem with minimal repair and random repair costs," *Journal of Statistical Computation and Simulation*, **15**, 1–7
44. Boland, P. J. (1982), "Periodic replacement when minimal repair costs vary with time," *Naval Research Logistics Quarterly*, **29**, 541–546
45. Boland, P. J. and Proschan, F. (1982), "Periodic replacement with increasing minimal repair costs at failure," *Operations Research*, **30**, 1183–1189
46. Block, H. W., Borges, W. S. and Savits, T. H. (1985), "Age–dependent minimal repair," *Journal of Applied Probability*, **22**, 370–385
47. Beichelt, F. (1993), "A unifying treatment of replacement policies with minimal repair," *Naval Research Logistics Quarterly*, **40**, 51–67
48. Tahara, A. and Nishida, T. (1975), "Optimal replacement policy for minimal repair model," *Journal of the Operations Research Society of Japan*, **18**, 113–124
49. Phelps, R. I. (1983), "Optimal policy for minimal repair," *Journal of the Operational Research Society*, **34**, 425–427
50. Segawa, Y., Ohnishi, M. and Ibaraki, T. (1992), "Optimal minimal–repair and replacement problem with average dependent cost structure," *Computers & Mathematics with Applications*, **24**, 91–101
51. Ohnishi, M. (1997), "Optimal minimal–repair and replacement problem under average cost criterion: optimality of (t, T)–policy," *Journal of the Operations Research Society of Japan*, **40**, 373–389
52. Bergman, B. (1980), "On the optimality of stationary replacement strategies," *Journal of Applied Probability*, **17**, 178–186
53. Barlow, R. E. and Hunter, L. C. (1961), "Reliability analysis of a one–unit system," *Operations Research*, **9**, 200–208
54. Osaki, S. and Nakagawa, T. (1975), "A note on age replacement," *IEEE Transactions on Reliability*, **R–24**, 92–94
55. Fox, B. L. (1966), "Age replacement with discounting," *Operations Research*, **14**, 533–537
56. Nakagawa, T. and Osaki, S. (1977), "Discrete time age replacement policies," *Operational Research Quarterly*, **28**, 881–885
57. Glasser, G. J. (1967), "The age replacement problem," *Technometrics*, **9**, 83–91
58. Scheaffer, R. L. (1971), "Optimum age replacement policies with an increasing cost factor," *Technometrics*, **13**, 139–144
59. Cleroux, R. and Hanscom, M. (1974), "Age replacement with adjustment and depreciation costs and interest charges," *Technometrics*, **16**, 235–239
60. Osaki, S. and Yamada, S. (1976), "Age replacement with lead time," *IEEE Transactions on Reliability*, **R–25**, 344–345
61. Nakagawa, T. and Osaki, S. (1976), "Reliability analysis of a one–unit system with unrepairable spare units and its optimization applications," *Operational Research Quarterly*, **27**, 101–110
62. Derman, C. and Sacks, J. (1960), "Replacement of periodically inspected equipment," *Naval Research Logistics Quarterly*, **7**, 597–607

63. Ansell, J., Bendell, A. and Humble, S. (1984), "Age replacement under alternative cost criteria," *Management Science*, **30**, 358–367
64. Christer, A. H. (1978), "Refined asymptotic costs for renewal reward process," *Journal of the Operational Research Society*, **29**, 577–583
65. Christer, A. H. (1987), "Comments on finite–period applications of age–based replacement models," *IMA Journal of Mathematics Applied in Business and Industry*, **1**, 111–124
66. Christer, A. H. and Jack, N. (1991), "An integral–equation approach for replacement modelling over finite time horizons," *IMA Journal of Mathematics Applied in Business and Industry*, **3**, 31–44
67. Wiggins, A. D. (1967), "A minimum cost model of spare parts inventory control," *Technometrics*, **9**, 661–665
68. Allen, S. G. and D'Esopo, D. A. (1968), "An ordering policy for repairable stock items," *Operations Research*, **16**, 669–674
69. Allen, S. G. and D'Esopo, D. A. (1968), "An ordering policy for stock items when delivery can expedited," *Operations Research*, **16**, 880–883
70. Simon, R. M. and D'Esopo, D. A. (1971), "Comments on a paper by S. G. Allen and D. A. D'Esopo: 'an ordering policy for repairable stock items'," *Operations Research*, **19**, 986–989
71. Nakagawa, T. and Osaki, S. (1974), "Optimum replacement policies with delay," *Journal of Applied Probability*, **11**, 102–110
72. Osaki, S. (1977), "An ordering policy with lead time". *International Journal of Systems Science*, **8**, 1091–1095
73. Thomas, L. C. and Osaki, S. (1978), "A note on ordering policy," *IEEE Transactions on Reliability*, **R–27**, 380–381
74. Thomas, L. C. and Osaki, S. (1978), "An optimal ordering policy for a spare unit with lead time," *European Journal of Operational Research*, **2**, 409–419
75. Kaio, N. and Osaki, S. (1978), "Optimum ordering policies with lead time for an operating unit in preventive maintenance," *IEEE Transactions on Reliability*, **R–27**, 270–271
76. Kaio, N. and Osaki, S. (1978), "Optimum planned maintenance with salvage costs," *International Journal of Production Research*, **16**, 249–257
77. Kaio, N. and Osaki, S. (1979), "Discrete–time ordering policies," *IEEE Transactions on Reliability*, **R–28**, 405–406
78. Kaio, N. and Osaki, S. (1980), "Optimum planned maintenance with discounting," *International Journal of Production Research*, **18**, 515–523
79. Osaki, S., Kaio, N. and Yamada, S. (1981), "A summary of optimal ordering policies," *IEEE Transactions on Reliability*, **R–30**, 272–277
80. Dohi, T., Kaio, N. and Osaki, S. (1998), "On the optimal ordering policies in maintenance theory – survey and applications," *Applied Stochastic Models and Data Analysis*, **14**, 309–321
81. Dohi, T., Kaio, N. and Osaki, S. (1995), "Continuous review cyclic inventory models with emergency order," *Journal of the Operations Research Society of Japan*, **38**, 212–229
82. Dohi, T., Shibuya, T. and Osaki, S. (1997), "Models for 1–out–of–Q systems with stochastic lead times and expedited ordering options for spares inventory," *European Journal of Operational Research*, **103**, 255–272

83. Shibuya, T., Dohi, T. and Osaki, S. (1998), "Spare part inventory models with stochastic lead times," *International Journal of Production Economics*, **55**, 257–271
84. Nakagawa, T. (1986), "Modified discrete preventive maintenance policies," *Naval Research Logistics Quarterly*, **33**, 703–715
85. Nakagawa, T. and Osaki, S. (1976), "Analysis of a repairable system which operates at discrete times," *IEEE Transactions on Reliability*, **R–25**, 110–112
86. Barlow, R. E., Hunter, L. C. and Proschan F. (1963), "Optimum checking procedures," *SIAM Journal of Applied Mathematics*, **11**, 1078–1095
87. Kaio, N. and Osaki, S. (1984), "Some remarks on optimum inspection policies," *IEEE Transactions on Reliability*, **R–33**, 277–279
88. Kaio, N. and Osaki, S. (1984), "Analytical considerations on inspection policies," in Stochastic Models in Reliability Theory (Osaki, S. and Hatoyama, Y. eds.). 53–71, Springer–Verlag, Heidelberg
89. Munford, A. G. and Shahani, A. K. (1972), "A nearly optimal inspection policy," *Operational Research Quarterly*, **23**, 373–379
90. Munford, A. G. and Shahani, A. K. (1973), "An inspection policy for the Weibull case," *Operational Research Quarterly*, **24**, 453–458
91. Tadikamalla, P. R. (1979), "An inspection policy for the gamma failure distributions," *Journal of the Operational Research Society*, **30**, 77–80
92. Nakagawa, T. and Yasui, K. (1980), "Approximate calculation of optimal inspection times," *Journal of the Operational Research Society*, **31**, 851–853
93. Keller, J. B. (1974), "Optimum checking schedules for systems subject to random failure," *Management Science*, **21**, 256–260
94. Kaio, N. and Osaki, S. (1986), "Optimal inspection policies: A review and comparison," *Journal of Mathematical Analysis and Applications*, **119**, 3–20
95. Kaio, N. and Osaki, S. (1989), "Comparison of inspection policies," *Journal of the Operational Research Society*, **40**, 499–503

4. A Review of Delay Time Analysis for Modelling Plant Maintenance

A. H. Christer
Centre for OR & Applied Statistics,
University of Salford
Greater Manchester, U.K.

Summary.

Delay time analysis is a pragmatic mathematical concept readily embraced by engineers, which has been developed as a means to model maintenance decision problems. Attention is focused upon the maintenance engineering decisions of what to do, as opposed to the logistical decisions of how to do it. This paper reviews the cumulative knowledge and experience of delay time modelling. The decision environment within which Delay time models are intended as decision aids is briefly reviewed, and the initial development of simple DT models for a repairable component and a complex plant presented. Variations on the basic model are outlined and discussed, including perfect and nonperfect inspection, steady state and non-steady state conditions, and homogeneous and non-homogeneous Poisson arrival rate of defects. Attention is given to the parameter estimation process, and both subjective and objective estimation techniques are outlined. Case sketches present practical experience in using the DT concept to model actual plant, to assess the benefits obtained, and to validate modelling and parameter assessment. References are given throughout to related work as well as to future developments.

Keywords: delay time concept, maintenance, reliability, engineering, production

4.1 Introduction

This chapter is essentially a paper previously published in the *Journal of the Operational Research Society* Vol 50, 1999, 1120-1137 and reproduced here by kind permission of the Editor.

From time to time estimates are published for particular countries indicating the level of national expenditure on maintenance. One particular study conducted in the Netherlands (Geraerds [1]) considered the split of GDP across expenditure sectors of health, education, defence, ..., and within each of these sectors estimated the level of expenditure ultimately being expended on maintenance activities. The overall conclusion was that

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

some 14% of the Netherlands GDP was consumed by maintenance activities. The breakdown between expenditure areas was 34% industrial plant, 22% buildings, 19% housing, 19% transport, and 6% roads. Interestingly, the author later commented that this distribution had remained relatively stable over subsequent years. Interestingly, the World Development Report, World Band 1998/99 indicates that a 6.7% of GDP within the Netherlands is expended upon public health, which places maintenance expenditure within a relative context at twice the level of public health expenditure. Comparable orders of magnitude can be expected for other developed countries, with perhaps an increase in total level for developing countries.

Two points follow immediately from such estimates. The first is that the figures are enormous, and therefore the topic is worthy of serious and sustained study. The second point is that there is little further one can do with these figures, since one cannot say if the order of magnitude is too high, too low, or about right. Although the implication when such macro figures are presented is invariably of excessive expenditure, in reality the appropriateness of any maintenance expenditure needs to be assessed within the context of a specific industry, plant and circumstance. Whilst exploring maintenance expenditure at the macro level is of interest, it will not of itself solve any resource problems or produce improvements. To do this we need to address specifics and develop maintenance decision models.

Over the last thirty years the OR/MS community world wide has contributed to the area of maintenance decision making through the publication of many hundreds of papers under the general umbrella title of "maintenance". Here 'maintenance' is a Catholic term embracing aspects of plant management including servicing, inspection, repair, overhaul, reliability and replacement. Review papers within the area are numerous [2], [3], [4], [5], [6], [7], [8], [9], [10]. Yet despite this outflow, valid evidence of impact and of modelling being productively used within industry, or even considered for use, is thin on the ground. Very few publications are based upon an actual maintenance situation as opposed to a postulated scenario. Only a small subset use case data, and of these very few present a validation or post study check upon the model or consider observing its influence upon decision making.

There are other views on the scale of impact of OR modelling in maintenance. In a study based upon a postal survey of 200 randomly selected Fortune 500 industrial firms, Hsu [11], it was reported that 89% of firms used modelling in some form for equipment replacement decisions, though it may only have been a discounted cash flow or payback period calculation. This positive picture of modelling use contrasted sharply with the findings of a survey by Christer and Waller [12], who, based upon 20 in-depth studies within the U.K., raised concerns for the perceived role and quality of equipment replacement modelling where it was used. For the more costly equipment surveyed, they reported that modelling often had little actual influence upon decision making because issues were dominated by factors omitted from the analysis. Where modelling was influential, usually for the lower unit value items, the authors raised concerns for the appropriateness of the models and the relevance of the modelling process to the replacement decision. Although some form of discounted cash flow or payback period calculation may have been used, it was seen more as an accounting ritual than as an influential factor informing a decision. Such cases would have contributed to the count in the 89% of model users in Hsu's survey, which highlights the difficulties in forming a meaningful and accurate

picture of modelling use in the maintenance area based upon arms-length postal research. Either way, the modelling concept of delay-time analysis and modelling introduced below does, it is argued, provide a modelling framework readily applicable to a class of actual industrial maintenance problems.

To clarify the objective of the type of modelling we are concerned with here, consider a plant item with a maintenance practice, or concept [13], of servicing every period T, with repair of failures as they arise. The service consists of a check list of activities to be undertaken, and a general inspection of the operational state of the plant. Any defect identified leads to immediate repair, and the objective of the maintenance concept is to minimise operational downtime.

Conceptually, there is a relationship between the expected downtime per unit of time $D(T)$, and the service period T (Figure 4.1). If T were small, the downtime per unit of time would be large because the plant would frequently be unavailable due to servicing, and if T were sufficiently large, the downtime per unit of time would essentially be that under a breakdown maintenance policy. If the chosen service period is T^*, all that we can expect to be known of $D(T)$ is the observed value $D(T^*)$, that is the current downtime measure. One wishes to reduce $D(T^*)$, and if a model such as Figure 4.1 were available, there would be little difficulty in identifying a good operational period for T, which need not be finite. Unfortunately, in the absence of modelling, all that is generally available is the data of Figure 4.2. To move from Figure 4.2 to Figure 4.1 requires

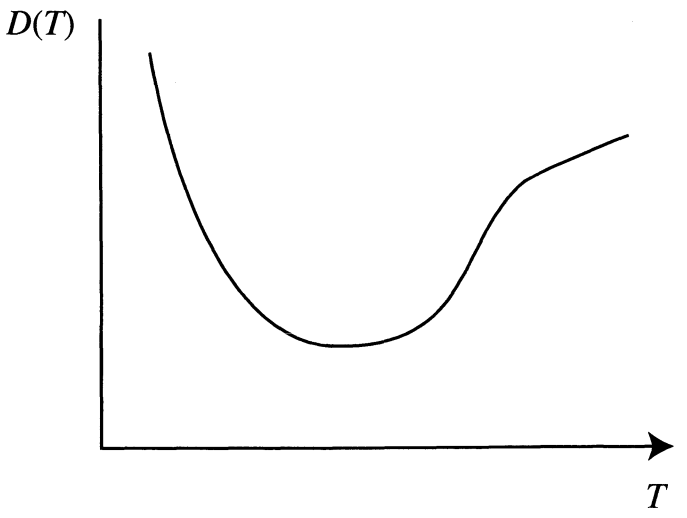


Fig. 4.1. Downtime model: service period T

maintenance modelling, and Figure 4.1 is a graphical representation of the model. The model could equally well be cost based or reliability based. Of course, the above model addresses an engineering decision for a single plant. If there were a large number of items of different plant then even

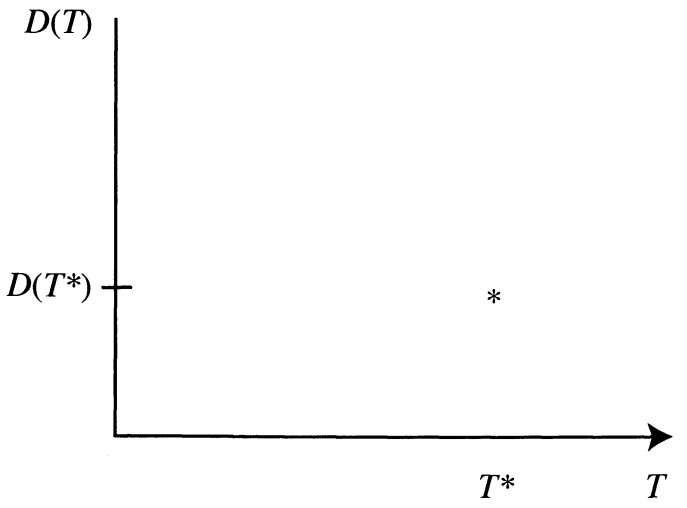


Fig. 4.2. Downtime information

knowing the most appropriate period, T, for each item might still present considerable logistical problems of supplying the correct spares and materials with competent fitters and correct instructions and briefing notes to the appropriate plant at the time scheduled for service. This requires logistical support, scheduling, and spares provisioning which, whilst potentially aided by a mature literature on modelling to guide decision making, is not of the maintenance engineer's decision type being addressed here, but of a more general class of scheduling and inventory control modelling applied within a maintenance context. It has been recognised that in the integrated management of maintenance concepts for sets of plant, economic dependence for set-up costs is possible, and the logistics may require a compromise between the optimal maintenance concept for particular plant items and the associated logistical costs [14], [15], [16] [17]. However, these operational procedures are considerably aided and simplified by the existence of valid maintenance models for plant items such as Figure 4.1. Here we are interested in modelling engineering decisions of what to do, as opposed to modelling the implementation of the maintenance decisions. In short, we wish to model the decision of what to do as opposed to how to do it. Both types of modelling are important to maintenance, but the former is less understood and, in the view of the author, has the greater potential for impact upon industrial performance. Delay time modelling has been developed to have relevance within the operating culture currently found within industry, and before introducing the delay time concept, it is appropriate to characterise the decision environment in which it is to be used. Also, if as suggested above, maintenance modelling such as Figure 4.1 is a useful guide to decision making, we should perhaps consider why its use is not common practice, and how maintenance decisions are currently made in the absence of modelling. This is a relevant question to the modeller, since modelling is common place in related areas of activity, including inventory control, logistics and scheduling.

4.2 Maintenance Practice

The role of management within the maintenance area is traditionally the province of the engineer, with decision making strongly influenced by educational background and industrial experience. Maintenance modelling is unlikely to have featured in either of these environments. Few university engineering programmes address maintenance engineering, and even fewer consider rudimentary concepts of modelling the decision process of maintenance. Currently, the concept of maintenance modelling is a scarce commodity within industry. Engineering judgement and sound common sense have, with good reason, held sway in the area of maintenance for generations. Accumulated experience is essentially in engineering detail, which may be captured for modelling. One observes that the management of maintenance practice within industry has evolved over time, but slowly and with little input from modelling.

In the 1960s maintenance by 'indices' was very popular, with numerous indices of performance being defined [18]. However, without modelling the significance of an index level is unknown. Planned preventive maintenance (PPM) dominated the 1970s, with advocates prescribing what proportion of maintenance activity should be planned [19]. Virtually none of these advocates attempted to model or even consider modelling the problem. Information technology started its impact in the 1980s with the development of computer-based maintenance information systems. The belief was that instant access to past data would solve problems. With relatively few exceptions, one observed that after a while such systems, even when kept updated, were seldom accessed for data and, perhaps not surprisingly, were ultimately degraded. It is argued here that such systems provide more information on the already known point of Figure 4.2, but without skills in data analysis and the concept of maintenance modelling, cannot aid the manager with his key decision problems, that is provide the model of Figure 4.1. This remains true today.

Total production maintenance (TPM; Nikojima [20]) rose in popularity in the 1990s and is based upon much sound engineering practice. But again, we observe that in the absence of maintenance modelling, a TPM search for improvement cannot stop when the best is attained, because the state of 'best' is unknown. Any such search for improvement needs to be guided by modelling if the inefficient consumption of resource in a situation of diminishing returns is to be avoided. Reliability Centre Maintenance (RCM [21]), has also become very popular over the past decade, and has some comparable features to the delay time concept which will be developed below. However, unlike DTM, RCM is a procedure, not a modelling methodology, and is therefore subject to the same criticism as TPM.

All the above 'fashions' have two common features. First they are prescriptive in that they propose procedures allegedly leading to improvement, and secondly, there is a lack of any underpinning scientific concept, testing, verification or validation. Fashions driving management practice appear to have sometimes been embraced remarkably quickly. When directors responsible for annual maintenance spends of $30M to several billion have been asked how much is spent researching the effectiveness of the spend, the answer has been zero, or worse, 'I don't understand the question'. Compare this with the marking function, where a spend of a few hundred thousand dollars will typically be followed up with a further expenditure exploring

the effectiveness of the spend and lessons learnt. Clearly there is a problem here and a culture to be changed.

Condition-based maintenance is currently very common within industry. Here, maintenance action, or lack of it, is based upon measures of condition information which could be provided by oil analysis, vibration analysis, thermal techniques, ... Although there is an abundance of technical literature on these techniques, the associated decision problem is relatively untouched [22]. Aghjagan [23] reported that since introducing a condition monitoring regime based upon oil analysis of gear boxes, the incident of gearbox failure whilst in use of all the Canadian Pacific locomotives had fallen by 90%. This is a notable accolade for condition monitoring. However, it transpired during reconditioning of 'defective' gearboxes that in 50% of occasions there was no evident gearbox fault. Seemingly, condition monitoring can be at the same time very effective and rather inefficient. This is another example where the maintenance decision process requires modelling.

To increase the uptake of valid modelling within the maintenance function, two developments are necessary. First, maintenance engineers need to be informed of the service OR/MS can supply to aid the management of the maintenance process. This would be aided by increasing the number of validated case histories. The second is to interest OR/MS scientists in addressing actual industrial maintenance problems by highlighting the observed scope for contributing in practice [24]. This paper has been written to stimulate the second of these, and the remainder of this paper introduces an approach to modelling the engineering aspects of maintenance, called delay time analysis and modelling, and presents a review of developments. Though still being developed by, amongst others, colleagues at Salford University, this concept has nevertheless enjoyed success in modelling and has influenced maintenance practice.

4.3 The Delay Time Concept

We are interested in the relationship between the performance of equipment and maintenance intervention, and to capture this the conventional reliability analysis of time to first failure, or time between failures [25], [26] requires enrichment. Consider a repairable item of plant. It could be, say, a component, a machine, or an integrated set of machines forming a production line, but viewed by management as a plant unit. The interaction between maintenance concept and equipment performance may be captured using the delay time concept presented below.

Let the item of plant be maintained on a breakdown basis. The time history of breakdown or failure events is a random series of points (see Figure 4.3). For any one of these failures, the likelihood is that had the

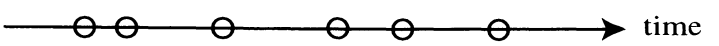


Fig. 4.3. Failure points 'o'

plant been inspected at some point just prior to failure, it could have been seen that all was not well and a defect was present which, though the plant

was still working, would ultimately lead to a failure. Such signals include excessive vibration, unusual noise, excessive heat, surface staining, smell, reduced output, increased quality variability, ... The first instance where the presence of a defect might reasonably be expected to be recognised by an inspection is called the initial point u of the defect, and the time h to failure from u is called the delay time of the defect, see Figure 4.4. Had an

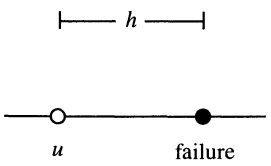


Fig. 4.4. The delay time for a defect

inspection taken place in $(u, u+h)$, the presence of a defect could have been noted and corrective action taken prior to a failure. Given that a defect arises, its delay time represents a window of opportunity for preventing a failure. Clearly, the delay time h is a characteristic of the plant concerned, the type of defect, the nature of any inspection, and perhaps the person inspecting. For example, if the plant was a vehicle, and the maintenance practice was to respond when the driver reported a problem, then there is in effect a form of continuous monitoring inspection of cab-related aspects of the vehicle, with a reasonably long delay time consistent with the rate of deterioration of the defect. However, should the exhaust collapse because a support bracket was corroded through, the likely warning period for the driver, the delay time, would be virtually zero, since he would not normally be expected to look under the vehicle. At the same time, had an inspection been undertaken by a service mechanic, the delay time might have been measured in weeks or months. Had the exhaust collapsed because securing bolts became loose before falling out, then the driver could have a warning period of excessive vibration, and perhaps noise, and the defect have a driver-related delay time measured in days or weeks.

To see why the delay time concept is of use, consider Figure 4.5 incorporating the same failure point pattern as Figure 4.3 along with the initial points associated with each failure arising under a breakdown system. Had

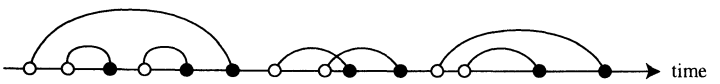


Fig. 4.5. 'o' initial points; '•' failure points

an inspection taken place at point (A), one defect could have been identified and the seven failures reduced to six. Likewise, had inspection taken place at points (B) and point (A), four defects could have been identified and the seven failures now reduced to three. Figure 4.5 demonstrates that provided it is possible to model the way defects arise, that is the rate of arrival of defects $\lambda(u)$, and their associated delay time h, then the delay time concept can capture the relationship between inspection frequency

and the number of plant failures. As first explored in the appendix to Christer [27], knowing the relationship then enables the construction of a delay-time-based maintenance model.

4.4 Basic Delay Time Maintenance Model: Complex Plant

A complex plant, or multi-component plant, is one where a large number of failure modes arise, and the correction of one defect or failure has nominal impact in the steady state upon the overall plant failure characteristics. Consider the following basic complex plant maintenance modelling scenario where:

(a) An inspection takes place every T time units, costs C_I units and requires D_I time units, where $D_I << T$.
(b) Inspections are perfect in that all (and only) defects present will be identified.
(c) Defects identified will be repaired during the inspection period.
(d) Defects arise at a constant rate λ per unit of time.
(e) The probability density function for delay time of faults $f(h)$ is independent of the initial point u.
(f) Failure will be repaired immediately at an average cost c_f and downtime d_f.
(g) The plant has operated sufficiently long since new to be effectively in a steady state.
(h) Defects and failures only arise whilst plant is operating.

These assumptions characterise the simplest non-trivial inspection problem. Under these assumptions, for a defect with delay time h, the expected number of breakdowns over $(0, T)$, $EN_f(T)$, is given by Christer and Wang [28],

$$EN_f(T) = \lambda \int_0^T F(T-u)du$$

and therefore, the probability $b(T)$ that a fault arising causes a breakdown given inspection period T is

$$b(T) = \left(\frac{EN_f(T)}{\lambda T}\right) = \frac{1}{T}\int_0^T F(T-u)du \tag{4.1}$$

which increases from 0 to 1 as T increases from 0 to infinity. This expression is a basic building block for modelling the inspection aspect of maintenance.

Accepting these assumptions, the expected number of failures over an inspection period is $\lambda T b(T)$, and the expected cost per unit time $C(T)$ and downtime per unit time $D(T)$ become

$$C(T) = \left\{\frac{C_I + C_f \lambda T b(T)}{T + D_I}\right\} \tag{4.2}$$

and

$$D(T) = \left\{ \frac{D_I + d_f \lambda T b(T)}{T + D_I} \right\} \tag{4.3}$$

respectively. In this formulation, it has been assumed that $\lambda d_f b(T) << 1$, that is in estimating the number of failures over an inspection period T, it is not necessary to make allowance for the loss of operating period due to failures. If this condition is not valid, a modification to the model is readily made [29], [30]. Equations (4.2) and (4.3) are delay-time-based maintenance models of the inspection practice, and (4.3) is a particular form of the conceptual curve of Figure 4.1. Both criterion models (4.2) and (4.3) clearly exhibit the expected characteristics of having large values for small T, and having an asymptotic value for large T corresponding to the breakdown system value of $C = \lambda c_f$ and $D = \lambda d_f$.

The existence of a delay time period associated with failures is not new to engineers. Its existence has been the rationale underpinning PM and inspection systems for years, and more recently the concept of reliability-centred maintenance [21]. What is new, however, is an attempt to capture the relationship quantitatively and exploit it in modelling maintenance practice. Planned maintenance activity will usually embrace two activities. First, preventive actions such as cleaning, greasing, oil changes and possibly age- or use-based replacements, all of which are designed to prevent defects arising, that is influence and hopefully reduce $\lambda(u)$. Secondly, the inspection element of PM, which accepts defects arise, but seeks to reduce the consequences, namely failures, by identifying defects for corrective action before failure. As such, inspections do not influence defects arising, but they should influence the number of failures. The nature of inspections and the definition of failure adopted will usually be consistent with custom and practice within a client organisation.

4.5 Basic Maintenance Model: Component Tracking

A second basic type of delay time model concerns the inspection maintenance of a repairable component which is assumed to have a single failure mode. At most one defect can exist at any given time, and the time from stochastically new to the initial point of a defect, u, is now governed by a pdf $g(u)$ with c.d.f. $G(u)$ say. The c.d.f. of time to failure given no intervention, $P(t)$ say, is the convolution of $g(u)$ and $f(h)$.

$$P(t) = \int_{u=0}^{t} g(u) F(t-u)\, du$$

and the component reliability at time t is

$$R(t) = 1 - P(t).$$

Suppose a component is replaced or repaired both at failure and when detected at inspection to be in a defective state. A similar case has been considered by Cox [31] based upon component wear level. We assume perfect inspections, and the special case where $g(u)$ is negative exponential, which implies inspections are effectively renewal points. Assume a failure replacement, inspection replacement and inspection require downtimes of

d_f, d_r and d_i respectively. The expected downtime $D(T)$ of an inspection cycle of length T is given by the sum of the expected downtimes over $(0, T)$ when no defect arises, when a defect arises and is identified at T, and when a failure arises. That is

$$\begin{aligned} D(T) &= d_f P(T) + (d_r + d_i) \int_{u=0}^{T} g(u)[1 - F(T-u)]\, du \\ &\quad + d_i(1 - G(t)) \\ &= (d_f - d_r - d_i)P(T) + d_r G(T) + d_i. \end{aligned}$$

Similarly, the expected cycled length $M(T)$ is given by

$$\begin{aligned} M(T) &= (T + d_i)(1 - G(T)) + (T + d_r + d_i)P(T) \\ &\quad + \left(d_f P(T) + \int_0^T R(t)\, dt - TR(T) \right), \end{aligned}$$

and the downtime per unit time measure to inform the choice of T becomes

$$\frac{D(T)}{M(T)}.$$

This expression is equivalent to Figure 4.1 for the basic D.T. component tracking case.

The assumptions of this case may be relaxed in the same way as those of the basic maintenance model for complex plant. Indeed, Baker and Wang [32] have investigated the case where component age influences the initial point distribution $g(u)$ and delay time distribution $f(h)$, where u and h are possibly correlated, and where an inspection can have a hazardous effect upon component life. However, the relaxation will be discussed mainly in the complex plant case, since this is the more common case encountered within industry.

4.6 Relaxation of Assumptions

4.7 Non-perfect Inspection

Perhaps the most suspect assumption in the above modelling is that of perfect inspection. If a defect has a probability r of being identified at a non-perfect inspection, then the above delay time models of (4.2) and (4.3) for $C(T)$ and $D(T)$ are still valid for the steady state, but with modified $b(T)$. Of the λT defects expected to arise over an inspection period T, the expected number resulting in failures is [69]

$$EN_f(T) = \lambda \sum_{n=1}^{\infty} \int_{u=0}^{T} r(1-r)^{n-1} R(nT - u) du.$$

Consequently, the probability $b(T)$ that a defect will give rise to a failure given non-perfect inspections of period T is

$$b(T) = \frac{1}{T} \sum_{n=1}^{\infty} r(1-r)^n \int_0^T F(nT - u) du, \tag{4.4}$$

where

$$F(x) = \int_0^x f(h)dh.$$

4.8 Non-steady-state Condition

All the plant actually modelled to date using the delay time concept has been sufficiently old to satisfy assumption (g) concerning steady state. Should this assumption be in question, it is a relatively straightforward task to relax it. This is only necessary, of course, if $r < 1$ since $\lambda(u)$ is currently assumed to be constant. Suppose, for example, a plant item has been inspected every period T since new, and its age is now nT. We concern ourselves with both the average downtime per unit of time to date, and during the last inspection cycle $((n-1)T, nT)$ (see Figure 4.6).

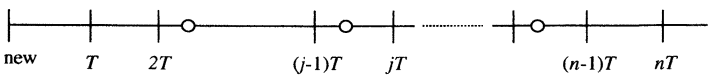


Fig. 4.6. Non-steady-state inspection case

Let N_j denote the expected number of failures occurring in operating period $((j-1)T, jT)$. Since $N_j = N_{j-1}+$ failure arising from defects originating in the first inspection period $(0, T)$, we have

$$N_j = \begin{cases} N_{j-1} + (1-r)^{j-1}\lambda \int_{u=0}^{T} (F(jT-u) \\ \quad -F((j-1)T-u))du & \text{for } j > 1, \\ \lambda \int_{u=0}^{T} F(T-u)du, & \text{for } j = 1. \end{cases} \tag{4.5}$$

The total expected number of failure repairs over $(0, nT)$ is $\sum_{j=1}^{n} N_j$, and from (4.5) it can be shown that

$$N_j = \begin{cases} \lambda(1-r)^{j-1} \int_0^T F(jT-u)du \\ \quad +\lambda \sum_{i=1}^{j-1} (r(1-r)^{i-1} \int_0^T F(iT-u)du) & \text{for } j > 1, \\ \lambda \int_{u=0}^{T} F(T-u)du, & \text{for } j = 1. \end{cases}$$

Clearly, the long-term probability of a defect arising as a breakdown is

$$\lim_{n\to\infty} \left(\frac{\sum_{j=1}^{n} N_j}{\lambda nT} \right). \tag{4.6}$$

If interest was in selecting inspection period T to control the total downtime over a fixed time period τ, then for any τ, n^* inspection cycles would be completed where

$$n^*(T + D_I) \leq \tau < (n^* + 1)(T + D_I).$$

The total downtime over $(0, \tau)$ now becomes

$$\Gamma(\tau; T) = \left(n^* D_I + d_f \sum_{j=1}^{n^*} N_j + \epsilon \right), \tag{4.7}$$

where ϵ represents the contribution to expected downtime due to failures arising over $(n^*(T+D_I), \tau)$. The form of ϵ depends upon operating practice when failures occur near the end point τ. Should τ be known to within a density function $\Theta(\tau)$, then an objective function which has been used in a related context [33] is

$$\min_T \left\{ \int_\tau \Gamma(\tau; T) d\Theta(\tau) \right\}.$$

Of interest in some cases for monitoring and accounting purposes would be the expected downtime per unit of time measure over the last, or next, operating period $((n-1)T, nT)$, namely

$$\frac{D_I + d_f N_n}{T + D_I}. \tag{4.8}$$

It can be shown that downtime measure (4.8) will tend to the steady state unit of time measure of (4.3) as n increases. Cases where the steady state assumption may not be valid are evident when the objective function is the reliability of plant or a repairable component. Here interest is usually restricted to a finite operating period or to a mission time. Exploratory modelling work has established the potential for DTM to aid decision making in such cases, [34], [35], [36], [37], Reliability modelswill more naturally be of the component tracking type discussed below. The important point here is that the requirements for a steady state situation may be readily relaxed in delay-time modelling at the expense of some additional mathematical detail.

4.9 Non-homogeneous Defect Arrival Rate λ

It has been the author's experience that the assumption of constant defect arrival rate for mature plant has been justified for most plant, but not all. In a study of the maintenance practice of the roll change mechanism of a high-tech steel rolling mill [38], evidence was found that the rate of arrival of defects in the plant was higher just after a periodic planned maintenance intervention (PM), but subsequently settled down to a steady level prior to the next PM. If there was to be any change in λ, one would expect it to be lower just after PM and rising with time, not vice versa. Clearly something is wrong with the PM, and having been highlighted, the problem becomes predominantly an engineering issue. However, the point is that cases exist where because of perhaps wear mechanisms or the nature of human intervention, the arrival rate λ of defects is time dependent, that is $\lambda = \lambda(u)$.

Suppose all the above basic modelling assumptions (a)–(g) are valid excepting (d), where now $\lambda = \lambda(u)$, and u is the operating time from when

the plant was stochastically as new. Defects and failures are assumed not to arise or occur during an inspection period downtime. Criterion measures of interest such as the downtime per unit time, will now depend upon the time zone $(0 - t)$ over which the measurement is made, or the time t at which a local measure is taken. If the time of interest is t, then for some integer n, we have $nT \leq t < (n+1)T$ (see Figure 4.7). Because of perfect inspections, the expected number of failures occurring over $((j-1)T, jT)$ is given by

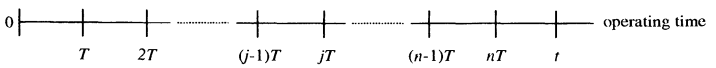


Fig. 4.7. $n \leq T < (n+1)T$

$$N_j = \int_{(j-1)T}^{jT} \lambda(u)F(jT-u)du \tag{4.9}$$

with the expected number of defects identified at the j-th inspection being

$$\int_{(j-1)T}^{jT} \lambda(u)du - N_j.$$

Over the period t, we have the overall expected downtime $D(t)$ given by

$$D(t) = nD_I + d_f \left[\sum_{j-1}^{n} \int_{(j-1)T}^{jT} \lambda(u)F(jT-u)du + \int_{nT}^{t} \lambda(u)F(t-u)du \right], \qquad nT \leq t < (n+1)T. \tag{4.10}$$

Likewise, the expected downtime per unit time measured over the last complete inspection period is

$$\frac{D_I + d_f \int_{(n-1)T}^{nT} \lambda(u)F(nT-u)du}{T + D_I}. \tag{4.11}$$

If conditions are non-steady, the assumptions of constant inspection period T needs to be relaxed. This assumption is normally not relaxed lightly since to operate a variable inspection period policy requires evidence of sufficient benefit to justify the operational inconvenience. However, it is possible to assume an inspection policy $\underline{T}(T_1, T_2, \ldots)$ and optimise the objective function with respect to the inspection vector $\underline{T}$. This has been investigated in the case of a component tracking model [39] [42] and as expected, the nature of the problem changes. If defects arise at an increasing rate and inspections accordingly become more frequent, the point could arise where it is appropriate to replace the plant. Now, instead of just searching for the best inspection practice, one has to consider integrating inspection modelling with the modelling of a replacement decision (or overhaul to a much improved condition). In this case, it may be necessary to consider a finite modelling horizon and consider the influence of taxation upon capital expenditure [40], [41].

4.10 Condition-dependent Cost and Downtime for Repair

Although most of the developments of the delay time concept in maintenance modelling have been in the context of industrial mechanical plant, where characteristic timescales are days and perhaps weeks, initially the concept was developed in the area of building maintenance, where delay time scales may be measured in months and perhaps years [43], [44]. Over a long delay time period, a building defect would characteristically increase in severity and therefore cost of repair (see Figure 4.8).

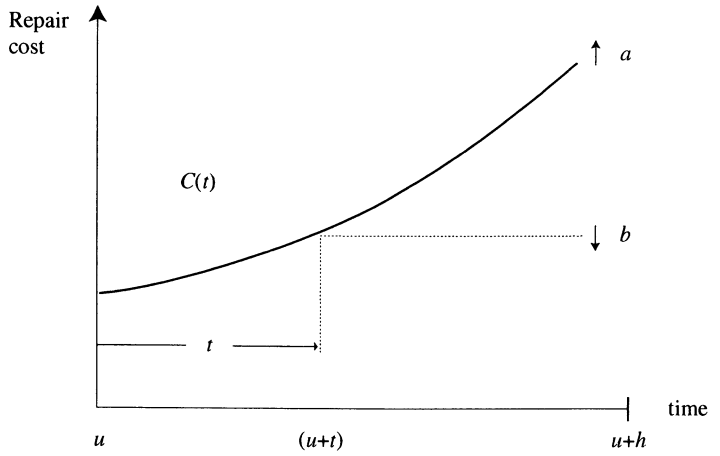


Fig. 4.8. Repair cost over delay time h; u = initial point

If $C(\tau)$ is the cost of repair at time $(\tau + u)$, then by inspecting at time τ after the initial point a saving of $(C(h) - C(\tau))$ is possible. In this case, repair cost savings as well as failure avoidance are possible. This aspect has been developed further and applied in a civil engineering context to model the degradation and maintenance of concrete structures [45], but for industrial plant the curve $C(\tau)$ is believed in the main to be relatively flat, with perhaps a discontinuous increase at the failure point because of the inconvenience and possible additional cost of associated failure damage. In the remainder of this paper, we will assume the repair cost and repair downtime of a defect to be constant over the delay time period.

Other assumptions in the modelling may likewise be relaxed, but the important point for the present is that providing $\lambda(u)$ and $f(h)$ can be estimated, the delay time concept should be capable of producing the target inspection model of Figure 4.1.

The option might exist to subdivide defect types into independent clusters with common arrival patterns $\lambda_i(u)$ or $g_i(u)$, and pdf's $f_i(h)$, and thereby construct more refined models. Such options are developments in modelling detail and not principle, and as such will not be considered further here. As an aid to qualitative modelling, the delay time concept only has practical value if delay time parameters can be estimated. We now consider the techniques established to estimate $f(h)$, $\lambda(u)$ and r.

4.11 Case Experience Using Subjective Data: Case Experience

The first industrial study requiring the estimation of delay time parameters was the downtime modelling of a complex high-speed canning line [46]. Cans can be filled and sealed at a rate of up to 1,000 per minute, and the line operated 24 hours per day, 7 days a week, 50 weeks a year. For the previous 5 years, since new, the maintenance concept had been to stop production once every 24 hours to inspect the line and rectify identified faults. These brief events (20–30 minutes) were called pit-stops. Breakdowns were rectified when they occurred. Management wanted to know if the 24-hour inspection period could be improved upon to further reduce production downtime.

Before attempting to model the inspection of an existing plant, thought needs to be given to identifying engineering solutions to reduce the rate of defects arising, that is $\lambda(u)$. This entails considering what actually causes defects, as opposed to the usually recorded maintenance information system data on what the defect was. A snap-shot survey system was established to collect detailed information and assessment at every maintenance intervention over a 6-week period [47]. Part of this survey was designed to provide an estimate of the delay time for each defect as follows:

At every breakdown, the maintenance fitter repairing the plant would be asked to estimate

HLA: How long ago the defect causing the failure may first have been expected to have been recognised at an inspection.

If a defect was identified at an inspection, then in addition to HLA, the fitter would be asked to estimate

HML: How much longer the defect could be left unattended before repair was essential.

The estimates are given by (see Figures 4.9 and 4.10) $\hat{h}$ =HLA for a break-

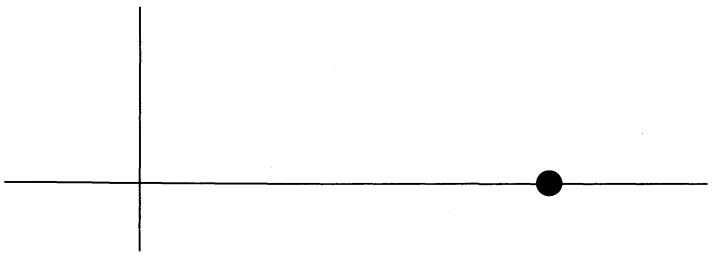


Fig. 4.9. Breakdown: $\hat{h}$ = HLA

down and $\hat{h}$ = HLA + HML for an inspection repair. $f(h)$ is estimated from the synthesis of estimation of $\{\hat{h}\}$. At the time of repair, the maintenance fitter has information available to inform his estimate. In addition to his

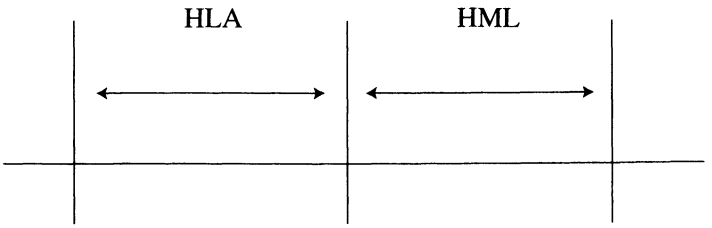


Fig. 4.10. Inspection: $\hat{h}$ = HLA + HML

experience, the defect is present and the plant may be examined and operatives questioned. Even so, in this study there were problems.

Estimating HLA and HML, as well as providing judgements as to causes of defects, was a novel task for the engineers, and the first attempt highlighted misconceptions. Typically the HLA estimate would be very short for inspection repairs, a matter of minutes. The implication of the engineers' estimates is that perhaps one should not inspect, since defects seemingly entered the plant just before an inspection (see Figure 4.11). Once the consequences of their estimates was accepted by the engineers, they also accepted that the HLA measure, as well as the HML, was being underestimated and appropriate 'revisions' were made.

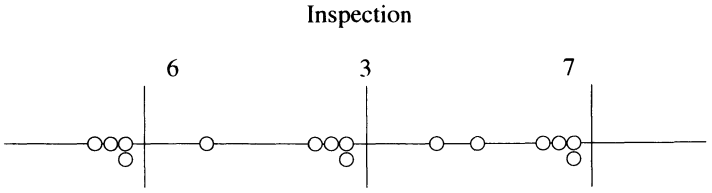


Fig. 4.11. ○, initial points

On occasions, engineers would provide a range estimate of HLA and HML, not a point estimate. This led to an optimistic and pessimistic estimate for $\hat{h}$, and therefore for $f(h)$, and consequently for the estimate of the probability of a defect arising as a breakdown, $b(T)$, (4.1). The observed value of $b(T = 24)$ lay reassuringly between the optimistic and pessimistic estimates of $b(T)$ for $T = 24$ hours. This was welcomed, but not expected since the $b(T)$ curves are a synthesis of subjective assessments, and the $b(T = 24)$ is the observed and objective point of current practice.

The subsequent model for downtime per unit time as a function of inspection period, $D(T)$, corresponding to (4.3), is reproduced in Figure 4.12. Evidently, the 24-hour pit-stops based upon engineers' judgement, with just over 7 hours downtime per week, was the best one could sensibly do with the plant since the marginal improvement in moving to $T = 30$ hours could not justify the effort of an unnatural period. Management had chosen the best inspection period, but without modelling did not know it.

As a consequence of a snap-shot survey, the key plant component where engineering solutions could reduce the rate of defect arrival was the set of

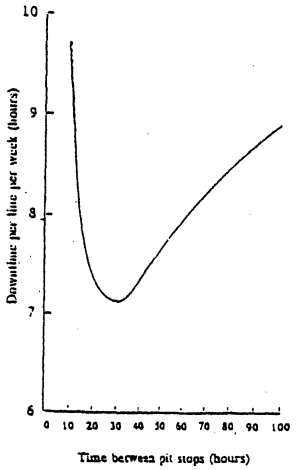


Fig. 4.12. Maintenance model for canning line: original plant

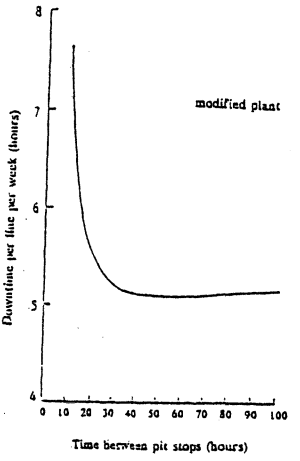


Fig. 4.13. Maintenance model for canning line: modified plant

40 filling valves on each line. It was known that valves were troublesome, but the extent had never been quantified. Over the 6-week survey, 72 valve faults were identified, causing management to task three engineers to propose design improvements to the valves. A redesigned valve with a troublesome flange region removed was developed, tested and ready for use, and the analysts were asked what the inspection period for the redesigned plant should be. It proved possible by removing from the delay time estimates $\{\hat{h}\}$ all estimates associated with defects believed designed out, and adjusting the rate of arrival of defects λ accordingly, to provide within a day a model of $D(t)$ for the new, and as yet untried, plant. The new curve, Figure 4.13, now bottomed out at just over 5 hours per day, but remained very flat between $T = 40$ and $T = 100$ before slowly increasing.

This revised curve was used by management to argue and win the case for growing out to weekly pit-stops. A year later, on a return visit to monitor developments, pit-stops were still weekly, the downtime level was as predicted, and none of the 'designed out' valve faults had occurred. Two years later, a visiting check revealed that pit-stops had now moved to two-weekly, but were of a more substantial nature.

This study shows how the delay time concept can model plant downtime, even untried modifications, and be influential in improving maintenance practice. It is to be noted that without the intervention of OR modelling (i) the appropriateness of a 24-hour pit-stop policy, or its relative gain over any other period, would be unknown, (ii) the initiative to redesign the valve would not have started in the foreseeable future, (iii) savings of 2 hours per week in downtime would not have been realised, (iv) a change in pit-stop period would not have occurred. In this case, the model represented by (4.3) was simple to solve, but graphical techniques are also possible for this and more complex cases [48]. However, perhaps the most enduring feature of this first study is a verified and validated ex-

ample that the delay time concept can be utilised to model and influence actual maintenance practice.

4.12 Revision of Subjectively Estimated Delay Time Distribution

In the canning line problem, the underestimation problem of HLA and HML was resolved by presenting to engineers the implication of their estimates, and inviting a revision. There was no need for further adjustment since the subsequent model of $b(T)$, or $D(T)$, was satisfactorily close to the known point for $T = 24$ hours. In general, a further 'adjustment' will be necessary, since as observed elsewhere, there is no guarantee the subjectively based model will model the status quo [29], [49]. Underestimating delay time measures will imply the known point of Figure 4.3 lies below the model of Figure 4.1, or above if h is overestimated.

The task of revising both the delay time distribution and the delay time maintenance model has been addressed by Christer and Redmond [50], who transform the estimated delay time to $h = \alpha\hat{h} + w$, where w and the stretch factor α are assumed constant. In the case $w = 0$, a unique value of α exists to match a delay time model to the status quo measure irrespective of the quality of inspection r. The model may also be updated by changing the assumed wrong initial value of r, and a unique revised r value exists provided $b^* > b(T_0; r = 1)$, where b^* is the observed status quo value of $b(T_0)$ and T_0, the current inspection period.

Other cases including varying w, α and r together can lead to unique solution, no solution, or multiple solutions to status quo equation such as $b^* = b(T_0)$. Some possible means of resolving the latter case are also discussed in the paper. There are numerous options here, and the actual updating technique selected in any case will probably depend upon the context, and sensitivity analysis of the decision consequence to different updating options. Experience to date indicates decisions to be robust to updating procedures. Even so, there is a case for exploring Bayesian techniques for updating prior distribution and models [51].

4.13 Correction for Sampling Bias

It is possible that estimates of delay time measures of specific defects for use in the synthesis approach to estimating delay time parameters may only be obtainable at either inspections, $\hat{h}_i$ = (HLA + HML) or at breakdown, $\hat{h}_b$ = (HLA) or that the estimator may only be experienced in one of these cases. Both sets of estimate $\{\hat{h}_i\}$ and $\{\hat{h}_b\}$ lead to estimates of density function $F_i(h)$ and $F_b(h)$, but neither is the process delay time distribution $F(h)$. This is because there is a bias towards small delay time in $\{h_b\}$ for breakdown estimates, and larger delay time estimates in $\{\hat{h}_i\}$, since defects with larger delay time are more likely to span an inspection. Christer and Redmond [52] address this sampling bias problem and propose ways of estimating $F(h)$ from either $\{\hat{h}_i\}$ or $\{\hat{h}_b\}$. It may still, however, be necessary to revise the delay time distribution estimate or model to match

the status quo condition. This bias is related to the well-known waiting time paradox [53].

4.14 Subjective Estimation of the Delay Time Distribution Directly

An alternative subjective technique for estimating delay time parameters is to estimate the delay time distribution directly [54]. This is based upon the work of Cooke [55] and in essence entails providing a mesh of class intervals and asking a group of experts to indicate, for 100 random faults arising within the plant, how many delay times would lie within each interval. This provides a histogram from which to estimate $f(h)$. An advantage of the method over the previous one of synthesis is speed, ofen reducing to an afternoon that which may otherwise require a survey spanning weeks or months. The contextual difference is in the nature of the experience and the evidence influencing the final delay time estimate. However, where both methods have been used, and subsequently revised to satisfy status quo conditions, the resulting maintenance models have been very similar in form and in decision consequence [49]. In practice, when a data collection survey is necessary to define the maintenance problem and identify engineering solutions, it is recommended that the delay time distribution be estimated using the synthesis technique in addition to the probability estimating technique. The latter could aid modelling decisions to extend or terminate the longer timescale subjective data collection survey.

4.15 Objective Estimation of Delay Time Parameters

Component Models

Baker and Wang [56] were the first to estimate delay time parameters using objective data. Interest is restricted to a component tracking DT model of independent components subject to inspection, where defects and failures are effectively repaired to replacement level. Using the multiplication law of likelihood, expressions are developed for the likelihood of observing the recorded sequence of events. The likelihood is conditional upon assumed forms of $g(u)$, the pdf for the initial point u, and $f(h)$, for each component. Inspections as such may or may not be perfect, and when the plant consists of more than one component, opportunistic inspection may take place at failure repairs. Parameter fits for $g(u)$, $f(h)$ and the quality of inspection r, are obtained by maximising the likelihood expression. The authors advocate choosing between different possible parametric forms of $g(u)$ and $f(u)$ using the Akaike information criterion, and medical equipment maintenance data is used to demonstrate the technique.

This important paper on component tracking models was developed further by Baker and Wang [32]. Still addressing component modelling, the mathematical form of the likelihood was extended to allow the initial point u and delay time h to depend upon the age of the parent plant,

and to permit inspection to have adverse as well as beneficial effects. An implicit assumption in the likelihood formulation is that the required data exists. This may not be the case, and ways of combining existing objective data with subjective data with subjective data within a likelihood formulation for delay time parameter estimation of component tracking models is discussed in Baker and Christer [57].

Multi-component Plant

Component tracking models for maintenance modelling are appropriate for equipment that has very few inspected or otherwise monitored repairable components. Though components can be combined to form a plant, for complex equipment consisting of many components the implied complexity becomes too great. A preferred approach for modelling complex plant is to directly model the rate of arrival of defects at time u as $\lambda(u)$, with the delay time density function $f(h)$ as before. If sufficient data exist for given forms of $f(h)$ and $\lambda(u)$, the maximum likelihood technique may be used for parameter estimation.

A common set of seven steps is proposed for deriving objective estimates of delay time parameter in the case of both component tracking and multi-component models. These are

(1) Clarify the definition accuracy, completeness and content of available data. This is important for stages (2) or (3).
(2) Formulate and confirm assumptions defining the operating custom and practice over the period for which objective data have been collected.
(3) Identify candidates forms for $\lambda(u)$, $f(h)$ and r, the probability of detection of a fault at inspection.
(4) Formulate the likelihood expression for observed data.
(5) Obtain maximum likelihood estimates of parameter for assumed forms of $\lambda(u)$, $f(h)$ and r.
(6) Use an information criterion to make choice between different forms of $\lambda(u)$, $f(h)$ and r.
(7) Check goodness of fit of choice to observed data.

The mathematical form of a likelihood will, of course, depend upon the data available and the operating practice. Suppose, for example, inspection takes place every k days, say, with identified defects removed, l inspection cycles of data are available, M_n defects are identified at the n-th inspection, $1 \leq n \leq l$, and m_{nj} failure occurred on the j-th day of the n-th cycle, $1 \leq j \leq k$, see Figure 4.14.

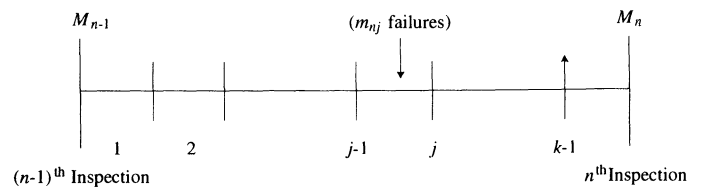


Fig. 4.14. Possible data format

The likelihood expression for this data set is

$$L = \prod_{n=1}^{l} \Big\{ P(M_n \text{ defects identified and removed at } n\text{-th inspection}) \times \prod_{j=1}^{k} P(m_{nj} \text{ failure on } j\text{-th day of } n\text{-th cycle}) \Big\}. \tag{4.12}$$

In formulating such expressions for likelihood, the task is considerably aided by the assumption of a Poisson process for $\lambda(u)$, and by extension to the theory of NHPP applicable to the delay time context.

First, there are intuitive arguments supporting the assumption of a Poisson process of defects arising within complex plant, which are essentially those supporting Poisson arrivals of failure in complex plant maintained on a breakdown basis [26]. For established plant, the probability of a fault arriving in a small time interval will be hardly influenced by the pattern of previous fault arrivals and subsequent failures and repairs (and subsequent change in the risk of future faults arising) since the arrivals will relate to a negligible fraction of the total complex plant. This independence of previous epochs characterises an NHPP, and one might reasonably anticipate fault arrivals for established complex plant to follow an NHPP, or a homogeneous Poisson Process (HPP) over a short time period. Further support is given here by Barlow and Proschan [25] who proved that for a complex plant with negligible repair times and subject to breakdown maintenance, the failure process (and in our cases the corresponding fault arrival process) follows a HPP in the steady state.

An important feature of the failure process we wish to model here is that the plant is not only repaired upon breakdown, but also subject to periodic inspection and repair of identified defects. Also, at breakdowns, on occasion the failed plant can be subject to an opportunistic inspection. Suffice it to say, assuming defects arise as a HPP, the probabilistic results of Ross [58] can be extended in the current context using the following theorems, Christer and Wang [28], Christer *et al.* [58]:

(1) For perfect or imperfect inspections, in the absence of opportunistic inspection for the plant at a breakdown, the rate of arrival of failures (breakdowns) between inspections follows a NHPP.
(2) If perfect or imperfect PM inspections take place, the number of defects identified at PM is Poisson distributed.
(3) When opportunistic inspections take place at a breakdown, then both the above statements still hold provided the Poisson property of independent increments of the failure process may be assumed [58]:

To be more specific, suppose a complex plant new at time $t = 0$ has PM inspections at times T_i, $i = 1, 2, \ldots$. The period between PMs may not be constant. Defects are assumed to arise at constant rate λ, and a defect has probability r of being detected at an inspection. The pdf and cdf of delay time are as before. We have from the above theorems that in this case

(i) the failure rate arrival process follows a NHPP with failure rate function at time t given by

$$\nu(t) = \lambda \Big\{ \sum_{n=1}^{i-1} (1-r)^{i-n} (F(t - T_{n-1}) - F(t - T_n)) + F(t - T_{i-1}) \Big\}, \qquad T_{i-1} < t \leq T_i, \tag{4.13}$$

(ii) the expected number of defects identified at the i-th PM is $E(N_p(T_i))$

$$E(N_p(T_i)) = \lambda \sum_{n=1}^{i} r(1-r)^{i-n} \int_{T_{n-1}}^{T_n} (1 - F(T_i - u))du, \qquad (4.14)$$

and

(iii) the number of defects identified at the PM in the steady state is Poisson distributed with mean $\lim_{i\to\infty} E(N_p(T_i))$

For example, if $(T_{i+1} - T_i) = T$ for all $i \geq 1$, that is inspections are regular, we have the steady state value of $E(N_p(T_i))$ becoming $E(N_p(T))$ where

$$E(N_p(T)) = \left\{ \lambda T - \lambda \sum_{n=1}^{\infty} r(1-r)^{n-1} \int_0^T F(nT - u)du \right\}. \qquad (4.15)$$

This expression is consistent with (4.4) and gives the parameter of the Poisson distribution of defects identified at T. In the case of finite i when condition may not yet be steady, the distribution of the number of defects identified at T_i is Poisson with parameter $E(N_p(T_i))$ given by (4.14).

Expressions (4.3)–(4.5) readily generalise to where defects identified at PM may not always be rectified, but be left to be attended to at a subsequent PM or cause a future failure [60]. The above theoretical results help considerably in formulating likelihood expressions such as (4.12).

The likelihood formulation (4.12) is based upon one format of possible available data. Unfortunately, in some cases data relating to activities undertaken and defects rectified at a PM inspection is not recorded. This considerably complicates the estimation problem because of the correlation between parameters to be estimated. For example, a failure epoch pattern may be attributable to a high λ and r value with defects being filtered out, or to a relatively low λ and r value, with most defects giving rise to failures. This situation arises [59], [60], [62]. One procedure here is to obtain a subjective estimate for either λ or r, and then seek maximum likelihood estimates for the remaining parameters.

We now comment on the practical experience of using the above modelling and parameter estimating methods.

4.16 Case Experience Using Objective Data: HPP of Defect Arrival

The first modelling study undertaken using objective data was the downtime modelling of a 1700 ton extrusion press used in copper products manufacture. It operated 16 hours a day (2 shifts), was a key plant, was 35 years old at the start of the project, and had had a history of frequent breakdowns. Two years previous, a PM system lasting 2 hours per week had been introduced consisting of a thorough plant inspection along with subsequent adjustment or repair of faults found. A full account of this study is given in Christer, Wang, Baker and Sharp [59].

After a time it was agreed with management that the following assumptions describe the press maintenance practice

(1) Faults may be assumed to arise according to a Poisson process of rate λ.

(2) All faults are assumed to be independent of each other and follow the same delay time distribution.
(3) The delay time h of a fault is independent of its time of origin and has pdf $f(h)$ and cdf $F(h)$.
(4) PMs carried out at PMs are assumed to be independent of each other and follow the same delay time distribution.
(5) All identified faults are rectified by repair or replacement during the PM, or as soon as possible after.
(6) Failures are identified immediately and repaired.
(7) PMs are performed on a regular basis of period T.

The objective here is to identify the best operating period T to minimise the downtime per unit time $D(T)$ for the plant, that is minimise the equivalent expression to (4.2).

The data format in this study is that described in Figure 4.14, with corresponding likelihood expression of (4.12). It was evident that some defects will always have a zero delay time, and others may occasionally [59]. Accordingly, a mixed delay time distribution of the form

$$F(h) = 1 - (1 - P)e^{-\alpha h}$$

was adopted, where $P = \Pr(h \equiv 0)$. The outcome of the above methodology for estimating parameters was, in this case,

$\alpha = 0.179$ per day
$P = 0.5546$
$r = 0.9$
$\lambda = 1.356$ per day.

It proved possible to derive a closed form expression for the steady state expected number of failures $E(N_f)$ over a PM period T, namely

$$E(N_f) = \lambda T - \frac{\lambda r(e^{\alpha T} - 1)(1 - p)}{\alpha(e^{\alpha T} - 1 + r)}. \tag{4.16}$$

The second term on the RHS of (4.16) represents the reduction in the number of failures arising over the inspection period attributable to the inspection process. The objective function corresponding to (4.3) can now be formulated.

Initially, PM was weekly and took two hours: $dp = 120$ minutes. Subsequently, maintenance engineers were allowed early morning access to the plant before production started, thereby reducing production time lost to PM to 30 minutes. This was later further reduced to zero when all PM work was scheduled to take place before production commenced. The downtime per unit of time model corresponding to these three cases is presented in Figure 4.15.

Clearly, if $dp = 120$, the best PM period is two or three weekly, but not weekly as practised. Weekly PM is appropriate when $dp = 30$, and obviously PMs are done as often as possible to reduce downtime when $dp = 0$. Changes in practice meant it was possible here to validate modelling prediction using observed values for different PM periods (Table 4.1).

The closeness of the observed and modelled downtime provides reassurance of the validity of the joint problem modelling and parameter estimation process. The greater variance between the observed and the predicted

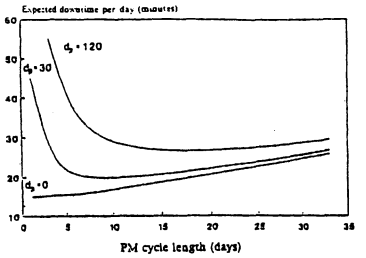


Fig. 4.15. Press downtime per day

Table 4.1. Observed and predicted % downtime per press hour.

% downtime per press hour	Observation	Model Prediction
No PM	5.47	5.53
Weekly PM ($dp = 120$)	4.06	4.05
Daily PM ($dp = 0$)	2.45	1.85

in the case where $dp = 0$ is believed due in part to a relatively short operating period providing the observed measure. There is, however, another explanation which sheds greater light upon the modelling process.

To gain insight into the objectively and subjectively based modelling process, a repeat study was undertaken for this same plant, but using subjective data [61]. Subjectively estimated parameters required revision in this case (as discussed previously) to satisfy status quo conditions, and the subsequent results proved remarkably consistent with those of the objectively based study, and are summarised in part in Figure 4.16.

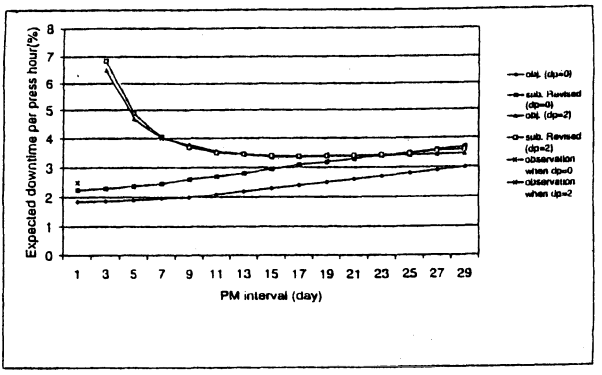


Fig. 4.16. Comparison of objective and subjective model

The closeness of the objective, and subjectively revised models in the case $dp = 2$ hours is evident and adds credibility to both the objective and subjective modelling procedures. More interesting is the difference in the models for $dp = 0$. Where the subjective model is revised to satisfy a status quo point corresponding to $dp = 0$ hours, it appears a better fit than the objective model. Whilst both models are similar in form and decision consequence, the difference in model prediction is of the order of 0.5% in

downtime per press hour. In both objective and subjective cases, the model assumes $dp = 0$ since PM is in downtime per press hour scheduled before the start of production. However, the plant was not always completely available for production at the scheduled time, and had an average of as little as 5 minutes delay to commencing production been assumed, the modelled point would coincide with the observation.

Often, there is a lack of data available on the findings and action at inspection, though breakdowns are recorded. This can lead to convergence problems when optimising a likelihood formulation, which so far have been resolved by blending both objective and subjective data [62]. Other problems caused by small sample sizes have been investigated by [63].

Case Experience Using Objective Data and NHPP of Defect Arrivals

Perhaps the most complex plant yet tackled using the delay time concept is a high-technology steel manufacturing plant which operates 24 hours a day seven days a week with the exception of maintenance downdays (2 weekly) when PM and inspection take place [38]. Between downdays, the plant, consisting of seven rolling mills each with three rollers, operates at such a rate that a hardened steel roller can wear within a period just over three hours and need replacing. A roll change mechanism dedicated to each mill performs this task, which is normally completed within 12 minutes. Sometimes the roll change mechanism does not work, which causes abnormal roll change periods of up to several hours. The modelling task was to study the performance of the roll change mechanism as a function of downday period.

Between downdays, defects arise which may or may not cause an abnormal roll change at some point within the role change mechanism, see Figure 4.17. Defects arise, and may be present during a role change, but if

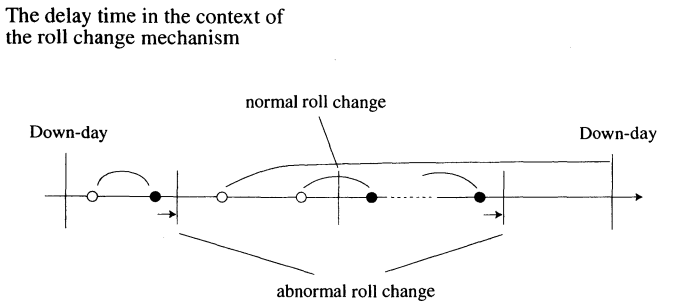


Fig. 4.17. Defect and failure arrival patterns

they have not led to a failure, the roll change will be normal. In this situation, a failure as such is not recognisable until the mechanism is called upon. We are, therefore, dealing with a problem of the preparedness type. Other modelling variations that arose in this case are that failures present at roll changes are not always repaired, but subject to temporary measures to replace rollers and restore production. The same defect will be present at the next, and therefore abnormal, role change. This was modelled by

introducing a probability Q that a failure present at a roll change is fixed. Also, it was evident from data analysis that downday maintenance was not perfect, and could inject faults. An appropriate form of λ was the NHPP expression $\lambda(u) = \lambda_1 + \lambda_2 e^{-\lambda_3 u}$. Data from over 1,000 roll changes led to maximum likelihood estimate of:

$Q = 0.5884$, *i.e.* 40% of failures temporarily fixed
$P = 0.1280$, 13% of defects have zero delay time
$\lambda_1 = 0.2004$, 1 defect arises every 5 hours on average (long term)
$\left.\begin{array}{l}\lambda_2 = 1.470\\ \lambda_3 = 0.065\end{array}\right)$ initially, 1 extra defect arises every 40 minutes
$\alpha = 0.010$, exponential delay time parameter.

The resulting model of downtime per day is shown in Figure 4.18, when $t_a = 19$ minutes is the mean time required for a abnormal repair. Before the study commenced, management had moved from a two- to a three-week PM period, but had done so based upon judgement and not quantitative modelling. This model shows that as far as the roll change mechanism is concerned, the change was appropriate, and that an even larger increase in down-day period may be contemplated. The existence of the model could have considerably reduced the decision stress for management in their decision to extend the PM period. Insufficient information on and understanding of PM activity prevents us from drawing the conclusion that PM is ineffective, though clearly it can be made more effective through quality techniques to prevent the introduction of faults. The gain of modelling in

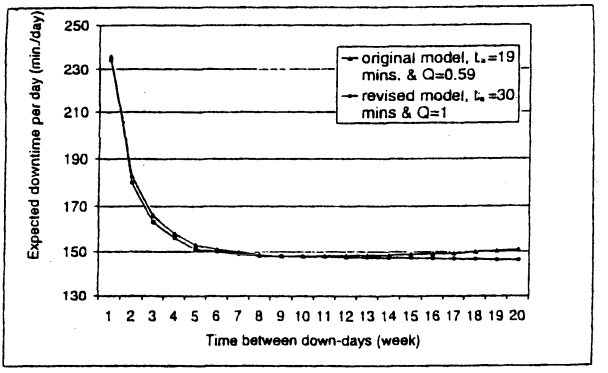


Fig. 4.18. Expected downtime per day for roll change mechanim as function of downday period

this context is the increased insight into the order of magnitude of contributing effects and the consequence of possible changes. For instance, if at a roll change failures were fully fixed, *i.e.* $Q = 1$, then presumably the expected roll change period would increase for the failure case, but decrease for the rest. Using the model, we see that there is virtually no change between performance given current practice and always fully repairing failures at roll changes, if the average time for all roll changes is $t_a = 30$ minutes. This indicates the scope for changing policy and improving the downtime. Other measures, such as improved practice at PM and design changes to reduce $\lambda(u)$, can likewise be examined. As it is, the study supported the

previous decision of management, identified weaknesses in the information system, and both highlighted and quantified areas for detailed engineering study.

4.17 Discussion of Further Developments in Delay Time Modelling

So far, much of the modelling discussed has assumed a uniform inspection/PM period. This is, of course, convenient in practice, but is not necessary and may be relaxed. The modelling task can then become one of inspection and replacement modelling [64]. Reliability as a criterion has also been considered, where interest is focused on the dependence of reliability upon inspection practice for a repairable component [34], [35], reliability over a mission period [36], and safety as measured by an undetected defect [37].

The estimation of parameters when modelling complex plant is considerably aided by the assumption of a Poisson arrival process and the consequential Poisson arrival process for failures. In the case where opportunistic inspections take place, the failure arrival rate is still Poisson provided the property of independent increments is valid [58]. This is a strong assumption and can be non-trivial to establish. Suffice it to say, where attempts have been made to check its validity within a modelling context by simulation, the assumption of independent increments appeared reasonable [28].

Case studies and associated theoretical developments are required in the area of non-uniform inspections, where the depth of inspection as well as the period of inspection are decision variables. Different depths of inspection correspond to a different delay time distribution $F(h)$. Opportunistic inspection can increase the technical complexity of maintenance modelling, but needs to be addressed in a more coherent fashion since it is a frequent occurrence in practice. Though simple to implement in practice, opportunistic inspection can considerably increase the complexity of maintenance modelling [28]. It is convenient to assume steady state conditions in modelling. However, if the time zone of requirement for plant is both finite and known, it could be productive to modify maintenance decisions accordingly, and finite horizon modelling is required. The relationship of equipment inspection modelling to health screening is frequently commented on, but as yet is not fully explored, though a useful start has been made by Baker [65]. Once Bayesian techniques for the revision of models and delay time parameters have been explored and sufficient understanding has been gained from the modelling of case studies, a long-term aspiration would be to couple maintenance information systems to intelligent decision support systems, as postulated by Kobbacy *et al.* [66] and Ascher and Kobbacy [67], to enable delay time analysis and modelling to become one of several modelling techniques whose modelling is almost automatic with a computer system.

Of the other techniques available within the literature for modelling the consequences of maintenance interactions, Markov-based techniques such as that of Winden and Dekker [68] are the most numerous, though the literature is thin if interest is in validated applications, or even applications. The high-speed canning line problem discussed above [46] could also

have been modelled as a Markov process. Had this been done, joint ownership of the model with the engineers would have been more problematic because of the distance between Markov concepts and current engineering practice. Such communication problems within maintenance have long been recognised [2]. Also, there would be little to guide the estimation of Markov transition probabilities for modified plant. However, if $g(u)$ is negative exponential and $f(h)$ arbitrary, a Markov or semi-Markov model may be constructed where the transition probabilities are functions of the delay time parameters [70]. This is to be expected since the DTM approach is of a more fundamental nature and models the process giving rise to state transitions. When $g(u)$ is non-exponential, inspections become imperfect, the period of inspection non-uniform, identified defects are not always repaired, or opportunistic inspections take place at, say, breakdowns, the resulting state space for a semi-Markov approach becomes prohibitively large, and the parameter estimation problem very complex. Another approach, such as DTM, is required.

4.18 Conclusions

The delay time concept is a natural one within the maintenance engineering context. More importantly, it can be used to build qualitative models of the inspection practice of plant, both existing and modified, which have proved in practice to be valid. Techniques exist to estimate delay time parameters given objective data, or subjective data, or a mixture, and where the objective and subjective techniques were jointly tested, have given consistent results. The theory is still developing, but so far there has been no technical barrier to developing a delay-time-based maintenance model for any plant studied. Though only industrial manufacturing has been addressed in the cases cited here, other modelling areas of equal applicability for the delay time approach could be cited. These include vehicle fleets, including lorries and buses [71], [49], transportation systems, roadways and motorways, high-rise housing, concrete structures [50], [72], [73], and housing estates [43].

Close collaboration is required with engineers to produce a snap-shot model, interpret data analysis, and agree a set of defining assumptions for a model. This team approach ensures joint ownership of a modelling exercise, which is essential for implementation. The DTM requirement that a model be capable of modelling the status quo situation and predicting the consequence of changes, which can later be observed and compared with the prediction is important. It increases both the quality of science within modelling and, through this validation, its impact. If a model proves invalid, one must return to earlier stages, perhaps to the initial snap-shot phase, to understand why and revise accordingly. Sometimes the conclusion is that attention to engineering issues and practice is required before maintenance decisions can be usefully modelled [62].

Engineers are generally not used to modelling decisions in management processes, especially maintenance decisions. As such, unless stimulated by some means, maintenance management are unlikely to request such modelling support. The author is aware of remarkably few OR groups who ever consider maintenance as an area of study. It is hoped, therefore, that this paper will highlight the fact that modelling real maintenance problems

is a substantive task where OR/MS support can both benefit the client, produce clearly auditable improvements and challenge the analyst.

The potential for OR/MS to impact upon the area of maintenance management is considerable, but relatively unrecognised within industry. It is believed that in the short term, and despite its importance, OR/MS will continue to have only a minor impact upon maintenance practice. The situation will change when the education of engineers embraces decision modelling concepts for maintenance, and when the OR/MS practitioners within industry start to address maintenance decision. In particular, the area is expected to gain considerably as more OR/MS and engineering academics with an interest in the area of maintenance collaborate to test modelling theories and jointly develop new theories and models in conjunction with case situations.

References

1. Geraerds, W. M. J. (1978), "Estimation of cost of maintenance expenditure within the Netherlands," Internal Report, Faculty of Technology Management, Eindhoven University of Technology, Netherlands
2. Turban, E. (1967), "The use of mathematical models in plant maintenance decision-making," *Management Science*, **13**, B342-358
3. Pierskalla, W. P. and Voelker, J. A. A. (1976), "A survey of maintenance models: the control and surveillance of deteriorating systems," *Naval Research Logistics Quarterly*, **223**, 53-88
4. Christer, A. H. (1984), "Operational Research applied to industrial maintenance and replacement," Development in OR (Eglese. and Rand. eds.). 31-58, Pergamon, ???
5. Thomas, L. C. (1986), "A survey of maintenance and replacement models for maintainability and reliability of multi-item systems," *Reliability Engineering*, **16**, 297-309
6. Valdez-Flores, C. and Feldman, R. M. (1989), "A survey of preventive maintenance models for stochastically deteriorating single unit system," *Naval Research Logistics Quarterly*, **36**, 419-446
7. Cho, D. L. and Parlar, M. (1991), "A survey of maintenance models for multi-unit systems," *European Journal of Operational Research*, **51**, 1-31
8. Pintelon, L. M. and Gelders, L. (1992), "Maintenance management decision-making," *European Journal of Operational Research*, **58**, 301-317.
9. Scarf, P. S. (1997), "On the application of mathematical models to maintenance," *European Journal of Operational Research*, **99**, 493-506
10. Dekker, R. and Scarf, P. A. (1998), "On the impact of optimisation models in maintenance decision-making: the state of the art," *Reliability Engineering and System Safety*, **66**, 111-119
11. Hsu, J. S. (1988), "Equipment replacement policy - a survey," *Production and Inventory Management Journal*, **29**, 23-27
12. Christer, A. H. and Waller, W. M. (1987), "A descriptive model of capital plant replacement," *Journal of the Operational Research Society*, **38**, 473-477
13. Gits, C. W. (1992), "Design of maintenance concepts," *International Journal of Production Economics*, **24**, 217-226
14. Dekker, R., Smit, A. C. and Losekoot, J. M. (1992), "Combining maintenance activities in an operational planning phase: a set partitioning problem," *IMA Journal of Mathematics Applied in Business and Industry*, **3**, 315-331

15. Dekker, R. (1995), "Integrating optimisation, priority setting, planning and combining of maintenance activities," *European Journal of Operational Research*, **82**, 225-240
16. Wildeman, R. E., Dekker, R. and Smit, A. C. J. M. (1997), "A dynamic policy for grouping maintenance activities, " *European Journal of Operational Research*, **99**, 530-551
17. Christer, A. H., MacCallum, K. J., Kobbacy, K. A. H., Bolland, D. and Hesset, C. (1989), "A system model of underwater inspection operation," *Journal of the Operational Research Society*, **40**, 551-565
18. Armitage, W. (1968), "Maintenance Effectiveness," OR in Maintenance, (Jardine, A K S MUP, ed.). 196-223
19. Corder, A. S. (1976), Maintenance Management Techniques. McGraw Hill, New York
20. Nikojima, S. (1989), Total Productive Maintenance Development Programme: Implementing Total Productive Maintenance. Production Press, Cambridge, Massachusetts
21. Mowbray, J. (1997), Reliability Centred Maintenance. Butterworth-Heineman
22. Christer, A. H., Wang, W. and Sharp, J. (1997), "A state space condition monitoring model for erosion furnace prediction and replacement," *European Journal of Operational Research*, **101**, 1-14
23. Aghjagan, H. N. (1989), "Lubeoil analysis expert system," *Proceedings of the Canadian Maint. Eng. Conference*, Toronto
24. Ormerod, R. J. (1993), "The OR/MS contribution to maintenance management: comments on maintenance management decision making," *European Journal of Operational Research*, **65**, 140-142
25. Barlow, R. E. and Proschan, F. (1996), Mathematical Theory of Reliability. SIAM, Philadelphia
26. Ascher, H. E. and Feingold, H. (1984), Repairable Systems Reliability: Modelling Inference, Misconceptions and Their Causes. Lecture Notes in Stasts, Marcel Dekker, New York
27. Christer, A. H. (1976), "Innovatory decision making," *Proceedings of the NATO Conference on Role and Effectiveness of Theories of Decision Theory in Practice* (Bowen, K. C. and White, D. J. eds.), Hodder and Sroughton, 368-377
28. Christer, A. H. and Wang, W. (1995), "A delay-time based maintenance model of a multi-component system," *IMA Journal of Mathematics Applied in Business and Industry*, **6**, 205-222
29. Chilcott, J. and Christer, A. H. (1991), "Modelling of condition based maintenance at the coal face," *International Journal of Production Economics*, **22**, 1-11
30. Christer, A. H. and Lee, C. S. (2000), "Refining the delay-time PM inspection model with non-negligible system downtime estimates of the expected number of failures," Department Report, CORAS, Salford University, to appear in *International Journal of Production Economics, 2000*
31. Cox, D. R. (1957), Renewal Theory. Chapman and Hall, London
32. Baker, R. D. and Wang, W. (1993), "Developing and testing the delay-time model," *Journal of the Operational Research Society*, **44**, 361-374
33. Christer, A. H. and Doherty, T. (1977), "Scheduling overhauls for soaking pits," *Operational Research Quarterly*, **28**, 915-926

34. Christer, A. H. (1987), "Delay-time models of reliability of equipment subject to inspection monitoring," *Journal of the Operational Research Society*, **38**, 329-334
35. Cerone, P. (1991), "On a simplified delay-time model of reliability of equipment subject to inspection monitoring," *Journal of the Operational Research Society*, **42**, 505-511
36. Christer, A. H. and Lee, S. K. (1997), "Modelling ship operational reliability over a mission under regular inspections," *Journal of the Operational Research Society*, **48**, 688-699
37. Wang, W. and Christer, A. H. (1997), "A modelling procedure to optimise component safety over a finite time horizon," *Quality and Reliability Engineering, An International Journal*, **13**, 217-224
38. Christer, A. H., Wang, W., Sharp, J. and Baker, R. D. (1997), "A stochastic modelling problem of high-tech steel production," in Stochastic Modelling in Innovative Manufacturing (Christer, A. H., Osaki, S. and Thomas, L. C. eds.). 196-214, Springer-Verlag, Berlin
39. Christer, A. H. (1992), "Prototype modelling of irregular condition monitoring of production plant," *IMA Journal of Mathematics Applied in Business and Industry*, **3**, 219-232
40. Scarf, P. S. and Christer, A. H. (1997), "Applications of capital replacement models with finite planning horizons," *Internatinal Journal of Technology Management*, **13**, 25-36
41. Christer, A. H. and Waller, W. M. (1987), "Tax adjusted replacement models," *Journal of the Operational Research Society*, **38**, 993-1006
42. Sun Wei, Yang Hong Wei, Qu Fazeng, (1999), "The determination of unequal inspection intervals for repairable mechanical product," *Proceedings of the 4th International Conference Reliability Maintainability and Safety*, (ICRNS '99) Shanghai, China Machine Press, Beijing, 531-535
43. Christer, A. H. (1982), "Modelling inspection policies for building maintenance," *Journal of the Operational Research Society*, **33**, 723-752
44. Christer, A. H. (1988), "Condition based inspection models of major civil engineering structures," *Journal of the Operational Research Society*, **39**, 71-82
45. Redmond, D. and Christer, A. H. (1997), "An application of delay time analysis to concrete structures," *European Journal of Operational Research*, **99**, 619-631
46. Christer, A. H. and Waller, W. M. (1984), "Reducing production downtime using delay-time analysis," *Journal of the Operational Research Society*, **35**, 499-512
47. Christer, A. H. and Whitelaw, J. (1983), "An OR approach to breakdown maintenance: problem recognition," *Journal of the Operational Research Society*, **34**, 1041-1052
48. Pellegrin, C. C. (1991), "A graphical procedure for on-condition maintenance policy: imperfect inspection model and interpretation," *IMA Journal of Mathematics Applied in Business and Industry*, **3**, 177-191
49. Desa, M. I. and Christer, A. H. (1992), "Maintenance and availability modelling of bus transport in Malaysia: Issue and problems," *Proceedings of the International Conference on OR in Developing Countries*, Ahmadabad, India

50. Christer, A. H. and Redmond, D. F. (1992), "Revising models of maintenance and inspection," *International Journal of Production Economics*, **24**, 227-234
51. Singpurwalla, N. P. and Percy, D. F. (1998), "Bayesian calculations in maintenance modelling," Technical report, **03**, Department of Computer & Mathematical Sciences, University of Salford, United Kingdom
52. Christer, A. H. and Redmond, D. F. (1990), "A recent mathematical development in maintenance theory," *IMA Journal of Mathematics Applied in Business and Industry*, **2**, 97-108
53. Feller, W. (1971), An Introduction to Probability Theory and Its Applications, **II**. John Wiley & Sons
54. Wang, W. (1997), "Subjective estimation of the delay-time distribution in maintenance modelling," *European Journal of Operational Research*, **99**, 515-529
55. Cooke, R. M. (1991), "Experts in uncertainty: expert opinion and subjective probability in science," OUP
56. Baker, R. D. and Wang, W. (1992), "Estimating the delay-time distribution of faults in repairable machinery for failure data," *IMA Journal of Mathematics Applied in Business and Industry*, **3**, 259-282
57. Baker, R. D. and Christer, A. H. (1994), "Operational research modelling of engineering aspects of maintenance," *European Journal of Operational Research*, **73**, 407-422
58. Ross, S. M. (1996), Stochastic Process. John Wiley & Sons
59. Christer, A. H., Wang, W., Baker, R. and Sharp, J. (1995), "Modelling maintenance practice of production plant using the delay time concept," *IMA Journal of Mathematics Applied in Business and Industry*, **6**, 67-83
60. Christer, A. H., Wang, W. and Choi, K. M. (1998), "The delay-time modelling of preventive maintenance of plant given limited PM data and selective repair at PM," *IMA Journal of Mathematics Applied in Business and Industry*, **9**, 4
61. Christer, A. H., Wang, W., Sharp, J. and Baker, R. (1998), "A case study of modelling preventive maintenance of production plant using subjective data," *Journal of the Operational Research Society*, **49**, 210-219
62. Christer, A. H., Lee, C. S. and Wang, W. (2000), "A delay-time parameter estimation problem and case study in PM modelling," to appear in *International Journal of Production Economics*
63. Baker, R. D. and Scarf, P. A. (1995), "Can models fitted to small data samples lead to maintenance policies with near optimum cost?," *IMA Journal of Mathematics Applied in Business and Industry*, **6**, 3-12
64. Christer, A. H. (1992), "Prototype modelling of irregular condition monitoring of production plant," *IMA Journal of Mathematics Applied in Business and Industry*, **3**, 219-232
65. Baker, R. D. (1998), "What can industrial inspection models learn from medical analogues? Determining the optimal screening policy for breast cancer," *IMA Journal of Mathematics Applied in Business and Industry*, **9**, 4
66. Kobbacy, K. A. H., Proudlove, N. C. and Harper, M. (1995), "An intelligent maintenance optimisation system," *Journal of the Operational Research Society*, **46**, 831-853

67. Ascher, H. E. and Kobbacy, K. A. H. (1995), "Modelling preventive maintenance for deteriorating repairable systems," *IMA Journal of Mathematics Applied in Business and Industry*, **6**, 85-99
68. Winder, C. van and Dekker, R. (1998), "Rationalisation of building maintenance by Markov decision models: a pilot case study," *Journal of the Operational Research Society*, **49**, 928-935
69. Christer, A. H. and Waller, W. M. (1984), "Delay time models of industrial inspection maintenance," *Journal of the Operational Research Society*, **35**, 401-406
70. Christer, A. H. and Wang, W., Choi, K. M. and Van der Duyn Schouten. (1998), "The robustness of the semi-Markov and delay time maintenance models to the Markov assumption," to appear in *European Journal of Operational Research*
71. Christer, A. H. and Waller, W. M. (1984), "An OR approach to Planned Maintenance Modelling (PM) for a vehicle fleet," *Journal of the Operational Research Society*, **35**, 967-984
72. Redmond, D.F., Christer, A. H., Rigden, S. R. and Burley, E. (1997), "OR modelling of the deterioration and maintenance of concrete structures," *European Journal of Operational Research*, **99**, 619-631
73. Burley, E., Christer, A. H. and Rigden, S. D. (1989), "Inspection practice for concrete structure in the U.K.," *Proceedings of the Conference British Cement Association, Life of Concrete Structures*
74. Wang, W. (1999), "A model of multiple nested inspections of production plant," to appear in *Computer and Operations Research*

5. Imperfect Preventive Maintenance Models

Toshio Nakagawa
Department of Industrial Engineering,
Aichi Institute of Technology,
Toyota 470-0392, Japan

Summary.

Two imperfect preventive maintenance (pm) models where (i) the age of the unit becomes x units of time younger at pm and (ii) the age t or the failure rate $r(t)$ reduces to at or $ar(t)$ at pm have been well-known. This chapter applies the notion (ii) of imperfectness to a sequential policy where the pm is done at sequential intervals. The expected costs of two models and optimal intervals are analytically derived. Further, we also apply this notion to a cumulative damage model where the total damage Y_k reduces to $a_k Y_k$ at the k-th pm. The expected cost is obtained, and optimal policies which minimize it are discussed. To make it possible to understand these results easily and correctly, some numerical examples of each model are given.

Keywords: preventive maintenance, imperfect pm, sequential pm, shock model, age of younger, decrease of damage

5.1 Introduction

The maintenance of an operating unit after failure is costly, and sometimes it requires a long time to repair failed units. It would be an important problem to determine when to maintain the unit preventively before it fails. However, it would not be wise to maintain the unit too often. From this point of view, commonly considered maintenance policies are replacement policies for the unit with no repair and preventive maintenance policies for the unit with repair. It may be wise to maintain the unit to prevent failure when the failure rate increases with age.

The usual preventive maintenance (pm) of the unit is done before failure at a specified time T after its installation. The mean time to failure (MTTF), the availability and the expected cost are derived as the measures of reliability for maintained units. Optimal pm policies which maximize or minimize these measures have been summarized in [1], [2], [3]. All models have assumed that "after pm the unit is as good as new." Actually, this assumption might not be true. The unit after pm usually might be younger at pm, and occasionally, it might be worse than before pm because of faulty procedures, *e.g.*,

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

wrong adjustments, bad parts, and damage done during pm. Generally, the improvement of the unit by pm would depend on the resources spent for pm.

Weiss [4] first assumed that the inspection to detect failures may not be perfect. Similar models, in that inspections, test and detection of failures are uncertain, were treated in Coleman and Abrams [5] and Noonan and Fain [6]. Chan and Downs [7], Nakagawa [8], [9] and Murthy and Nguyen [10] considered the imperfect pm where after pm the unit is not as new with a certain probability, and discussed the optimal policies which maximize the availability or minimize the expected cost.

It is imperative to check a computer system and remove as many faults, failures and degradations as possible by providing fault tolerant techniques. Imperfect maintenances for a computer system were first treated by Ingle and Siewioreck [11]. Helvic [12], Yak *et al.* [13] and Nakagawa and Yasui [14] considered that while the system is usually renewed after pm, it sometimes remains unchanged, and obtained the MTTF and the availability. Chung [15] studied the imperfect test of intermittent faults which had occurred in digital systems.

Nakagawa [16], [17] considered two imperfect pm models of the unit:
(i) The age becomes x units of time younger at each pm and (ii) the failure rate is reduced in proportion to that before pm or the pm cost. Lie and Chun [18] and Jayabalan and Chaudhuri [19] introduced an improvement factor in failure rate or age after maintenance, and Canfield [20] considered the system degradation with time where the pm restores the hazard function to the same shape.

Brown and Proschan [21], Fontenot and Proschan [22], and Bhattacharjee [23] assumed that a failed unit is as good as new with a certain probability and investigated some properties of a failure distribution. Similar imperfect repair models were studied by Ebrahimi [24], Natvig [25], Makis and Jardine [26], and Zhao [27]. Further, Shaked and Shanthikumar [28], and Sheu *et al.* [29], [30] derived multivariate distributions and studied probabilistic quantities of imperfect repair models. Recently, Wang and Pham [31], [32] considered the extended pm models with imperfect repair and discussed the optimal policies which minimize the expected cost and maximize the availability.

This chapter summarizes the results of sequential imperfect pm models which could be applied to actual systems and would be helpful for future studies in research fields.

Section 5.2 considers a sequential imperfect pm model where the pm is done at successive times and the age or the failure rate reduces in proportional to those before pm. The expected cost rates are obtained and optimal policies which minimize them are discussed. It is shown in numerical examples that optimal intervals are uniquely determined when the failure time has a Weibull distribution.

Section 5.3 applies a sequential pm policy to a cumulative damage model.

5.2 Sequential Imperfect Preventive Maintenance

5.2.1 Introduction

A sequential policy where preventive maintenance (pm) is done at fixed intervals x_k was proposed in [33], [34]. This policy could be applied to real systems,

because most systems need more frequent maintenances with age. In one approximation model of imperfect pm, the system has different failure rates in the period k of pm, while they increase with the number of pm's [35].

It is reasonable to postulate the system where pm reduces the failure rate or the age. The improvement by pm depends on cost of pm and/or age of the system.

By introducing improvement factors [17], [18] in failure rate and age for a sequential pm policy [33], [34], we consider the following two pm policies: The pm is done at fixed intervals x_k $(k = 1, 2, \cdots, N-1)$ and the system is replaced at the N-th pm. If the system fails between pm's it undergoes only minimal repair. The pm is imperfect as follows:

(i) The age after the k-th pm falls to $a_k t$ when it was t before pm.
(ii) The failure rate in the k-th pm becomes $b_k r(t)$ when it was $r(t)$ in the previous period of pm.

The expected cost rates of two models are obtained and optimal sequences $\{x_k^*\}$ are derived. When the failure time has a Weibull distribution, optimal intervals are computed explicitly.

5.2.2 Model A - age

Consider the sequential pm policy for a one-unit system which has to operate for an infinite span. It is assumed that [36]:

1. The pm is done at fixed intervals x_k $(k = 1, 2, \cdots, N-1)$ and the system is replaced at the N-th pm, *i.e.*, the unit is maintained preventively at successive times $x_1 < x_1 + x_2 < \cdots < x_1 + x_2 + \cdots + x_{N-1}$ and is replaced at time $x_1 + x_2 + \cdots + x_N$ where $x_0 \equiv 0$.
2. The unit undergoes only minimal repair at failures between replacements and is as good as new at replacement.
3. The age after the k-th pm reduces to $a_k t$ when it was t before pm, *i.e.*, the unit of age t becomes $t(1 - a_k)$ units of time younger at the k-th pm, where $0 = a_0 < a_1 \leq a_2 \leq \cdots \leq a_N < 1$.
4. The cost of each minimal repair is c_1, the cost of each pm is c_2, and the cost of replacement at the N-th pm is c_3.
5. The times for pm, repair and replacement are negligible.

The unit is aged from $a_{k-1}(x_{k-1} + a_{k-2}x_{k-2} + \cdots + a_{k-2}a_{k-3}\cdots a_2 a_1 x_1)$ after the $k-1$-pm to $x_k + a_{k-1}(x_{k-1} + a_{k-2}x_{k-2} + \cdots + a_{k-2}a_{k-3}\cdots a_2 a_1 x_1)$ before the k-th pm, *i.e.*, from $a_{k-1}Y_{k-1}$ to Y_k, where $Y_k \equiv x_k + a_{k-1}x_{k-1} + \cdots + a_{k-1}a_{k-2} + \cdots + a_2 a_1 x_1$ $(k = 1, 2, \cdots)$, which is the age of the unit immediately before the k-th pm. Thus, the expected cost rate is

$$C_A(Y_1, Y_2, \cdots, Y_N) \equiv \frac{c_1 \sum_{k=1}^{N} \int_{a_{k-1}Y_{k-1}}^{Y_k} r(t)dt + (N-1)c_2 + c_3}{\sum_{k=1}^{N-1}(1 - a_k)Y_k + Y_N},$$

$$N = 1, 2, \cdots, \qquad (5.1)$$

since $x_k = Y_k - a_{k-1}Y_{k-1}$ and $\sum_{k=1}^{N} x_k = \sum_{k=1}^{N-1}(1 - a_k)Y_k + Y_N$.

To find an optimal sequence $\{Y_k^*\}$ which minimizes C_A, differentiating

$C_A(Y_1, Y_2, \cdots, Y_N)$ with respect to Y_k and setting it equal to zero, we have

$$\frac{r(Y_k) - a_k r(a_k Y_k)}{1 - a_k} = r(Y_N), \qquad k = 1, 2, \cdots, N-1, \tag{5.2}$$

$$c_1 r(Y_N) = C_A(Y_1, Y_2, \cdots, Y_N). \tag{5.3}$$

Suppose that Y_N $(0 < Y_N < \infty)$ is fixed. If $r(t)$ is strictly increasing then there exists some Y_k $(0 < Y_k < Y_N)$ which satisfies (5.2), since

$$\frac{r(0) - a_k r(0)}{1 - a_k} < r(Y_N),$$

$$\frac{r(Y_N) - a_k r(a_k Y_N)}{1 - a_k} > r(Y_N).$$

Further, if $r'(t)$ is also strictly increasing then a solution to (5.2) is unique.

Thus, substituting each Y_k into (5.3), its equation becomes a function only of Y_N which is

$$r(Y_N)\left[\sum_{k=1}^{N-1}(1 - a_k)Y_k + Y_N\right] - \sum_{k=1}^{N}\int_{a_{k-1}Y_{k-1}}^{Y_k} r(t)dt = \frac{(N-1)c_2 + c_3}{c_1}, \tag{5.4}$$

where each Y_k $(k = 1, 2, \cdots, N-1)$ is given by some function of Y_N. If there exists a solution Y_N to (5.4), then a sequence $\{Y_k\}$ minimizes the expected cost $C_A(Y_1, Y_2, \cdots, Y_N)$.

Finally, suppose that $Y_1, Y_2, \cdots, Y_N$ are determined from (5.2) and (5.4). Then, from (5.3), the resulting cost is $c_1 r(Y_N)$ which is a function of N. To complete an optimal pm schedule, we may seek an optimal number N^* which minimizes $r(Y_N)$.

From the above discussions, we can specify the computing procedure for obtaining the optimal pm schedule:

1. Solve (5.2) and express Y_k $(k = 1, 2, \cdots, N-1)$ by a function of Y_N.
2. Substitute Y_k into (5.4) and solve it with respect to Y_N.
3. Determine N^* which minimizes $r(Y_N)$.
4. Compute x_k $(k = 1, 2, \cdots, N^*)$ from $x_k = Y_k - a_{k-1}Y_{k-1}$.

5.2.3 Model B - failure rate

3. The failure rate in the k-th pm becomes $b_k r(t)$ when it was $r(t)$ in the $(k-1)$-th pm, *i.e.*, the unit has the failure rate $B_k r(t)$ in the k-th pm period, where $1 = b_0 < b_1 \leq b_2 \leq \cdots \leq b_{N-1}$, $B_k \equiv \prod_{j=0}^{k-1} b_j$ $(k = 1, 2, \cdots, N)$ and $1 = B_1 < B_2 < \cdots < B_N$.

1,2,4,5. Same as the assumptions of Model A.

The expected cost rate until replacement is

$$C_B(x_1, x_2, \cdots, x_N) \equiv \frac{c_1 \sum_{k=1}^{N} B_k \int_0^{x_k} r(t)dt + (N-1)c_2 + c_3}{x_1 + x_2 + \cdots + x_N}, \qquad N = 1, 2, \cdots. \tag{5.5}$$

Differentiating $C_B(x_1, x_2, \cdots, x_N)$ with respect to x_k and setting it equal to zero, we have

$$B_1 r(x_1) = B_2 r(x_2) = \cdots = B_N r(x_N), \tag{5.6}$$

$$c_1 B_k r(x_k) = C_B(x_1, x_2, \cdots, x_N), \qquad k = 1, 2, \cdots, N. \tag{5.7}$$

When the failure rate is strictly increasing to infinity, we can specify the computing procedure for obtaining an optimal schedule [36]:

1. Solve $B_k r(x_k) = D$ and express x_k $(k = 1, 2, \cdots, N)$ by a function of D.
2. Substituting x_k into (5.7) and solve it with respect to D.
3. Determine N^* which minimizes D.

Next, if the unit has the different failure rate $r_k(t)$ in the k-th pm, the expected cost rate is [35]

$$\widetilde{C}_B(x_1, x_2, \cdots, x_N) \equiv \frac{c_1 \sum_{k=1}^{N} \int_0^{x_k} r_k(t)dt + (N-1)c_2 + c_3}{x_1 + x_2 + \cdots + x_N}, \qquad N = 1, 2, \cdots. \tag{5.8}$$

Equations (5.6) and (5.7) are rewritten as follows:

$$r_1(x_1) = r_2(x_2) = \cdots = r_N(x_N), \tag{5.9}$$

$$c_1 r_k(x_k) = \widetilde{C}_B(x_1, x_2, \cdots, x_N), \qquad k = 1, 2, \cdots, N. \tag{5.10}$$

Similar discussions to the above are possible.

5.2.4 Numerical examples

Suppose that the failure time of the unit has a Weibull distribution, *i.e.*, $r(t) = \alpha t^{\alpha-1}$ for $\alpha > 1$.

From the computing procedure of Model A, by solving (5.2), we have

$$Y_k = \left[\frac{1 - a_k}{1 - a_k^{\alpha}}\right]^{1/(\alpha-1)} Y_N, \qquad k = 1, 2, \cdots, N-1. \tag{5.11}$$

Substituting Y_k into (5.4) and arranging it,

$${Y_N}^{\alpha} = \frac{(N-1)c_2 + c_3}{(\alpha - 1)c_1 \sum_{k=0}^{N-1} d_k}, \tag{5.12}$$

where

$$d_k \equiv (1 - a_k)\left[\frac{1 - a_k}{1 - a_k^{\alpha}}\right]^{1/(\alpha-1)}, \qquad k = 0, 1, 2, \cdots, N-1.$$

Next, we consider the problem which minimizes

$$C_A(N) \equiv \frac{(N-1)c_2 + c_3}{\sum_{k=0}^{N-1} d_k}, \qquad N = 1, 2, \cdots, \tag{5.13}$$

and which is the same problem as minimizing $r(Y_N)$, *i.e.*, $C_A(Y_1, Y_2, \cdots, Y_N)$. From the inequality $C_A(N+1) \geq C_A(N)$, we have

$$L_A(N) \geq \frac{c_3}{c_2}, \qquad N = 1, 2, \cdots, \tag{5.14}$$

where

$$L_A(N) \equiv \sum_{k=0}^{N-1} \frac{d_k}{d_N} - (N-1), \qquad N = 1, 2, \cdots. \tag{5.15}$$

If d_k is decreasing in k then $L_A(N)$ is increasing in N. Thus, there exists a finite and unique minimum N^* which satisfies (5.14) if $L_A(\infty) > c_3/c_2$.

Show that d_k is decreasing in k from the assumption that $a_k < a_{k+1}$. Let $g(x) \equiv (1-x)^\alpha/(1-x^\alpha)$ $(0 < x < 1)$ for $\alpha > 1$. Then, $g(x)$ is decreasing from 1 to 0, and hence,

$$\frac{(1-a_k)^\alpha}{1-a_k^\alpha} > \frac{(1-a_{k+1})^\alpha}{1-a_{k+1}^\alpha},$$

which follows that $d_k > d_{k+1}$. Further, if $a_k \to 1$ as $k \to \infty$ then

$$\lim_{k\to\infty} d_k = \lim_{x\to 1} [g(x)]^{1/(\alpha-1)} = 0,$$

i.e., $L_A(N) \to \infty$ as $N \to \infty$, and a finite N^* exists uniquely.

Therefore, if $a_k \to 1$ as $k \to \infty$ then an N^* is a finite and unique minimum which satisfies (5.14), and the optimal intervals are $x_k = Y_k - a_{k-1}Y_{k-1}$ $(k = 1, 2, \cdots, N^*)$, where Y_k and Y_N are given in (5.11) and (5.12), respectively.

For Model B, by solving $B_k r(x_k) = D$, we have

$$x_k = \left[\frac{D}{\alpha B_k}\right]^{1/(\alpha-1)}, \qquad k = 1, 2, \cdots, N. \tag{5.16}$$

Substituting x_k into (5.7) and arranging it,

$$D^{\alpha/(\alpha-1)} = \frac{(N-1)c_2 + c_3}{c_1(1-1/\alpha)\sum_{k=1}^{N} [1/(\alpha B_k)]^{1/(\alpha-1)}}, \tag{5.17}$$

which is a function of N. Let denote D by $D(N)$. Then, from the inequality $D(N+1) \geq D(N)$, an N^* to minimize D is given by a unique minimum which satisfies

$$L_B(N) \geq \frac{c_3}{c_2}, \tag{5.18}$$

where

$$L_B(N) \equiv \sum_{k=1}^{N} \left(\frac{B_{N+1}}{B_k} \right)^{1/(\alpha-1)} - (N-1), \qquad N = 1, 2, \cdots,$$

which is increasing in N since B_k is increasing in k. Further, if $B_k \to \infty$ as $k \to \infty$ then $L_B(N) \to \infty$ as $N \to \infty$, and hence, a finite N^* exists uniquely in (5.18), and the optimal intervals are given in (5.16) and (5.17).

Tables 5.1 and 5.2 give the optimal number N^* and the pm intervals x_1, x_2, $\cdots$, x_{N^*} for $c_3/c_2 = 2, 5, 10, 20, 40$ where $c_1/c_2 = 3$, $\alpha = 2$, and $\alpha_k = k/(k+1)$, $b_k = 1 + k/(k+1)$ $(k = 0, 1, 2, \cdots)$. These examples indicate that $x_1 > x_2 > \cdots > x_{N^*}$ for Model B, but $x_1 > x_{N^*} > x_2$ for $c_3/c_2 = 10, 20, 40$ of Model A. This indicates that it would be reasonable to do frequent pm with age, but it would be better to do the last pm as late as possible because the unit should be replaced at the next pm.

Table 5.1. Optimal N^* and pm intervals of Model A when $c_1/c_2 = 3$

	c_3/c_2				
	2	5	10	20	40
N^*	1	2	4	7	11
x_1	0.54	0.82	1.07	1.40	1.84
x_2		0.82	0.43	0.56	0.74
x_3			0.28	0.36	0.48
x_4			0.92	0.27	0.35
x_5				0.21	0.28
x_6				0.18	0.23
x_7				1.13	0.20
x_8					0.17
x_9					0,15
x_{10}					0.14
x_{11}					1.45

5.3 Shock Model with Imperfect Preventive Maintenance

5.3.1 Introduction

A sequential pm policy where the pm is done at fixed intervals $x_k (k = 1, 2, \cdots, N)$ has been proposed in [33], [34]. This could be useful for real systems because we might need more frequent maintenances with age. In many practical situations, however, the pm seems imperfect only in the sense that it does not make a system like new. Some types of imperfect pm were considered

Table 5.2. Optimal N^* and pm intervals of Model B when $c_1/c_2 = 3$

	c_3/c_2				
	2	5	10	20	40
N^*	2	3	4	5	6
x_1	0.77	1.06	1.37	1.82	2.45
x_2	0.52	0.71	0.92	1.21	1.64
x_3		0.43	0.55	0.73	0.98
x_4			0.31	0.42	0.56
x_5				0.23	0.31
x_6					0.17

in [18], [20]. In this chapter, we apply a sequential pm policy to a shock model (cumulative damage model) where each pm is imperfect.

A system is subject to shocks which occur randomly in time, and upon occurrence of shocks it suffers random damage which is additive. Each shock causes a system failure with probability $p(z)$ when the total damage is z. If the system fails between pm's, it undergoes only minimal repair [1]. We introduce an improvement factor in damage in order to describe imperfect pm actions: The amount of damage after the k-th pm becomes $a_k Y_k$ when it was Y_k before pm, *i.e.*, the k-th pm reduces the amount Y_k of damage to $a_k Y_k$. This is an extension of shock models [37], [38], [39], [40], [41], and would be applied to related reliability models [42], [43].

In this chapter, we describe the model under consideration and obtain the expected cost rate when shocks occur at a Poisson process and $p(z)$ is exponential [44]. Further, we discuss three types of optimal policies which minimize the expected cost rate, when the pm is done at periodic times and an improvement factor is constant, *i.e.*, $x_k = x$ and $a_k = a$. Optimal number $N^*(x)$, optimal interval $x^*(N)$, and optimal pair (N^*, x^*) are derived. Numerical examples are given to demonstrate potential usefulness of this study. Further discussions of optimal policies are referred to [44].

5.3.2 Model and expected cost

Consider a sequential pm policy for the system where the pm is done at fixed intervals $x_k (k = 1, 2, \cdots, N)$ where $x_0 \equiv 0$. We call an interval from the $(k-1)$-th pm to the k-th pm *period* k.

Suppose that shocks occur at a Poisson process with rate λ. Random variables $X_k (k = 1, 2, \cdots, N)$ denote the number of shocks in period k, *i.e.*, $Pr\{X_k = j\} = [(\lambda x_k)^j / j!] \exp(-\lambda x_k) (j = 0, 1, 2, \cdots)$. Further, we denote by Z_{kj} the amount of damage caused by the j-th shock in period k. It is assumed that Z_{kj} are non-negative, independent and identically distributed, and have an identical distribution $G(z)$ for all k and j. The damage is additive, and $G^{(j)}(z)$ is the j-fold Stieltjes convolution of $G(z)$ with itself $(j = 1, 2, \cdots)$, where $G^{(0)}(z) \equiv 1$ for $z \geq 0$. Then, it follows that

$$Pr\{Z_{k1} + Z_{k2} + \cdots + Z_{kj} \leq z\} = G^{(j)}(z), \qquad j = 0, 1, 2, \cdots. \tag{5.19}$$

The system fails with probability $p(z)$ when the total amount of damage becomes z at each shock. If the system fails between pm's, it undergoes only

minimal repair, and hence, the amount of damage remains unchanged by minimal repair.

Next, introduce an improvement factor in pm: Suppose that the k-th pm reduces $100(1-a_k)\%$ of the total damage. Letting Y_k be the total amount of damage in the end of period k, *i.e.*, just before the k-th pm, the k-th pm reduces it to a_kY_k. During period k, the total damage is additive and is not removed since a failed system undergoes only minimal repair. Thus, since the amount $\sum_{j=1}^{X_k} Z_{kj}$ of damage is incurred during period k, we evidently have

$$Y_k = a_{k-1}Y_{k-1} + \sum_{j=1}^{X_k} Z_{kj}, \qquad k = 1, 2, \cdots, N, \tag{5.20}$$

where $Y_0 \equiv 0$ and $\sum_{j=1}^{0} \equiv 0$.

Let c_1 be the cost of minimal repair, c_2 be the cost of each pm and c_3 be the cost of replacement at the N-th pm ($c_3 > c_2$). Then, since the system fails with probability $p(\cdot)$ at each shock, the total expected cost in period k is

$$\hat{C}(k) = c_2 + c_1 \sum_{j=1}^{X_k} p(a_{k-1}Y_{k-1} + Z_{k1} + Z_{k2} + \cdots + Z_{kj}),$$
$$k = 1, 2, \cdots, N-1. \tag{5.21}$$

Similarly, the total expected cost in period N is

$$\hat{C}(N) = c_3 + c_1 \sum_{j=1}^{X_N} p(a_{N-1}Y_{N-1} + Z_{N1} + Z_{N2} + \cdots + Z_{Nj}). \tag{5.22}$$

To obtain the expectations of (5.21) and (5.22), we assume that probability $p(z)$ is exponential, *i.e.*, $p(z) = 1 - e^{-sz}$ for some constant $s > 0$. Letting $g(s)$ be the Laplace-Stieltjes transform of $G(z)$, we have

$$E\{\exp[-s(Z_{k1} + Z_{k2} + \cdots + Z_{kj})]\} = \int_0^\infty e^{-sz} dG^{(j)}(z) = [g(s)]^j. \tag{5.23}$$

The probability that the system fails at the first shock is

$$\int_0^\infty p(z)dG(z) = \int_0^\infty (1 - e^{-sz})dG(z) = 1 - g(s). \tag{5.24}$$

Using the law of total probability in (5.21), the expected cost in period k is

$$E\{\hat{C}(k)\} = c_2 + c_1 E\left\{\sum_{j=1}^{X_k} p(a_{k-1}Y_{k-1} + Z_{k1} + Z_{k2} + \cdots + Z_{kj})\right\}$$
$$= c_2 + c_1 \sum_{n=1}^{\infty} Pr\{X_k = n\} \sum_{j=1}^{n} E\{B(k)\}, \tag{5.25}$$

where

$$B(k) \equiv 1 - \exp[-s(a_{k-1}Y_{k-1} + Z_{k1} + Z_{k2} + \cdots + Z_{kj})].$$

Let $\beta_k(s) \equiv E\{\exp(-sY_k)\}$. Then, since Y_{k-1} and Z_{kj} are independent of each other, we have, from (5.23),

$$E\{B(k)\} = 1 - \beta_{k-1}(sa_{k-1})[g(s)]^j.$$

Thus, from the assumption that X_k has a Poisson distribution,

$$\begin{aligned} E\{\hat{C}(k)\} &= c_2 + c_1 \sum_{n=1}^{\infty} \frac{(\lambda x_k)^n}{n!} \exp(-\lambda x_k) \sum_{j=1}^{n} \{1 - \beta_{k-1}(sa_{k-1})[g(s)]^j\} \\ &= c_2 + c_1 \left[\lambda x_k - \frac{g(s)}{1 - g(s)} \beta_{k-1}(sa_{k-1}) \{1 - \exp[-\lambda x_k(1 - g(s))]\} \right], \\ &\qquad k = 1, 2, \cdots, N-1. \end{aligned} \tag{5.26}$$

Similarly, the expected cost in period N is

$$E\{\hat{C}(N)\} = c_3 + c_1 \left[\lambda x_N - \frac{g(s)}{1 - g(s)} \beta_{N-1}(sa_{N-1}) \{1 - \exp[-\lambda x_N(1 - g(s))]\} \right]. \tag{5.27}$$

It remains to determine $\beta_{k-1}(sa_{k-1})$. Let ${A_j}^k \equiv \prod_{i=j}^{k} a_i$ for $j \leq k$, and $\equiv 1$ for $j > k$. Then, from (5.20),

$$a_{k-1}Y_{k-1} = a_{k-1}a_{k-2}Y_{k-2} + a_{k-1} \sum_{i=1}^{X_{k-1}} Z_{(k-1)i} = \sum_{j=1}^{k-1} \left({A_j}^{k-1} \sum_{i=1}^{X_j} Z_{ji} \right),$$

and hence,

$$\begin{aligned} \beta_{k-1}(sa_{k-1}) &= E\{\exp(-sa_{k-1}Y_{k-1})\} \\ &= E\left\{ \exp\left[-s \sum_{j=1}^{k-1} \left({A_j}^{k-1} \sum_{i=1}^{X_j} Z_{ji} \right) \right] \right\}. \end{aligned}$$

Recalling that Z_{ji} is independent and has a distribution $G(z)$, we have

$$\begin{aligned} &E\left\{ \exp\left(-s{A_j}^{k-1} \sum_{i=1}^{X_j} Z_{ji} \right) \right\} \\ &\quad = \sum_{n=0}^{\infty} Pr\{X_j = n\} E\left\{ \exp\left(-s{A_j}^{k-1} \sum_{i=1}^{n} Z_{ji} \right) \right\} \\ &\quad = \sum_{n=0}^{\infty} \frac{(\lambda x_j)^n}{n!} \exp(-\lambda x_j)[g(s{A_j}^{k-1})]^n \\ &\quad = \exp\{-\lambda x_j[1 - g(s{A_j}^{k-1})]\}, \end{aligned}$$

and consequently,

$$\beta_{k-1}(sa_{k-1}) = \exp\left\{-\sum_{j=1}^{k-1}\lambda x_j[1-g(sA_j{}^{k-1})]\right\}. \tag{5.28}$$

Substituting (5.28) into (5.26) and (5.27), respectively, the expected costs in period k are

$$E\{\hat{C}(k)\} = c_2 + c_1\left[\lambda x_k - \frac{g(s)}{1-g(s)}\exp\left\{-\sum_{j=1}^{k-1}\lambda x_j[1-g(sA_j{}^{k-1})]\right\}\right.$$
$$\left.\times\{1-\exp[-\lambda x_k(1-g(s))]\}\right], \quad k = 1, 2, \cdots, N-1, \tag{5.29}$$

and

$$E\{\hat{C}(N)\} = c_3 + c_1\left[\lambda x_N - \frac{g(s)}{1-g(s)}\exp\left\{-\sum_{j=1}^{N-1}\lambda x_j[1-g(sA_j{}^{N-1})]\right\}\right.$$
$$\left.\times\{1-\exp[-\lambda x_N(1-g(s))]\}\right]. \tag{5.30}$$

Therefore, the expected cost rate until replacement is

$$C(x_1, x_2, \cdots, x_N) \equiv \frac{\sum_{k=1}^{N-1} E\{\hat{C}(k)\} + E\{\hat{C}(N)\}}{\sum_{k=1}^{N} x_k}$$
$$= \frac{\left(\begin{array}{l}(N-1)c_2 + c_3 + c_1\left[\sum_{k=1}^{N-1}\lambda x_k - \frac{g(s)}{1-g(s)}\right.\\ \times\sum_{k=1}^{N}\exp\left\{-\sum_{j=1}^{k-1}\lambda x_j[1-g(sA_j{}^{N-1})]\right\}\\ \left.\times\{1-\exp[-\lambda x_k(1-g(s))]\}\right]\end{array}\right)}{\sum_{k=1}^{N} x_k}. \tag{5.31}$$

5.3.3 Optimal policies

The expected cost rate in (5.31) is very complicated, and we cannot analyze optimal policies. Suppose that $x_k = x$ and $a_k = a$, *i.e.*, the pm is done at period times kx ($k = 1, 2, \cdots, N$) and an improvement factor a_k is constant. Then, the expected cost rate is rewritten as

$$C(N,x) = \lambda c_1 + \frac{(N-1)c_2 + c_3 - c_1 \frac{g(s)}{1-g(s)} Q_N(x)}{Nx}, \tag{5.32}$$

where

$$Q_N(x) \equiv [1 - e^{-\lambda x[1-g(s)]}] \sum_{k=1}^{N} \exp(-\lambda x \xi_k), \qquad N = 1, 2, \cdots,$$

and

$$\xi_1 \equiv 0,$$
$$\xi_k \equiv \sum_{j=1}^{k-1} [1 - g(sa^j)], \qquad k = 2, 3, \cdots.$$

When $a = 0$, *i.e.*, the pm is perfect, $\xi_k = 0$ and the expected cost rate is

$$C(\infty, x) = \lambda c_1 + \frac{c_2 - c_1 \frac{g(s)}{1-g(s)} [1 - e^{-\lambda x[1-g(s)]}]}{x}. \tag{5.33}$$

Since $c_3 > c_2$, $C(N,x)$ in (5.32) is decreasing in N, and hence, $N^* = \infty$. Thus, the optimal interval x^* is easily derived by differentiating (5.33) and setting it equal to zero.

Before discussing optimal policies, we define a function which plays an important role in the following parts. Let

$$T_k(x) \equiv c(k) - c_1 \frac{g(s)}{1-g(s)} [1 - e^{-\lambda x[1-g(s)]}] \exp(-\lambda x \xi_k), \quad k = 1, 2, \cdots, \tag{5.34}$$

where $c(1) = c_3$ and $c(k) = c_2$ $(k = 2, 3, \cdots, N)$. Then, (5.32) can be simplified as

$$C(N,x) = \lambda c_1 + \frac{1}{Nx} \sum_{k=1}^{N} T_k(x). \tag{5.35}$$

(1) Optimal number $N^*(x)$

We seek an optimal number $N^*(x)$ which minimizes $C(N,x)$ in (5.35) for $0 < a < 1$ when x is fixed. From the inequality $C(N+1,x) \geq C(N,x)$, we have

$$L(N|x) \geq \frac{c_3 - c_2}{c_3 - T_1(x)}, \qquad N = 1, 2, \cdots, \tag{5.36}$$

where $L(0|x) \equiv 0$ and

$$L(N|x) \equiv \sum_{k=1}^{N} \exp(-\lambda x \xi_k) - N \exp(-\lambda x \xi_{N+1}), \qquad N = 1, 2, \cdots. \tag{5.37}$$

Evidently,

$$L(N|x) - L(N-1|x) = N[\exp(-\lambda x \xi_N) - \exp(-\lambda x \xi_{N+1})] > 0,$$

since ξ_N is strictly increasing in N. Hence, $L(N|x)$ is increasing in N. Thus, if $L(\infty|x) \equiv \lim_{N\to\infty} L(N|x) > (c_3 - c_2)/[c_3 - T_1(x)]$ then there exists a finite and unique minimum $N^*(x)$ which satisfies (5.36).

Example 5.3.1 Suppose that the amount of damage at each shock has an exponential distribution $G(z) = 1 - e^{-\mu z}$ and $g(s) = \mu/(s+\mu)$. Then,

$$\xi_k = \sum_{j=1}^{k-1} [a^j/(a^j + \mu/s)], \qquad k = 2, 3, \cdots.$$

Assume that the amount of damage after pm is reduced in proportion to pm cost c_2, *i.e.*, $c_2/c_3 = 1 - a$. Table 5.3 gives the optimal number $N^*(x)$ and the resulting cost rates $C(N^*, x)/(\lambda c_1)$ for $a = 0.9 \sim 0.1$ and $c_3/c_1 = 3, 5, 10$ when $\lambda x = 7$ and $g(s) = 0.9$, *i.e.*, $\mu/s = 9$. This indicates that $N^*(x)$ are not increasing with respect to a against our expectation. But, this can be explained because $L(N|x)$ depends on a through c_2/c_3. For example, suppose that $x = 7$ days, *i.e.*, the pm is scheduled on the weekend and shocks occur on the average once a day. Further, if $a = 0.5$ and $c_3/c_1 = 5$, *i.e.*, the pm cost is half of the replacement cost and the damage is reduced to half by pm, then the system should be replaced at three weeks. When a is small, several $N^*(x)$ become infinite. These cases indicate that the damage is removed greatly by pm and the system should undergo only pm rather than replacement.

Table 5.3 Optimal $N^*(x)$ and expected cost rates $C(N^*, x)/(\lambda c_1)$ when $g(s) = 0.9, \lambda x = 7$ and $c_2/c_3 = 1 - a$

	c_3/c_1					
	3		5		10	
a	$N^*(x)$	$C(N^*,x)/(\lambda c_1)$	$N^*(x)$	$C(N^*,x)/(\lambda c_1)$	$N^*(x)$	$C(N^*,x)/(\lambda c_1)$
0.9	2	0.7408	3	0.8917	7	1.1203
0.8	2	0.7508	3	0.9192	6	1.2084
0.7	2	0.7597	3	0.9443	6	1.2869
0.6	2	0.7674	3	0.9671	9	1.3569
0.5	2	0.7739	3	0.9876	∞^*	1.4086
0.4	2	0.7790	3	1.0062	∞	1.4656
0.3	1	0.7813	3	1.0229	∞	1.5324
0.2	1	0.7813	∞^*	1.0367	∞	1.6081
0.1	1	0.7813	∞	1.0487	∞	1.6915

∞^* indicates that $N^*(x)$ may not be infinite, but is very large.

(2) Optimal number $x^*(N)$

We seek an optimal interval $x^*(N)$ which minimizes $C(N, x)$ in (5.35) when N is fixed. Differentiating $C(N, x)$ with respect to x and setting it equal to zero, we have

$$x \sum_{k=1}^{N} T_k{}'(x) = \sum_{k=1}^{N} T_k(x), \tag{5.38}$$

i.e.,

$$\sum_{k=1}^{N}\left[1+\lambda x\xi_k\{1+\lambda x[1-g(x)+\xi_k]\}e^{-\lambda x[1-g(s)]}\right]\exp(-\lambda x\xi_k) = \frac{(N-1)c_2+c_3}{c_1}\frac{1-g(s)}{g(s)}. \tag{5.39}$$

Note that the term with $k=1$ in the left-hand side of (5.39) is a gamma distribution of order 2 so that it increases from 0 to 1. Other terms with k ($k=2,3,\cdots,N$) are unimodal, each having the mode which is a unique solution of

$$e^{-\lambda x(1-g(s))} = \left(\frac{\xi_k}{1-g(s)+\xi_k}\right)^2. \tag{5.40}$$

Thus, the left-hand side in (5.39) is increasing from 0 first, then oscillating and finally decreasing to converge to 1. Therefore, there are possibly at most $(2N-1)$ solutions which satisfy (5.39). An optimal $x^*(N)$ is either one of solutions or $x^*(N)=\infty$. If there is no solution then $x^*(N)=\infty$. In particular, when $N=1$, there exists a unique solution if $g(s)/[1-g(s)] > c_3/c_1$.

Example 5.3.2 We compute $x^*(N)$ for $N=1,2,\cdots,7$ when $g(s)=0.9$, $c_3/c_1=5$ and $a=c_2/c_3=0.5$. Table 5.4 indicates the values of $x^*(N)$ and $C(N,x^*)/(\lambda c_1)$ when N varies. In this case, the optimal interval becomes infinity for $N\geq 7$.

Table 5.4 Optimal $x^*(N)$ and expected cost rates $C(N^*,x^*)/(\lambda c_1)$ when $g(s)=0.9, c_3/c_1=5$ and $a=c_2/c_3=0.5$

N	$x^*(N)$	$C(N,x^*)/(\lambda c_1)$
1	18.627	0.860271
2	13.358	0.909511
3	11.665	0.942916
4	10.816	0.965437
5	10.293	0.981115
6	9.933	0.992396
7	∞	1

(3) Optimal pair (N^*,x^*)

We seek both optimal x^* and N^* together. From (5.32), we can see that $C(N,\infty)=\lambda c_1$ for any $N\geq 1$. Thus, optimal (N^*,x^*) must satisfy $C(N^*,x^*)\leq \lambda c_1$. It also follows from (5.32) that a necessary condition for (N^*,x^*) is that $T_k(x^*)<0$ at least one $k\leq N^*$, since otherwise no contribution to the second term in (5.32) occurs.

Next, consider the inequality $T_k(x)\leq 0$. This is equivalent to

$$h_k(x)\geq \frac{c(k)}{c_1}\frac{g(s)}{1-g(s)}, \tag{5.41}$$

where

$$h_k(x) \equiv [1 - e^{-\lambda x[1-g(s)]}] \exp(-\lambda x \xi_k).$$

It is easy to see that $h_k'(x) = 0$ $(k = 2, 3, \cdots, N)$ has a unique solution m_k which satisfies

$$[1 - g(s) + \xi_k] e^{-\lambda x[1-g(s)]} = \xi_k. \tag{5.42}$$

Thus, $h_k(x)$ is unimodal with mode m_k, and hence, we have

$$h_k(x) \leq h_k(m_k) \left(1 - \frac{\xi_k}{1 - g(s) + \xi_k}\right) \left(\frac{\xi_k}{1 - g(s) + \xi_k}\right)^{\xi_k/[1-g(s)]} < 1. \tag{5.43}$$

It is proved that both m_k and $h_k(m_k)$ are decreasing in k so that m_∞ and $h_\infty(m_\infty)$ exist. It follows that

$$N^* < k^* = \min \left\{ k \geq 2 : h_k(m_k) \leq \frac{c_2}{c_1} \frac{1 - g(s)}{g(s)} \right\}. \tag{5.44}$$

Here, if $h_\infty(m_\infty) > (c_2/c_1)[1 - g(s)]/g(s)$, then we set $k^* = \infty$. Since m_k is decreasing, $x^* \geq m_{k^*-1}$. On the other hand, $x^* \leq max\{x^*(1), m_2\}$. To this end, suppose that x satisfies (5.36). Recall that $T_k'(x) < 0$ for $x < m_k$, $T_k' \geq 0$ for $x \geq m_k$, and m_k is decreasing in k. Hence, if $x^*(1) > m_2$ then either $x^* = x^*(1)$ with $N^* = 1$ or $x^* < x^*(1)$. If $x^*(1) < m_2$, then $x^* \geq m_2$ never happens since $\sum_{k=1}^{N} T_k'(x^*)/N > T_1'(x^*(1))$. Thus, $x^* \leq max\{x^*(1), m_2\}$, as desired.

Therefore, we have the following optimal policy: Suppose that $k^* < \infty$. Then, optimal pair (N^*, x^*) is confined as $N^* < k^*$ and $m_{k^*-1} \leq x^* \leq max\{x^*(1), m_2\}$, where k^* is given in (5.44) and m_k is a unique solution of (5.42). Therefore, the optimal pair is given by

$$x^*(N^*) = \inf_{1 \leq N < k^*} x^*(N) = \inf_{m_{k^*-1} \leq x \leq max\{x^*(1), m_2\}} N^*(x). \tag{5.45}$$

Example 5.3.3 Consider the model in Example 5.3.1 and determine an optimal pair (N^*, x^*) which minimizes $C(N, x)$. In this example, $h_4(m_4) \approx 0.2621 < 0.27$ so that $N^* \leq 3$. In fact, from Table 5.4, it is evident that $N^* = 1$ and $x^* = 18.627$.

5.4 Conclusions

Most pm's are imperfect, *i.e.*, units are not as new, but are younger at least at each pm. The improvement of units by pm depends on the resources spent for pm, *i.e.*, pm costs, pm times, repairmen's abilities, factory equipments and so on. But, it would be very difficult to estimate improvement factors and

reliability measures connected with them. It seems that there is no practical paper that has dealt with imperfect pm. I fully expect that imperfect pm models will be applied to real systems and some reports on these subjects will be made and published in the near future.

Recently, Pham and Wang [31], [32] have given good summaries of these models over the past 30 years. Such papers and this chapter are very helpful and useful for future studies. I believe that there are still many theoretical problems in imperfectness, and hope that many researchers will study these interesting subjects.

References

1. Barlow, R. E. and Proschan, F. (1965), Mathematical Theory of Reliability. John Wiley & Sons, New York
2. Nakagawa, T. (1977), "Optimum preventive maintenance policies for repairable systems," *IEEE Transactions on Reliability*, **R-26**, 166-173
3. Valdez-Flores, C. and Feldman, R. M. (1989), "A survey of preventive maintenance models for stochastically deteriorating single-unit systems," *Naval Research Logistics Quarterly*, **36**, 419-446
4. Weiss, G. H. (1962), "A problem in equipment maintenance," *Management Science*, **8**, 266-277
5. Coleman, J. J. and Abrams, I. J. (1962), "Mathematical model for operational readiness," *Operations Research*, **10**, 126-133
6. Noonan, G. C. and Fain, C. G. (1962), "Optimum preventive maintenance policies when immediate detection of failure is uncertain," *Operations Research*, **10**, 407-410
7. Chan, P. K. W. and Downs, T. (1978), "Two criteria for preventive maintenance," *IEEE Transactions on Reliability*, **R-27**, 272-273
8. Nakagawa, T. (1979), "Optimal policies when preventive maintenance is imperfect," *IEEE Transactions on Reliability*, **R-28**, 331-332
9. Nakagawa, T. (1979), "Imperfect preventive-maintenance," *IEEE Transactions on Reliability*, **R-28**, 402
10. Murthy, D. N. P. and Nguyen, D. G. (1981), "Optimal age-policy with imperfect preventive maintenance," *IEEE Transactions on Reliability*, **R-30**, 80-81
11. Ingle, A. D. and Siewiorek, D. P. (1977), "Reliability models for multiprocessor systems with and without periodic maintenance," *Proceedings of the 7th International Symposium Fault-Tolerant Computing*, 3-9
12. Helvic, B. E. (1980), "Periodic maintenance on the effect of imperfectness," *Proceedings of the 10th International Symposium Fault-Tolerant Computing*, 204-206
13. Yak, Y. W., Dillon, T. S. and Forward, K. E. (1984), "The effect of imperfect periodic maintenance of fault tolerant computer system," *Proceedings 14th International Symposium Fault-Tolerant Computing*, 66-70
14. Nakagawa, T. and Yasui, K. (1987), "Optimum policies for a system with imperfect maintenance," *IEEE Transactions on Reliability*, **R-36**, 631-633
15. Chung, K. J. (1995), "Optimal test-times for intermittent faults," *IEEE Transactions on Reliability*, **44**, 645-647
16. Nakagawa, T. (1980), "Mean time to failure with preventive maintenance," *IEEE Transactions on Reliability*, **R-29**, 341
17. Nakagawa, T. (1980), "A summary of imperfect preventive maintenance policies with minimal repair," *R.A.I.R.O. Operations Research*, **14**, 249-255

18. Lie, C. H. and Chun, Y. H. (1986), "An algorithm for preventive maintenance policy," *IEEE Transactions on Reliability*, **R-35**, 71-75
19. Jayabalan, V. and Chaudhuri, D. (1992), "Cost optimization of maintenance scheduling for a system with assured reliability," *IEEE Transactions on Reliability*, **R-41**, 21-25
20. Canfield, R. V. (1986), "Cost optimization of periodic preventive maintenance," *IEEE Transactions on Reliability*, **R-35**, 78-81
21. Brown, M. and Proschan, F. (1983), "Imperfect repair," *Journal of Applied Probability*, **20**, 851-859
22. Fontenot, R. A. and Proschan, F. (1984), "Some imperfect maintenance models," in Reliability Theory and Models. Academic Press, Orlando, Florida
23. Bhattacharjee, M. C. (1987), "New results for the Brown-Proschan model of imperfect repair," *Journal of Statistical Planning and Inference,* **16**, 305-316
24. Ebrahimi, N. (1985), "Mean time to achieve a failure-free requirement with imperfect repair," *IEEE Transactions on Reliability*, **R-34**, 34-37
25. Natvig, B. (1990), "On information based minimal repair and the reduction in remaining system lifetime due to the failure of a specific module," *Journal of Applied Probability*, **27**, 365-375
26. Makis, V. and Jardine, A. K. S. (1992), "Optimal replacement policy for a general model with imperfect repair," *Journal of Operational Research Society*, **43**, 111-120
27. Zhao, M. (1994), "Availability for repairable components and series systems," *IEEE Transactions on Reliability*, **43**, 329-334
28. Shaked, M. and Shanthikumar, J. G. (1986), "Multivariate imperfect repair," *Operations Research*, **34**, 437-448
29. Sheu, S. H. and Griffith, W. S. (1991), "Multivariate age-dependent imperfect repair," *Naval Research Logistics* , **38**, 839-850
30. Sheu, S. H. and Griffith, W. S. (1992), "Multivariate imperfect repair," *Journal of Applied Probability*, **29**, 947-956
31. Pham, H. and Wang, H. (1996), "Imperfect maintenance," *European Journal of Operational Research*, **94**, 425-438
32. Wang, H. and Pham, H. (1996), "Optimal age-dependent preventive maintenance policies with imperfect maintenance," *International Journal of Reliability, Quality and Safety Engineering*, **3**, 119-135
33. Nakagawa, T. (1986), "Periodic and sequential preventive maintenance policies," *Journal of Applied Probability*, **23**, 536-542
34. Nguyen, D. G. and Murthy, D. N. P. (1981), "Optimal preventive maintenance policies for repairable systems," *Operations Research*, **29**, 1181-1194
35. Nakagawa, T. (1989), "A summary of replacement models with changing failure distributions," *R.A.I.R.O. Operations Research*, **23**, 343-353
36. Nakagawa, T. (1988), "Sequential imperfect preventive maintenance policies," *IEEE Transactions on Reliability*, **37**, 295-298
37. Chikte, S. D. and Deshmukh, S. D. (1981), "Preventive maintenance and replacement under additive damage," *Naval Research Logistics Quarterly*, **28**, 33-46
38. Feldman, R. M. (1976), "Optimal replacement with semi-Markov shock models," *Journal of Applied Probability*, **13**, 108-117
39. Nakagawa, T. (1976), "On a replacement problem of a cumulative damage model," *Operational Research Quarterly*, **27**, 895-900
40. Taylor, H. M. (1975), "Optimal replacement under additive damage and other failure models," *Naval Research Logistics Quarterly*, **22**, 1-18

41. Zuckerman, D. (1977), "Replacement models under additive damage," *Naval Research Logistics Quarterly*, **24**, 549-558
42. Kijima, M. (1989), "Some results for repairable systems with general repair," *Journal of Applied Probability*, **26**, 89-102
43. Kijima, M., Morimura, H. and Suzuki, Y. (1988), "Periodic replacement problem without assuming minimal repair," *European Journal of Operational Research*, **37**, 194-203
44. Kijima, M. and Nakagawa, T. (1992), "Replacement policies of a shock model with imperfect preventive maintenance," *European Journal of Operational Research*, **57**, 100-110

6. Generalized Renewal Processes and General Repair Models

Masaaki Kijima
Faculty of Economics,
Tokyo Metroporitan University
Tokyo 192-0397, Japan

Summary.

This chapter considers a generalized renewal process (g-renewal process for short) and its applications to reliability theory. The g-renewal process was first developed by Kijima and Sumita [21] and used by Kijima, Morimura and Suzuki [18] to construct the failure process of a repairable system with general repair. Since then, many authors have proposed various general repair models and investigated the associated failure processes. See, for example, Kijima [13], Baxter, Kijima and Tortorella [3], Last and Szekli [23], [24] and references therein. Stadje and Zuckerman [30] and Makis and Jardine [26] considered optimal maintenance strategies, while Dagpunar [6] developed some computational methodology for a general repair functional. See also Guo and Love [8] and Dorado, Hollander and Sethuraman [7] for statistical issues on this subject.

In this chapter, the intention is to outline the connection between the theory of g-renewal processes and general repair models. A general repair model with full generality is constructed using a general point process, as demonstrated by Last and Szekli [23], [24]. Here, we do not intend to cover such results. See Kijima, Li and Shaked [17] for a survey of general repair models.

Keywords: Markov process, renewal process, point process, virtual age, effective age, stochastic ordering, hazard rate, g-renewal function

6.1 Background and Motivation

In reliability models, we often encounter the situation that the underlying stochastic process is described by a Markov partial-sum process. To be more specific, let X_n, $n = 1, 2, \cdots$, be a sequence of random variables and define

$$S_n = \sum_{i=1}^{n} X_i, \qquad n = 0, 1, 2, \cdots, \tag{6.1}$$

where $S_0 = 0$. The process $\{S_n, n = 0, 1, 2, \cdots\}$ (or $\{S_n\}$ for short) is called a *partial-sum process* with increments X_n; and if the probability distribution of

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

S_{n+1} depends only on the value of S_n, then it is called Markovian. For example, suppose that the increments X_n are independent, identically distributed (IID) random variables. Then, since

$$S_{n+1} = S_n + X_{n+1}, \qquad n = 0, 1, 2, \cdots,$$

the partial-sum process $\{S_n\}$ is a Markov process in discrete time on the real line which is temporally and spatially homogeneous.

Throughout this chapter, in order to state the motivation of our study from the context of reliability theory, we consider the following simple *basic model* (BM).

(BM) Consider a system which is maintained by repair. A new system is put in operation at time 0 and, when it fails, a repair activity takes place immediately, which is executed by a negligible time. We denote the working time between the n-th and $(n+1)$-th failures by X_{n+1}, where the 0-th failure is assumed to occur at time 0. The working-time distribution of the new system is denoted by $F(t)$, $t \geq 0$, *i.e.* $F(t)$ is the distribution function of X_1. Distribution functions of X_n other than X_1 may differ according to repair activities. In any case, however, S_n describes the total working time of the system up to the n-th failure, or the elapsed time until the n-th failure since installation of the new system.

In this section, we provide several examples of reliability models in which Markovian partial-sum processes appear naturally.

First, suppose that the increments X_n are non-negative and IID. Then, the partial-sum process $\{S_n\}$ is a Markov process on the non-negative real line. In this case, we are interested in the associated *counting process* $\{N(t),\ t \geq 0\}$ (or $\{N(t)\}$ for short) defined by

$$N(t) = \sup\{n : S_n \leq t\}, \qquad t \geq 0. \tag{6.2}$$

Since X_n are non-negative, $N(t)$ is non-decreasing in t and jumps up by a positive integer almost certainly. The random quantity $N(t)$ registers the number of failed systems up to time t in the basic model, and the counting process $\{N(t)\}$ is called a *renewal process*. Classical renewal theory has been studied extensively in the applied probability literature and appears in any standard textbook such as Çinlar [5] or Ross [29].

Example 6.1.1 Consider a system subject to failure and in which failures, if any, are fixed by repair which is executed by a negligible time. Suppose that the repair activity is *perfect* in the sense that it provides a functioning system which is as good as new. This means that the working times X_n are IID and we are in the situation that classical renewal theory is applicable. In practice, the perfect repair assumption may be reasonable for systems with one unit which is structurally simple.

In the classical renewal model, the increments X_n are IID regardless of the state of the system. In certain situations, however, it would be more realistic to assume that they are not identical nor independent. For example, if the working times become shorter as the number of failures increases, then a repair model in which the increments X_n are not identically distributed may be more plausible. See Kijima [12] for such a repair model. On the other hand, if the increment X_{n+1} is considered to depend on the total working time or the level of cumulative damage up to the n-th failure, then the partial-sum process $\{S_n\}$ becomes Markovian with temporal homogeneity but not with spatial

homogeneity. In the following sections, we study such a generalized renewal model with some details.

Example 6.1.2 Consider a system subject to failure, but in this example, repair activities are assumed to be *minimal*. That is, repair restores the system to its functioning condition just prior to failure. More formally, suppose that $F(t)$ is absolutely continuous with density function $f(t)$, and let

$$r(t) = \frac{f(t)}{\overline{F}(t)}, \qquad t \geq 0, \tag{6.3}$$

where $\overline{F}(t) = 1 - F(t)$ is the survival distribution of X_1. The function $r(t)$ is called the *hazard rate function* of $F(t)$. Under the minimal repair assumption, the hazard rate remains undisturbed by repair. See, *e.g.*, Barlow and Proschan [1]. Now suppose that $S_n = y$. Then, the increment $X_{n+1} = S_{n+1} - S_n$ is distributed by

$$P\{X_{n+1} \leq x | S_n = y\} = \frac{F(x+y) - F(y)}{1 - F(y)}, \qquad x \geq 0. \tag{6.4}$$

Hence, the partial-sum process $\{S_n\}$ is a Markov process in discrete time on the non-negative real line with temporal homogeneity but not with spatial homogeneity. The associated counting process $\{N(t)\}$ is a *non-homogeneous* Poisson process (NHPP) with intensity function $r(t)$. Especially, if the working-time distribution $F(t)$ is exponential then, from (6.4), its memoryless property leads to

$$P\{X_{n+1} \leq x | S_n = y\} = F(x), \qquad x \geq 0,$$

and the associated counting process becomes an ordinary Poisson process. The minimal repair assumption is plausible for systems consisting of many components each having its own failure mode.

Example 6.1.3 In many practical instances, repair activities may not result in the extreme situations considered in the above two examples, but in a complicated intermediate one. Let V_n denote the *virtual age* of the system immediately after the n-th repair, and suppose that, given $V_n = y$, the functioning system obtained has the $(n+1)$-th working time X_{n+1} distributed by

$$P\{X_{n+1} \leq x | V_n = y\} = \frac{F(x+y) - F(y)}{1 - F(y)}, \qquad x \geq 0. \tag{6.5}$$

Recall that $F(x)$ is the working-time distribution of a new system, and a new system is assumed to have the virtual age $V_0 = 0$. Note the difference between (6.4) and (6.5). Let A_n, $n = 1, 2, \cdots$, be the *degree* of the n-th repair. Kijima [13] considered, as his Model I, a *general repair model* in which

$$V_{n+1} = V_n + A_{n+1} X_{n+1}, \qquad n = 0, 1, 2, \cdots.$$

That is, the n-th repair does not remove the damage incurred before the $(n-1)$-th repair, but reduces the additional age X_n to $A_n X_n$. Typically, $0 \leq A_n \leq 1$, but the case $A_n > 1$, *i.e.* a *clumsy* repair, may exist in practice. Note that if $A_n = 0$ then one has a perfect repair, while $A_n = 1$ means a minimal repair.

Suppose that $A_n = a$ for all n as considered in Kijima, Morimura and Suzuki [18]. Then, $V_n = aS_n$ and so, from (6.5), we obtain

$$P\{X_{n+1} \leq x | S_n = y\} = \frac{F(x+ay) - F(ay)}{1 - F(ay)}, \qquad x \geq 0. \tag{6.6}$$

Hence, in this case, the partial-sum process $\{S_n\}$ is a Markov process with temporal homogeneity. Note that if A_n are random variables, then $\{S_n\}$ may not be Markovian, although the virtual age process $\{V_n\}$ is *always* so.

Example 6.1.4 Baxter, Kijima and Tortorella [3] considered a general repair model in which, given the *effective age* $Y = y$, the n-th working time X_n is distributed by

$$P\{X_n \leq x | Y = y\} = \frac{F(x+y) - F(y)}{1 - F(y)}, \qquad x \geq 0, \tag{6.7}$$

and the effective age Y may depend on the partial sum S_n and others. The effective age is interpreted as the difference between the degree of effort required to perform a perfect repair and the degree of effort actually expended. Now, let $G_y(x)$ be the distribution function of the effective age Y when $S_n = y$. Then, from (6.7), the working-time distribution is given by

$$P\{X_{n+1} \leq x | S_n = y\} = \int_0^\infty \frac{F(x+u) - F(u)}{1 - F(u)} dG_y(u), \qquad x \geq 0. \tag{6.8}$$

Thus, the partial-sum process $\{S_n\}$ is a Markov process with temporal homogeneity. It should be noted that if, given $S_n = y$, the effective age is almost certainly $Y = ay$, *i.e.*

$$G_y(u) = \begin{cases} 1, & u \geq ay, \\ 0, & u < ay, \end{cases} \tag{6.9}$$

then (6.8) reduces to (6.6).

In some situations, however, increments need not be positive, as the next example shows. A general renewal theory that permits negative increments has been developed by Keener [11], where the partial-sum process $\{S_n\}$ is Markovian on the real line with temporal homogeneity but not with spatial homogeneity. Note that, because of the two-sided nature of the increments, the real time t cannot be explicitly incorporated and the meaning of the counting process defined by (6.2) is lost in this model.

Example 6.1.5 This example considers a periodic preventive maintenance (PM) model in which a certain amount of the damage level is reduced by maintenance activities. The PM is called *imperfect* if it does not make a system as new but younger. Many types of imperfect PM have been considered in the literature, and the reader is referred to, *e.g.*, Pham and Wang [28] and Nakagawa [27]. Suppose that a new system is put in operation at time 0 and the PM is scheduled at periodic times $n = 1, 2, \cdots$. Let S_n denote the cumulative damage at the end of the n-th period and let D_n be the amount of damage incurred during the n-th period. We assume that each PM reduces $100b\%$, $0 \leq b \leq 1$, of the total damage. Such a periodic imperfect PM model is considered in Kijima and Nakagawa [19], [20]. Formally, the PM mechanism is described by

$$S_{n+1} = (1-b)S_n + D_{n+1}, \qquad n = 0, 1, 2, \cdots,$$

where $S_0 = 0$. If the damage D_n depends only on the level of the cumulative damage S_n, then the process $\{S_n\}$ is Markovian with temporal homogeneity but not with spatial homogeneity. To see this, suppose $S_n = y$. Then, the increment $X_{n+1} = S_{n+1} - S_n$ is distributed by

$$P\{X_{n+1} \leq x | S_n = y\} = P\{D_{n+1} \leq x + by | S_n = y\}.$$

Note that, in this model, the increment X_{n+1} may be negative. Now, suppose that the system fails when the accumulated damage first exceeds a level ξ. Assuming the system failure is detected only by PM, the distribution of the time to failure, T say, is obtained through the equations

$$P\{T > n\} = P\{S_n \leq \xi\}, \qquad n = 1, 2, \cdots.$$

The random variable T is called the *first passage time* into the upper set (ξ, ∞).

6.2 Generalized Renewal Processes

Let X_n, $n = 1, 2, \cdots$, be a sequence of *positive* random variables, and define the partial-sum process $\{S_n\}$ by (6.1). Associated with $\{S_n\}$ is the counting process $\{N(t)\}$ defined by (6.2). In this section, we assume that, as in Kijima and Sumita [21], the distribution of X_{n+1} is dependent on the value of S_n only, *i.e.*

$$P\{X_{n+1} \leq x | S_n = y\} = F(x|y), \qquad x \geq 0, \tag{6.10}$$

where $\{F(x|y),\ y \geq 0\}$ denotes the family of distribution functions indexed by y. We note that the probability in (6.10) is independent of n. Hence $\{S_n\}$ is a Markov process in discrete time defined on the non-negative real line, which is temporally homogeneous but not spatially homogeneous. Also, since S_n is almost certainly non-decreasing in n, the process $\{S_n\}$ is a *pure jump process*. Many authors have studied the first passage times of pure jump processes into the upper set of a pre-specified level, since it represents the time to failure of a system (cf. Example 6.1.5). The reader is referred to a recent survey article by Li and Shaked [25].

Definition 6.2.1 The counting process $\{N(t)\}$ is called a *generalized renewal process* (or g-renewal process for short) generated from the family of conditional distributions $\{F(x|y)\}$ if, given $S_n = y$, the increment X_{n+1} follows the conditional distribution $F(x|y)$.

In the following, in order to eliminate technical difficulties, we assume that the conditional distribution $F(x|y)$ is absolutely continuous in x with density

$$f(x|y) = \frac{\partial}{\partial x} F(x|y), \qquad x \geq 0.$$

The survival distribution of $F(x|y)$ is denoted by $\overline{F}(x|y)$. Note that, if $F(x|y)$ is absolutely continuous, then the counting process $N(t)$ is *simple*, *i.e.* it admits no multiple jumps, and $N(0+) = 0$.

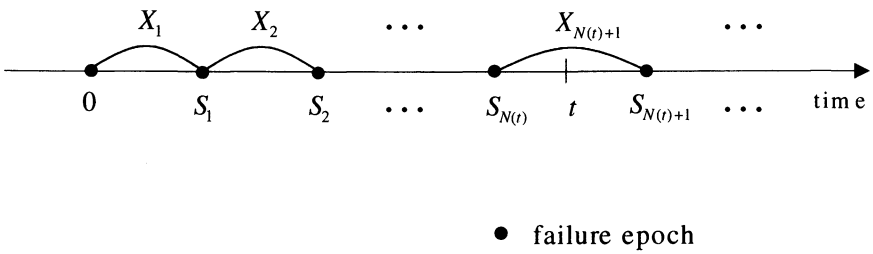


Fig. 6.1. Failure epochs and the partial-sum process $\{S_n\}$

For the g-renewal process $\{N(t)\}$, let

$$M(a,b) = N(b) - N(a), \qquad a < b,$$

and $M(a,b) = 0$ for $a \geq b$. The random quantity $M(a,b)$ counts the jumps of $N(t)$ in the interval $(a,b]$. Since, referring to Figure 6.1, we have

$$S_{N(t)} \leq y \iff N(y) \geq N(t),$$

it follows that

$$P\{S_{N(t)} \leq y\} = P\{M(y,t) = 0\} \equiv G(y|t), \qquad y \geq 0. \tag{6.11}$$

In reliability context, $G(y|t)$ is called the *interval reliability* since it is the probability that there is no failure in the interval $(y,t]$. It should be noted that $G(y|t)$ has a mass at $y = 0$, *i.e.* $G(0|t) = \overline{F}(t|0)$ for all $t \geq 0$.

Of interest in this chapter is a *generalized renewal function* (g-renewal function for short) defined by

$$H(t) = E[N(t)], \qquad t \geq 0. \tag{6.12}$$

If $H(t)$ is absolutely continuous, then one has a *generalized renewal density* $h(t)$ satisfying

$$h(t) = \frac{d}{dt}H(t), \qquad t \geq 0. \tag{6.13}$$

We call $h(t)$ a g-renewal density for short. The g-renewal density plays a key role in subsequent developments. For example, as we shall show later, the interval reliability $G(y|t)$ is expressed in terms of the g-renewal density, if it exists; see (6.21) below.

Let $L_n(y,t)$ denote the probability that there are at least n failures in the interval $(y,t]$ given that there is a failure at time y, *i.e.*

$$L_n(y,t) = P\{M(y,t) \geq n | S_{N(y+)} = y\}, \qquad 0 \leq y < t. \tag{6.14}$$

Then, we have $L_1(y,t) = F(t-y|y)$ and

$$L_{n+1}(y,t) = \int_y^t L_n(z,t) f(z-y|y)dz, \qquad n = 1, 2, \cdots. \tag{6.15}$$

Also, under the conditions stated above, the density

$$\ell_n(y,t) = \frac{\partial}{\partial t} L_n(y,t), \qquad y > 0,$$

exists and satisfies

$$\ell_{n+1}(y,t) = \int_y^t \ell_n(z,t) f(z-y|y) dz.$$

Moreover, it can be shown using a simple induction argument that

$$\ell_{n+1}(y,t) = \int_y^t \ell_n(y,z) f(t-z|z) dz.$$

See Kijima and Sumita [21] for the proofs.

Now, suppose that

(C1) For a given $t > 0$, there exists a distribution function $A_t(x)$ with $A_t(0+) = 0$ such that $F(x|y) \le A_t(x)$ for $0 < x \le t - y$ and $0 \le y \le t$.

Then, Kijima and Sumita [21] proved that the series

$$L(y,t) = \sum_{n=1}^{\infty} L_n(y,t), \qquad y < t,$$

is convergent and uniformly bounded. It follows from (6.14) that

$$L(y,t) = E[M(y,t)|S_{N(y+)=y}], \qquad 0 \le y < t, \tag{6.16}$$

and, from (6.15), we obtain

$$L(y,t) - F(t-y|y) = \int_y^t L(z,t) f(z-y|y) dz, \qquad y \le t. \tag{6.17}$$

Note from (6.12) and (6.16) that

$$H(t) = E[N(t)] = L(0,t).$$

Hence, taking $y = 0$ in (6.17), we have

$$H(t) = F(t|0) + \int_0^t L(z,t) f(z|0) dz, \qquad t \ge 0. \tag{6.18}$$

Under the given conditions, $L(y,t)$ is differentiable with respect to t and

$$\ell(y,t) = \frac{\partial}{\partial t} L(y,t) = \sum_{n=1}^{\infty} \ell_n(y,t), \qquad y < t.$$

Hence $H(t)$ is differentiable and, from (6.18) and (6.13), we have

$$h(t) = f(t|0) + \int_0^t \ell(z,t) f(z|0) dz = \ell(0,t).$$

Kijima and Sumita [21] proved that, under technical conditions, the g-renewal density $h(t)$ is a unique solution of the *generalized renewal equation* (g-renewal equation for short)

$$h(t) = f(t|0) + \int_0^t h(y)f(t-y|y)dy, \qquad t \geq 0. \tag{6.19}$$

Note that (6.19) can be viewed as a Volterra integral equation of the second kind which has the form

$$h(x) = g(x) + \int_0^x h(y)K(x,y)dy, \tag{6.20}$$

where the kernel is given by $K(x,y) = f(x-y|y)$ and $g(x) = f(x|0)$. Some numerical techniques to solve the Volterra equation (6.20) have been developed by Jagerman [10]. See Kijima, Morimura and Suzuki [18] for an approximation of the g-renewal function $H(t)$.

We next turn our attention to a relation between the interval reliability $G(y|t)$ and the g-renewal density $h(t)$. Observe that, for y, t, $\Delta > 0$, we have

$$\begin{aligned} &G(y+\Delta|t) - G(y|t) \\ &= P\{M(y+\Delta,t) = 0,\ M(y,y+\Delta) > 0\} \\ &= P\{M(y,y+\Delta) > 0\}P\{M(y+\Delta,t) = 0|M(y,y+\Delta) > 0\}, \end{aligned}$$

and that

$$\lim_{\Delta\to 0} \frac{1}{\Delta} P\{M(y,y+\Delta) > 0\} = h(y),$$

since the counting process $\{N(t)\}$ is simple. Since

$$\lim_{\Delta\to 0} P\{M(y+\Delta,t) = 0|M(y,y+\Delta) > 0\} = \overline{F}(t-y|y),$$

it follows that

$$g(y|t) \equiv \frac{\partial}{\partial y} G(y|t) = h(y)\overline{F}(t-y|y), \qquad y > 0.$$

Noting $G(0|t) = \overline{F}(t|0)$, integration of $g(x|t)$ over $[0,y]$ yields

$$G(y|t) = \overline{F}(t|0) + \int_0^y h(x)\overline{F}(t-x|x)dx, \qquad 0 \leq y \leq t. \tag{6.21}$$

It should be noted that since, from (6.19),

$$H(t) = F(t|0) + \int_0^t h(x)F(t-x|x)dx,$$

we obtain

$$G(t|t) = 1 - F(t|0) + H(t) - \int_0^t h(x)F(t-x|x)dx = 1,$$

as expected.

6.3 g-Renewal Processes in Discrete Time

In this section, we assume that the increments X_n are non-negative, discrete random variables taking values of multiples of Δt almost surely. That is, suppose

$$q_{ij} = P\{X_{n+1} = j\Delta t | S_n = i\Delta t\}, \qquad i, j = 0, 1, 2, \cdots . \tag{6.22}$$

It follows that the partial-sum process $\{S_n\}$ is a Markov chain on the non-negative lattice with time-homogeneous transition probabilities

$$p_{ij} \equiv P\{S_{n+1} = j\Delta t | S_n = i\Delta t\} = q_{i,j-i}, \qquad j \geq i. \tag{6.23}$$

As before, of interest is the associated counting process defined by

$$N_t = \sup\{n : S_n \leq t\}, \qquad t = 0, 1, 2, \cdots , \tag{6.24}$$

where $N_0 = 0$. Note that $T_z = N_z + 1$ is the first passage time for the upper set (z, ∞).

By definition (6.24), we observe the fundamental identity

$$\{N_t \geq n\} = \{S_n \leq t\}, \qquad t, n = 0, 1, 2, \cdots . \tag{6.25}$$

Since $\{S_n \leq t\} \subset \{S_n \leq t + 1\}$, it follows from (6.25) that

$$P\{N_t \geq n\} \leq P\{N_{t+1} \geq n\}, \qquad n = 0, 1, 2, \cdots , \tag{6.26}$$

so that N_t is stochastically non-decreasing in t. In the following, we obtain a monotonicity result of $\{N_t\}$ in terms of a stronger stochastic ordering under a condition on the conditional probabilities q_{ij}. Note that the stochastic monotonicity result (6.26) holds under no assumption. See Kijima [16] for details of stochastic ordering relations. It should be noted that, since distribution properties are preserved under limit, the results of this section also hold for the continuous-time case.

Definition 6.3.1 Let X and Y be non-negative, discrete random variables.

1. X is said to be greater than Y in the sense of *hazard rate ordering* if

$$P\{X \geq n\}P\{Y \geq n + 1\} \leq P\{X \geq n + 1\}P\{Y \geq n\}$$

for all $n = 0, 1, 2, \cdots$.

2. X is said to be greater than Y in the sense of *reversed hazard rate ordering* if

$$P\{X \leq n\}P\{Y \leq n + 1\} \leq P\{X \leq n + 1\}P\{Y \leq n\}$$

for all $n = 0, 1, 2, \cdots$.

Given the conditional probabilities q_{ij} in (6.22), let

$$Q_{ij} = \sum_{n=i}^{j} q_{i,n-i}, \qquad j \geq i, \tag{6.27}$$

and $Q_{ij} = 0$ for $j < i$. The next result is due to Kijima [14].

Theorem 6.3.1 Suppose that

$$Q_{ij}Q_{i+1,j+1} \geq Q_{i,j+1}Q_{i+1,j}, \qquad i, j = 0, 1, 2, \cdots, \tag{6.28}$$

where Q_{ij} are defined by (6.27). Then,

$$P\{S_n \leq t\}P\{S_{n+1} \leq t+1\} \geq P\{S_n \leq t+1\}P\{S_{n+1} \leq t\}$$

for all t, $n = 0, 1, 2, \cdots$.

Definition 6.3.1 and Theorem 6.3.1 together imply that, under the condition (6.28), the partial-sum process $\{S_n\}$ is non-decreasing in n in the sense of reversed hazard rate ordering. Since, from (6.25), we have

$$P\{N_t \geq n\} = P\{S_n \leq t\},$$

we also conclude that the g-renewal process $\{N_t\}$ is non-decreasing in t in the sense of hazard rate ordering.

Definition 6.3.2 Let X be a non-negative random variable with distribution function $F(x)$.

1. X (or $F(x)$) is said to be IHR (increasing hazard rate) if

$$\overline{F}(x+h)\overline{F}(y) \geq \overline{F}(x)\overline{F}(y+h), \qquad x < y,$$

for any $h > 0$.

2. X (or $F(x)$) is said to be DHR (decreasing hazard rate) if

$$\overline{F}(x+h)\overline{F}(y) \leq \overline{F}(x)\overline{F}(y+h), \qquad x < y,$$

for any $h > 0$.

In Definition 6.3.2, suppose that $F(t)$ is absolutely continuous with hazard rate function $r(t)$ defined in (6.3). Then, $F(t)$ is IHR (DHR, respectively) if and only if $r(t)$ is non-decreasing (non-increasing) in t. For discrete random variable X, its hazard rate function is defined by

$$r_n = \frac{P\{X = n\}}{P\{X \geq n\}}, \qquad n = 0, 1, 2, \cdots. \tag{6.29}$$

We note that X is greater than Y in the sense of hazard rate ordering if and only if the hazard rate function of X is *smaller* than that of Y, provided that the respective hazard rate functions exist. Hence, that the g-renewal process $\{N_t\}$ is non-decreasing in t in the sense of hazard rate ordering is equivalent to

$$\frac{P\{N_t = n\}}{P\{N_t \geq n\}} \geq \frac{P\{N_{t+1} = n\}}{P\{N_{t+1} \geq n\}}, \qquad n = 0, 1, 2, \cdots,$$

for all $t = 0, 1, 2, \cdots$.

The next result concerns a distribution property of the first passage time under the condition of Theorem 6.3.1. See Kijima [14] for the proof.

Theorem 6.3.2 Suppose that the condition (6.28) holds. Then, N_z as well as the first passage time T_z is IHR for every $z > 0$.

6.4 Monotonicity and Asymptotic Properties of the g-Renewal Density

First, in the ordinary renewal theory, it is well known that the renewal density is constant if and only if the underlying increments are exponential. This result can be extended to the generalized case as follows.

Theorem 6.4.1 Suppose that the g-renewal density $h(t)$ exists. Then, $h(t)$ is constant if and only if

$$f(x|y) = \theta e^{-\theta x}, \qquad x, y \geq 0,$$

for some $\theta > 0$.

The "if" part follows since, form (6.19), the equation

$$h(t) = \theta e^{-\theta t} + \int_0^t \theta e^{-\theta(t-y)} h(y) dy$$

implies $h(t) = \theta$. The "only if" part is much involved, and the reader is referred to Kijima and Sumita [21] for the proof. Theorem 6.4.1 states that, if $f(x|y)$ depends on y, then the g-renewal density $h(t)$ cannot be constant.

Next, in the ordinary renewal case, Brown [4] proved that if the increments are DHR then the renewal density is monotone and non-increasing, while from Kijima [15], if the increments follow a generalized Erlang distribution of any order, which is IHR, then it is monotone and non-decreasing. Such monotonicity results are useful in various applications, as observed in Hirayama and Kijima [9]. In this section, we consider monotonicity and asymptotic properties of the g-renewal density.

For the conditional distribution function $F(x|y)$, define the associated hazard rate function (see (6.3)) by

$$r(x|y) = \frac{f(x|y)}{\overline{F}(x|y)}, \qquad x \geq 0. \tag{6.30}$$

Recall that $F(x|y)$ is DHR if $r(x|y)$ is non-increasing in x, while it is IHR if $r(x|y)$ is non-decreasing in x where defined. According to Kijima and Sumita [21], the next result follows.

Theorem 6.4.2 Suppose that the g-renewal density $h(t)$ exists. If $F(x|y)$ is DHR (IHR, respectively) for every $y \geq 0$, and if

$$r(x|y) \geq (\leq)\ r(0|x+y), \qquad x, y \geq 0, \tag{6.31}$$

then $h(t)$ is monotone non-increasing (non-decreasing).

Unfortunately, the conditions in Theorem 6.4.2 become meaningless in the ordinary renewal context except the constant case (see Theorem 6.4.1). They are, however, meaningful for the generalized case, as the following examples suggest. These examples are taken from Kijima and Sumita [21].

Example 6.4.1 Consider the NHPP model given in Example 6.1.2. From (6.4), we obtain

$$f(x|y) = \frac{f(x+y)}{\overline{F}(y)}, \qquad \overline{F}(x|y) = \frac{\overline{F}(x+y)}{\overline{F}(y)}.$$

It follows from (6.30) that the conditional hazard rate is given by

$$r(x|y) = \frac{f(x+y)}{\overline{F}(x+y)} \equiv r_F(x+y),$$

whence $r(x|y) = r(0|x+y)$. Note that if $F(x)$ is IHR (DHR) then so is $F(x|y)$ for any $y \geq 0$. Therefore, if the underlying working-time distribution $F(x)$ is IHR (DHR) then the intensity function $h(t)$ is non-decreasing (non-increasing), as it should be.

Example 6.4.2 Let

$$f(x|y) = p(y)\theta e^{-\theta x} + (1 - p(y))\gamma e^{-\gamma x}, \qquad x \geq 0,$$

where $0 \leq p(y) \leq 1$ and $\theta > \gamma > 0$. Since a mixture of DHR distributions is DHR (see Barlow and Proschan [2]), $f(x|y)$ is DHR for each $y \geq 0$. The condition (6.31) for the non-increasing case is satisfied if $e^{(\theta-\gamma)x}p(x+y)$ is non-increasing in x. Hence, if the weighting function $p(y)$ is non-increasing in y reasonably fast, the conditions for $h(t)$ to be monotone non-increasing are satisfied.

Finally, we consider an asymptotic behavior of the g-renewal density $h(t)$.

Theorem 6.4.3 Suppose that condition (C1) holds, and that $F(x|y) \to F_0(x)$ as $y \to \infty$ where $F_0(x)$ is absolutely continuous with density $f_0(x)$ and finite mean μ_0. Suppose further that $\mu_y = \int_0^\infty x f(x|y)dx < \infty$ for all $y \geq 0$. Then $h(t) \to \mu_0^{-1}$ as $t \to \infty$.

From Theorem 6.4.3, we note that the g-renewal function $H(t)$ is asymptotically equal to t/μ_0 as $t \to \infty$. These results were proved in Kijima and Sumita [21] and can be observed in Example 6.4.2 explicitly.

6.5 On the g-Renewal Function

In the ordinary renewal context, it is well known that the renewal function $H(t)$ is *super-additive* (or *sub-additive*, respectively), *i.e.*

$$H(h) \leq (\geq)\ H(t+h) - H(t), \qquad t, h > 0, \tag{6.32}$$

if the underlying working-time distribution $F(x)$ is NBU (NWU), *i.e.*

$$\overline{F}(x+y) \leq (\geq)\ \overline{F}(x)\overline{F}(y), \qquad x, y \geq 0. \tag{6.33}$$

See, *e.g.*, Barlow and Proschan [2] for details. Knowing that $H(t)$ is super-additive or sub-additive gives us some information on the rate of increase of $H(t)$ with t. By a simple induction argument, it is easily seen that if $H(t)$ is super-additive (sub-additive) then, for any $h > 0$, we have

$$\frac{H(nh)}{nh} \geq (\leq)\ \frac{H(h)}{h}, \qquad n = 1, 2, \cdots.$$

Hence, roughly speaking, $H(t)$ is non-decreasing at a rate that is greater (less) than that of a linear function. Note that if $F(x)$ is DHR then it is NWU, and

$H(t)$ being concave implies that it is sub-additive. In the rest of this chapter, we derive these results in the generalized renewal context.

Before proceeding, it is convenient to obtain a couple of preliminary properties of $L_n(y,t)$ defined in (6.14).

Lemma 6.5.1 Suppose that $L_1(y,t) = F(t-y|y)$ is non-increasing in $y \le t$. Then, so is $L_n(y,t)$ for all $n = 1, 2, \cdots$.

Proof. Under the conditions stated earlier, we have $L_n(t,t) = F(0|t) = 0$ for any $t \ge 0$. Hence, on integration by parts, it follows from (6.15) that

$$L_{n+1}(y,t) = \int_y^t F(z-y|y) \frac{\partial}{\partial z}[-L_n(z,t)]dz.$$

Let $y < u < t$, and suppose that the lemma is proved up to some $n \ge 2$. Then, since both $L_n(z,t)$ and $F(t-z|z)$ are non-increasing in $z \le t$, we obtain

$$\begin{aligned} L_{n+1}(y,t) &\ge \int_u^t F(z-y|y) \frac{\partial}{\partial z}[-L_n(z,t)]dz \\ &\ge \int_u^t F(z-u|u) \frac{\partial}{\partial z}[-L_n(z,t)]dz = L_{n+1}(u,t), \end{aligned}$$

proving the lemma. ■

Lemma 6.5.2 Under the condition of Lemma 6.5.1, suppose further that, for any $h > 0$, $L_1(t,t+h) = F(h|t)$ is non-decreasing (non-increasing) in $t \ge 0$. Then, so is $L_n(t,t+h)$ for all $n = 1, 2, \cdots$.

Proof. Suppose that $F(h|t)$ is non-decreasing in $t \ge 0$. From (6.15), we have

$$L_{n+1}(t,t+h) = \int_0^h L_n(t+u,t+h) f(u|t)du.$$

Let $t < t'$ and suppose that $L_n(t,t+h)$ for some $n \ge 2$ is non-decreasing in $t \ge 0$. Then,

$$\begin{aligned} L_{n+1}(t,t+h) &\le \int_0^h L_n(t'+u,t'+h) f(u|t)du \\ &= \int_0^h F(u|t) \frac{\partial}{\partial u}\left[-L_n(t'+u,t'+h)\right] du \\ &\le \int_0^h F(u|t') \frac{\partial}{\partial u}\left[-L_n(t'+u,t'+h)\right] du \\ &= L_{n+1}(t',t'+h), \end{aligned}$$

where the last inequality follows since, from Lemma 6.5.1, $L_n(t+u,t+h)$ is non-increasing in $u \le h$ for all t. The other case follows similarly. ■

Define $\gamma(t) = S_{N(t)+1} - t$, the *forward recurrence time* at t. It is not hard to see that

$$P\{\gamma(t) > u\} = \overline{F}(t+u|0) + \int_0^t \overline{F}(t-x+u|x)h(x)dx, \qquad u \ge 0, \quad (6.34)$$

provided that the g-renewal density $h(t)$ exists. The next definition is parallel to (6.33).

Definition 6.5.1 The family $\{F(x|t)\}$ of conditional distribution functions is said to be g-NBU (or g-NWU, respectively) if

$$\overline{F}(x+y|t) \leq (\geq)\ \overline{F}(x|0)\overline{F}(y|t), \qquad x, y \geq 0, \tag{6.35}$$

holds for all $t \geq 0$.

A probabilistic interpretation of (6.35) is as follows. From (6.10), we observe that

$$P\{X_{n+1} > x+y|X_{n+1} > y,\ S_n = t\} = \frac{P\{X_{n+1} > x+y|S_n = t\}}{P\{X_{n+1} > y|S_n = t\}} = \frac{\overline{F}(x+y|t)}{\overline{F}(x|t)}.$$

Hence, if

$$P\{X_{n+1} > x+y|X_{n+1} > y,\ S_n = t\} \leq (\geq)\ P\{X_{n+1} > x|S_n = 0\},$$

then this assumption means that the remaining working time conditional on the event $\{X_{n+1} > y,\ S_n = t\}$ is stochastically smaller (greater) than the working time of a system with age 0. Especially, for the case that $t = 0$ in (6.35), we have the ordinary definition of NBU (NWU); see (6.33).

Lemma 6.5.3 Suppose that the family of conditional distribution functions $\{F(x|y)\}$ is g-NBU (g-NWU, respectively). Then,

$$P\{\gamma(t) > u\} \leq (\geq)\ \overline{F}(u|0), \qquad u \geq 0,$$

for any $t \geq 0$.

Proof. Suppose that $\{F(x|y)\}$ is g-NBU. Then, from (6.34), it is readily seen that

$$\begin{aligned} P\{\gamma(t) > u\} &\leq \overline{F}(u|0)\overline{F}(t|0) + \overline{F}(u|0)\int_0^t \overline{F}(t-x|x)h(x)dx \\ &= \overline{F}(u|0)P\{\gamma(t) \geq 0\}, \end{aligned}$$

and the result follows. The other case is proved similarly. ■

We are now in a position to prove the main result of this section. See Kotlyar [22] for some related results.

Theorem 6.5.1 Suppose that $F(t-y|y)$ is non-increasing in $y \leq t$, that $F(h|t)$ is non-decreasing (non-increasing, respectively) in $t \geq 0$ for any $h > 0$, and that the family of conditional distribution functions $\{\overline{F}(x|y)\}$ is g-NBU (g-NWU). Then, the g-renewal function $H(t)$ is super-additive (sub-additive).

Proof. Let $\Gamma_t(x) = P\{\gamma(t) \leq x\}$. It is easily seen that

$$H(t+h) - H(t) = \sum_{n=1}^{\infty} \int_0^h L_n(t+x, t+h)d\Gamma_t(x).$$

Also, on integration by parts, we obtain

$$\int_0^h L_n(t+x, t+h)d\Gamma_t(x) = \int_0^h \Gamma_t(x)\frac{\partial}{\partial x}[-L_n(t+x, t+h)]dx.$$

Suppose that $F(h|t)$ is non-decreasing in t and that $\{F(x|y)\}$ is g-NBU. Since $L_n(t+x, t+h)$ is non-increasing in $x \leq h$ from Lemma 6.5.1, it follows from Lemma 6.5.3 that

$$\int_0^h \Gamma_t(x) \frac{\partial}{\partial x}[-L_n(t+x, t+h)]dx \geq \int_0^h F(x|0) \frac{\partial}{\partial x}[-L_n(t+x, t+h)]dx,$$

whence

$$\int_0^h L_n(t+x, t+h)d\Gamma_t(x) \geq \int_0^h L_n(t+x, t+h)f(x|0)dx$$

for all $n = 1, 2, \cdots$. But, since $L_n(t, t+h)$ is non-decreasing in t from Lemma 6.5.2, we obtain

$$\int_0^h L_n(t+x, t+h)f(x|0)dx \geq \int_0^h L_n(x, h)f(x|0)dx, \qquad n = 1, 2, \cdots.$$

It follows that

$$H(t+h) - H(t) \geq \sum_{n=1}^{\infty} \int_0^h L_n(x, h)f(x|0)dx,$$

the right hand side being equal to $H(h)$, proving that $H(t)$ is super-additive. The sub-additive case can be proved similarly. ■

Finally, suppose that the conditions of Theorem 6.4.3 hold. Then, if $H(t)$ is super-additive (sub-additive, respectively), we have

$$H(h) \leq \ (\geq) \ \lim_{t\to\infty}[H(t+h) - H(t)] = \frac{h}{\mu_0}. \tag{6.36}$$

It should be noted that all the conditions of Theorems 6.4.3 and 6.5.1 are automatically satisfied in the ordinary renewal case. Hence, the results obtained so far in this section generalize the well-known results in the ordinary renewal context. Note, however, that the inequality (6.36) actually holds for the ordinary renewal case under a weaker assumption. See, *e.g.*, Barlow and Proschan [2].

6.6 A General Repair Model

In this section, we consider a general repair model given in Example 6.1.4. The family of conditional distribution functions to be considered is then given by

$$F(x|y) = \int_0^{\infty} \frac{F(x+u) - F(u)}{\overline{F}(u)} dG_y(u), \qquad x, y \geq 0, \tag{6.37}$$

where $F(x)$ is the working-time distribution of a new system and $G_y(x)$ is the distribution function of the effective age Y when $S_n = y$. The survival distribution is

$$\overline{F}(x|y) = \int_0^{\infty} \frac{\overline{F}(x+u)}{\overline{F}(u)} dG_y(u), \qquad x, y \geq 0. \tag{6.38}$$

Recall that $F(x)$ is IHR (DHR, respectively) if $\overline{F}(x+u)/\overline{F}(u)$ is non-increasing (non-decreasing) in $u \geq 0$ for all $x > 0$.

The results of this section were proved in Baxter, Kijima and Tortorella [3]; however, we recover their results from the general results obtained in Section 6.5. For comparisons of failure processes, *i.e.* g-renewal processes, in terms of repair activities, we refer to Baxter, Kijima and Tortorella [3]. See also Last and Szekli [23] in the general content. In the following, we assume that the condition (C1) is satisfied and the g-renewal density exists. Also, to avoid unnecessary technical difficulties, it is assumed that $G_y(0+) = 0$ and $G_y(\infty) = 1$ for all $y \geq 0$.

Let Y_y denote the effective age when $S_n = y$. For $y > 0$, we define $A_y = Y_y/y$. If $Y_y = ay$ almost certainly as in (6.9), then $A_y = a$ almost certainly for any y, which represents the degree of repair introduced in Example 6.1.3. Recall that, for non-negative random variables X and Y, X is said to be *stochastically smaller* than Y if $\overline{F}_X(x) \leq \overline{F}_Y(x)$ for all $x \geq 0$, where $\overline{F}_X(x)$ (or $\overline{F}_Y(x)$, respectively) is the survival distribution of X (Y).

Lemma 6.6.1 $F(x-y|y)$ is non-increasing in $y \leq x$ if one of the following conditions holds.

1. $F(x)$ is DHR and Y_y is stochastically non-decreasing in $y \geq 0$.
2. $F(x)$ is IHR, $0 \leq A_y \leq 1$ for all y, and A_y is stochastically non-increasing in $y \geq 0$.

Proof. For the first part of the lemma, let $y < y'$. Then, from (6.38) and integration by parts, we obtain

$$\begin{aligned}
\overline{F}(x-y|y) &\leq \int_0^\infty \frac{\overline{F}(x-y'+u)}{\overline{F}(u)} dG_y(u) \\
&= \overline{F}(x-y') + \int_0^\infty \overline{G}_y(u) \frac{\partial}{\partial u}\left[\frac{\overline{F}(x-y'+u)}{\overline{F}(u)}\right] du \\
&\leq \overline{F}(x-y') + \int_0^\infty \overline{G}_{y'}(u) \frac{\partial}{\partial u}\left[\frac{\overline{F}(x-y'+u)}{\overline{F}(u)}\right] du \\
&= \overline{F}(x-y'|y'),
\end{aligned}$$

where the second inequality holds since $F(x)$ is DHR and $\overline{G}_y(x) \leq \overline{G}_{y'}(x)$ under the assumption.

Next, to prove the second part, let $\Gamma_y(x)$ denote the distribution function of A_y. Then, since $0 \leq A_y \leq 1$ and from (6.38), we have

$$\overline{F}(x-y|y) = \int_0^1 \frac{\overline{F}(x-y+ay)}{\overline{F}(ay)} d\Gamma_y(a), \qquad y \leq x. \tag{6.39}$$

Note that

$$\begin{aligned}
&\frac{\partial}{\partial y}\left[\frac{\overline{F}(x-(1-a)y)}{\overline{F}(ay)}\right] \\
&\quad = \frac{\overline{F}(x-(1-a)y)}{\overline{F}(ay)}\{(1-a)r(x-(1-a)y) + ar(ay)\},
\end{aligned}$$

which is non-negative since $0 \leq a \leq 1$. Also,

$$\frac{\partial}{\partial a}\left[\frac{\overline{F}(x-y+ay)}{\overline{F}(ay)}\right] = y\frac{\overline{F}(x-y+ay)}{\overline{F}(ay)}\{r(ay) - r(ay+x-y)\},$$

which is non-positive since $F(x)$ is IHR. It follows that $\overline{F}(x-y+ay)/\overline{F}(ay)$ is non-decreasing in y and non-increasing in a. The rest of the proof is similar to that of the first part, and the proof is complete. ■

The assumption that the effective age Y_y depends on the cumulative operating time may be interpreted as a change in the quality of repair over time, perhaps reflecting an increase in the skill of the repairman with experience or improved repair procedure, in which case it is plausible to assume that Y_y is stochastically non-increasing in y. Alternatively, as a system ages, it may become increasingly difficult to perform a satisfactory repair, in which case we assume that Y_y is stochastically non-decreasing in y. The assumption $A_y \leq 1$ implies that repair never clumsy.

The assumption A_y is stochastically non-increasing in y is weaker than the assumption that Y_y is stochastically non-increasing in y. In fact, while Y_y is stochastically non-decreasing in y, A_y can be stochastically non-increasing in y (see Theorem 6.6.1 below). For example, consider the case given in Example 6.1.3. In this case, $A_y = a$ almost surely while $Y_y = ay$ is non-decreasing in y. We note that there is a flaw in Lemma 4.3 of Baxter, Kijima and Tortorella [3], which affects their Theorem 4.5. The following results correct the flaw.

Lemma 6.6.2 Suppose that Y_y is stochastically non-decreasing in $y \geq 0$. If $F(x)$ is DHR (IHR, respectively), then $F(h|t)$ is non-increasing (non-decreasing) in $t \geq 0$ for all $h > 0$.

Proof. The lemma follows since

$$\overline{F}(h|t) = \overline{F}(h) + \int_0^\infty \overline{G}_t(u)\frac{\partial}{\partial u}\left[\frac{\overline{F}(h+u)}{\overline{F}(u)}\right] du$$

and $\overline{G}_t(u) \leq \overline{G}_{t'}(u)$ whenever $t < t'$. ■

Lemma 6.6.3 The family of conditional distribution functions $\{F(x|y)\}$ is g-NWU (g-NBU, respectively) if the underlying working-time distribution $F(x)$ is NWU (NBU).

Proof. Suppose that $F(x)$ is NWU. Then, from (6.33),

$$\overline{F}(t+v+u) \geq \overline{F}(t+u)\overline{F}(v), \qquad t, v, u \geq 0.$$

It follows that

$$\int_0^\infty \frac{\overline{F}(t+v+u)}{\overline{F}(u)} dG_y(u) \geq \overline{F}(v)\int_0^\infty \frac{\overline{F}(t+u)}{\overline{F}(u)} dG_y(u),$$

which can be rewritten, from (6.38), as

$$\overline{F}(t+v|y) \geq \overline{F}(v|0)\overline{F}(t|y), \qquad t, v \geq 0,$$

for all $y \geq 0$. The other case follows similarly. ■

Corollary 6.6.1 If the underlying working-time distribution $F(x)$ is NBU (NWU), then

$$P\{\gamma(t) > u\} \leq \ (\geq)\ \overline{F}(u), \qquad u \geq 0,$$

for any $t \geq 0$. That is, the forward recurrence time at any time is stochastically smaller (greater) than the working time of a new system.

The next theorem follows from Lemmas 6.6.1 – 6.6.3 and Theorem 6.5.1.

Theorem 6.6.1 Suppose that Y_y is stochastically non-decreasing in $y \geq 0$.

1. If $F(x)$ is DHR, then the g-renewal function $H(t)$ is sub-additive.
2. If $F(x)$ is IHR, $0 \leq A_y \leq 1$ for all y, and A_y is stochastically non-increasing in $y \geq 0$, then $H(t)$ is super-additive.

References

1. Barlow, R. E. and Proschan, F. (1965), Mathematical Theory of Reliability. John Wiley & Sons, New York
2. Barlow, R. E. and Proschan, F. (1975), Statistical Theory of Reliability and Life Testing. Holt, Rinehart and Winston, New York
3. Baxter, L. A., Kijima, M. and Tortorella, M. (1996), "A point process model for the reliability of a maintained system subject to general repair," *Stochastic Models*, **12**, 37–65
4. Brown, M. (1980), "Bounds, inequalities, and monotonicity properties for some specialized renewal processes," *Annals of Probability*, **8**, 227–240
5. Çinlar, E. (1975), Introduction to Stochastic Processes. Prentice Hall, New Jersey
6. Dagpunar, J. S. (1998), "Some properties and computational results for a general repair process," *Naval Research Logistics*, **45**, 391–405
7. Dorado, C., Hollander, M. and Sethuraman, J. (1997), "Nonparametric estimation for a general repair model," *Annals of Statistics*, **25**, 1140–1160
8. Guo, R. and Love, C. E. (1992), "Statistical analysis of an age model for imperfectly repaired systems," *Quality and Reliability Engineering International*, **8**, 133–146
9. Hirayama, T. and Kijima, M. (1992), "Single machine scheduling problem when the machine capacity varies stochastically," *Operations Research*, **40**, 376–383
10. Jagerman, D. (1985), "Certain Volterra integral equations arising in queueing," *Stochastic Models*, **1**, 239–256
11. Keener, R. W. (1982), "Renewal theory for Markov chains on the real line," *Annals of Probability*, **10**, 942–954
12. Kijima, M. (1983), "Replacement policies in the case that failure distributions depend on the number of failures," *Journal of the Operations Research Society of Japan*, **26**, 347–356
13. Kijima, M. (1989), "Some results for repairable systems with general repair," *Journal of Applied Probability*, **26**, 89–102
14. Kijima, M. (1989), "Uniform monotonicity of Markov processes and its related properties," *Journal of the Operations Research Society of Japan*, **32**, 475–490
15. Kijima, M. (1992), "Further monotonicity properties of renewal processes," *Advances in Applied Probability*, **25**, 575–588
16. Kijima, M. (1997), Markov Processes for Stochastic Modeling. Chapman & Hall, London
17. Kijima, M., Li, H. and Shaked, M. (2001), "Stochastic processes in reliability," in Stochastic Processes: Theory and Methods (Rao, C.R. and Shanbhag, D.N. ed.). Handbook in Statistics, **18**, forthcoming.

18. Kijima, M., Morimura, H. and Suzuki, Y. (1988), "Periodical replacement problem without assuming minimal repair," *European Journal of Operational Research*, **37**, 194–203
19. Kijima, M. and Nakagawa, T. (1991), "A cumulative damage shock model with imperfect preventive maintenance," *Naval Research Logistics*, **38**, 145–156
20. Kijima, M. and Nakagawa, T. (1992), "Replacement policies of a shock model with imperfect preventive maintenance," *European Journal of Operational Research*, **57**, 100–110
21. Kijima, M. and Sumita, U. (1986), "A useful generalization of renewal theory: Counting processes governed by nonnegative Markovian increments," *Journal of Applied Probability*, **23**, 71–88
22. Kotlyar, V. Y. (1990), "Nonstationary bounds of the characteristics of Markovian renewal processes inclasses of aging distributions," *Kibernetika*, **88**, 74–78
23. Last, G. and Szekli, R. (1998), "Stochastic comparison of repairable systems by coupling," *Journal of Applied Probability*, **35**, 348–370
24. Last, G. and Szekli, R. (1999), "Time and Palm stationarity of repairable systems," *Stochastic Processes and Their Applications*, **79**, 17–43
25. Li, H. and Shaked, M. (1999), "Aging first-passage times," *Encyclopedia of Statistical Sciences* (Kotz, S., Read, C. B. and Banks, D. L. eds.). **11**, 11–20, John Wiley & Sons, New York
26. Makis, V. and Jardine, A. K. S. (1993), "A note on optimal replacement policy under general repair," *European Journal of Operational Research*, **69**, 75–82
27. Nakagawa, T. (2001), "Imperfect preventive maintenance models," in this book
28. Pham, H. and Wang, H. (1996), "Imperfect maintenance," *European Journal of Operational Research*, **94**, 425–438
29. Ross, S. M. (1983), Stochastic Processes. John Wiley & Sons, New York
30. Stadje, W. and Zuckerman, D. (1991), "Optimal maintenance strategies for repairable systems with general degree of repair," *Journal of Applied Probability*, **28**, 384–396

7. Two-Unit Redundant Models

Toshio Nakagawa
Department of Industrial Engineering,
Aichi Institute of Technology
Toyota 470-0392, Japan

Summary.

A two-unit system is the most fundamental redundant model in reliability theory, and it has been analyzed by many researchers over the past 40 years. In recent years, such systems have spread among industries and been used even in a daily life. In particular, computer systems have been composed of various types of redundant ones to achieve high reliability and accuracy with their rapid advancement and low cost. This chapter surveys the early results of a two-unit standby system, its preventive maintenance and other systems, using a unique modification of the regeneration point techniques of Markov renewal processes. Reliability quantities of a two-unit system are obtained, and optimum preventive maintenance policies which maximize the mean time to system failure and the availability are analytically discussed. Finally, other two-unit systems are briefly presented.

Keywords: two units, standby system, MTTF, availability, preventive maintenance, parallel system, Markov renewal process

7.1 Introduction

A system with high reliability can be achieved by the use of redundancy and maintenance. This chapter considers a redundant system of two repairable units. Such two-unit redundant systems are basic in reliability theory and can be found in many practical applications, *e.g.*, computer systems in industries and military applications, electric power stations and distributors, and electronic telephone exchange systems. Simple but basic two-unit standby and other two-unit redundant systems are discussed thoroughly in this chapter.

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

Let us sketch the early working of such two-unit redundant systems. The first contribution to reliability analysis of two-unit redundant systems was made by Epstein and Hosford [1]. Gaver [2], [3] treated similar models assuming general repair time distributions. Gnedenko [4], Srinivasan [5] and Liebowitz [6] discussed a two-unit standby system with more generalized assumptions. Gnedenko [7] and Solovyev [8] obtained the limiting theorems which gave the asymptotic distributions to system failure. Srinivasan [9] also discussed a similar system with noninstantaneous switchover. In other contributions, Mine and Osaki [10], Osaki [11], [12], [13], [14], [15], [16], [17] and Osaki and Asakura [18] discussed several variations of two-unit redundant systems by using Markov renewal processes. Moreover, several interesting two-unit redundant systems were discussed by Garkavi and Gogolovskiy [19], Harris [20], Mazumdar [21], Buzacott [22], Linton and Braswell [23], and others. Most of the content of this chapter is based on the authors' own original works that appeared in Osaki and Nakagawa [24], Nakagawa and Osaki [25], [26], [27], [28], [29], [30], [31], [32] and Yamada and Osaki [33]. Most of the results in [25] and [31] were rewritten by discrete time processes in [34].

Kodama and Deguchi [35], Kodama *et al.* [36], Adachi and Kodama [37], Ohashi and Nishida [38], Parathasarathy [39], Wiens [40], Jack [41], Wells [42] and Vanderperre [43] studied two-unit redundant systems under more general forms. Mine and Kawai [44] and Kumar *et al.* [45] discussed the maintenance policies for two-unit systems with degraded states. Further, Nakagawa and Osaki [46], Kumar and Agarwal [47] and Yearout *et al.* [48] reviewed many literatures of Goel, Gupta, Gopalan, Kapur, Kapoor, Kumar, Naidu, Ramanarayanan, Subramanian and so on. Using the results of two-unit systems, Laprie *et al.* [49], De and Krakan [50], Trivedi and Geist [51], Ng [52], Hu and Mouftah [53], Ibe *et al.* [54], Walker [55], Reibman [56] and Choi *et al.* [57] applied them to the analysis of computer systems and fault-tolerant systems. Further, Gai *et al.* [58] analyzed a full-authority, dual-redundant aircraft engine controller, and Tapiero and Hsu [59] studied a randomized two-station machining process with blocking.

In Section 7.2, we introduce a two-unit standby system and derive the following quantities:

(i) The first-passage time distributions.
(ii) The expected numbers of visits to a specified state during a given finite interval.
(iii) The transition probabilities.

The above quantities are obtained by using the results of a unique modification of Markov renewal processes.

In Section 7.3, we treat the preventive maintenance (pm) of a two-unit standby redundant system. Then, we obtain the optimum pm policies which maximize the mean time to system failure and the availability.

In Section 7.4, we describe several interesting variants of two-unit systems and briefly discuss the main results for such models. We finally show developments and scopes of other two-unit systems.

7.2 Two-Unit Standby System

A two-unit standby redundant system with a single repairman (or repair facility) is one of the most fundamental and important redundant systems in reliability theory: a system consists of two units, where one unit is used for operation and the other is in standby as an initial condition. If an operating unit fails, then it undergoes repair immediately and the other standby unit takes over its operation. Either of the two units is alternately operating. It can be said that a system failure occurs when both units are down simultaneously.

Gaver [3] obtained the Laplace-Stieltjes (LS) transform of the distribution of the time to system failure and the mean time to failure (MTTF) for the model with exponential failure times and general repair times. Gnedenko [4], [60] and Srinivasan [5] extended the results for the model with both general failure and repair times. Further, Osaki [12] obtained the same results by using the signal-flow graph method.

In the model above, we consider the system of two identical units and are interested in the following quantities: (i) The first-passage time distributions, (ii) the expected numbers of visits to a state during $(0, t]$, and (iii) the transition probabilities. We obtain the above three quantities and their associated means or limiting values, and derive some reliability properties of a two-unit standby system.

A unique modification of the regeneration point techniques of Markov renewal processes ([61], [62]) is applied for analyzing the system, since some epochs (or time instants) at which the system makes transitions into some states are not regeneration points.

7.2.1 Model and assumptions

Consider a two-unit standby redundant system of two identical units: an operating unit has a general failure time distribution $F(t)$ with finite mean $1/\lambda$, and a failed unit has a general repair time distribution $G(t)$ with finite mean $1/\mu$. If an operating unit fails and the other unit is in standby, the failed unit undergoes repair immediately and the standby unit takes over its operation immediately. However, if an operating unit fails while the other unit is under repair, the failed unit must wait for repair until a repairman is free. This situation means that a system failure has occurred. It is also assumed that a failed unit recovers its functioning upon repair completion and is as good as new. A unit in standby neither deteriorates nor fails in the standby interval. Two units are used alternately for its operation, as described

above. Even if a system failure occurs, the system can operate again upon repair completion. That is, the system repeats up and down alternately. All random variables considered here are independent of each other.

To analyze the above system, we define the following three system states which represent the states of the process:

State -1: One unit begins to operate and the other is in standby.

State 0: One unit is operating and the other unit is in standby.

State 1: One unit is operating and the other unit is under repair.

State 2: One unit is under repair and the other unit is waiting for repair.

The system states defined above form a Markov renewal process: state -1 represents an initial time instant. An epoch at which the system makes a transition into state 1 is a regeneration point. However, the epochs at which the system makes a transition into state j ($j = 0, 2$) are not regeneration points except all failure and repair times are exponential.

It is noted that the index i ($i = 0, 1, 2$) of system states denotes the number of failed units. Further, the epoch for state 0 is when a system recovery occurs, and the epoch for state 2 is when a system failure occurs.

Define a mass function (or one-step transition distribution) $Q_{ij}(t)$ from state i ($i = -1, 1$) to state j ($j = 0, 1, 2$) by the probability that after making a transition into state i, the process next makes a transition into state j, in a smaller amount of time than time t in a Markov renewal process. Then, from the theory of Markov renewal processes, the following mass functions are obtained:

$$Q_{-11}(t) = F(t), \tag{7.1}$$

$$Q_{10}(t) = \int_0^t \overline{F}(u)dG(u), \tag{7.2}$$

$$Q_{12}(t) = \int_0^t \overline{G}(u)dF(u), \tag{7.3}$$

where $\overline{F}(t) \equiv 1 - F(t)$ and $\overline{G}(t) \equiv 1 - G(t)$.

It is impossible to define the mass functions $Q_{01}(t)$ and $Q_{21}(t)$ since the epochs for states 0 and 2 are not regeneration points. Thus, define a new mass function $Q_{11}^{(0)}(t)$ (or $Q_{11}^{(2)}(t)$) which is the recurrence time distribution for state 1 via state 0 (or state 2). That is, $Q_{11}^{(0)}(t)$ (or $Q_{11}^{(2)}(t)$) is the probability that after making a transition into state 1, the process next makes a transition into state 0 (or state 2) and returns to state 1, in a smaller amount of time than time t. In a similar fashion, the new mass functions are given by

$$Q_{11}^{(0)}(t) = \int_0^t G(u)dF(u), \tag{7.4}$$

$$Q_{11}^{(2)}(t) = \int_0^t F(u)dG(u). \tag{7.5}$$

Throughout this chapter, we denote the LS transform of the function by the corresponding lowercase letter. For instance, $q_{ij}(s) \equiv \int_0^\infty e^{-st}dQ_{ij}(t)$ for

$Re(s) > 0$, and so on. The LS transforms $q_{ij}(s)$ and $q_{ij}^{(k)}(s)$ of the mass functions $Q_{ij}(t)$ and $Q_{ij}^{(k)}(t)$ can easily be obtained from (7.1) $\sim$ (7.5).

7.2.2 First-passage time distributions

We obtain the first-passage time distributions and the mean first-passage times by using the mass functions.

Let $H_{ij}(t)$ denote the first-passage time distributions from state i ($i = -1, 1$) to state j ($j = 0, 1, 2$). First, suppose that the process starts in the epoch for state i. Then, from (7.4) and (7.5), the recurrence time distribution $H_{11}(t)$ for state 1 is given by

$$H_{11}(t) = Q_{11}^{(0)}(t) + Q_{11}^{(2)}(t) = F(t)G(t), \tag{7.6}$$

which can be interpreted as the probability that the process comes back to state 1 via state 0 or state 2, whichever occurs first. Thus, the mean recurrence time l_{11} for state 1 is

$$l_{11} \equiv \int_0^\infty t dH_{11}(t) = 1/\lambda + 1/\mu - 1/\gamma, \tag{7.7}$$

where $1/\gamma \equiv \int_0^\infty \overline{F}(t)\overline{G}(t)dt$.

The first-passage time distributions $H_{ij}(t)$ ($j = 0, 2$) are given by solving the following renewal-type equation:

$$H_{10}(t) = Q_{10}(t) + Q_{11}^{(2)}(t) * H_{10}(t), \tag{7.8}$$

where the asterisk mark represents the Stieltjes convolution, *i.e.*, $A(t)*B(t) \equiv \int_0^t B(t-u)dA(u) = \int_0^t A(t-u)dB(u)$ for any $A(t)$ and $B(t)$. The first term of the right-hand side of (7.8) can be interpreted as the probability that the process goes to state 0 directly, and the second term is the probability that the process goes to state 0 after returning to state 1 via state 2. Similarly,

$$H_{12}(t) = Q_{12}(t) + Q_{11}^{(0)}(t) * H_{12}(t). \tag{7.9}$$

Next, suppose that the process starts from state -1. Then,

$$H_{-1j}(t) = Q_{-11}(t) * H_{1j}(t), \qquad j = 0, 2. \tag{7.10}$$

Thus, taking the LS transforms of (7.8), (7.9) and (7.10), and solving them,

$$h_{-10}(s) = \frac{q_{-11}(s)q_{10}(s)}{1 - q_{11}^{(2)}(s)}, \tag{7.11}$$

$$h_{-12}(s) = \frac{q_{-11}(s)q_{12}(s)}{1 - q_{11}^{(0)}(s)}. \tag{7.12}$$

The mean first-passage times l_{-1j} ($j = 0, 2$), starting from state -1, are

$$l_{-10} = \frac{1}{\lambda} + \frac{1}{\mu[1 - q_{11}^{(2)}(0)]}, \tag{7.13}$$

$$l_{-12} = \frac{1}{\lambda} + \frac{1}{\lambda[1 - q_{11}^{(0)}(0)]}, \tag{7.14}$$

which represent the mean first-passage times to system recovery and system failure, respectively. Note that (7.12) and (7.14) are the LS transforms of the distribution of the time to system failure and its mean time.

Further, it can be seen that $H_{10}(t)$ and l_{10} represent the distribution of the busy period of a repairman and its mean time, respectively. It is finally noted that the mean downtime l_d after system failure is

$$l_d = \int_0^\infty t dQ_{11}^{(2)}(t) - \int_0^\infty t dQ_{12}(t) = \int_0^\infty F(t)\overline{G}(t)dt. \tag{7.15}$$

7.2.3 Expected numbers of visits to state

Consider the expected numbers of visits to state j ($j = 0, 1, 2$) during the interval $(0, t]$. Let $M_{ij}(t)$ be the expected numbers of occurrences of state j during $(0, t]$, starting in the epoch for state i ($i = -1, 1$) at time 0, where the first visit to state j is not counted if $i = j$.

First, suppose that the process starts from state 1 at time 0. Then, we have the following renewal-type equation:

$$\begin{aligned} M_{10}(t) &= Q_{10}(t) - Q_{11}^{(0)}(t) + Q_{11}^{(0)}(t) * [1 + M_{10}(t)] + Q_{11}^{(2)}(t) * M_{10}(t) \\ &= Q_{10}(t) + H_{11}(t) * M_{10}(t), \end{aligned} \tag{7.16}$$

where the first term on the right-hand side is the probability that the process goes to state 0 directly and then stays at state 0, the second term is the expected number that the process returns to state 1 via state 0 and goes to state 0, and finally, the third term is the expected number that the process returns to state 1 via state 2 and then goes to state 0. In a similar fashion,

$$\begin{aligned} M_{11}(t) &= Q_{11}^{(0)}(t) * [1 + M_{11}(t)] + Q_{11}^{(2)}(t) * [1 + M_{11}(t)] \\ &= H_{11}(t) * [1 + M_{11}(t)], \end{aligned} \tag{7.17}$$

$$\begin{aligned} M_{12}(t) &= Q_{12}(t) - Q_{11}^{(2)}(t) + Q_{11}^{(0)}(t) * M_{12}(t) + Q_{11}^{(2)}(t) * [1 + M_{12}(t)] \\ &= Q_{12}(t) + H_{11}(t) * M_{12}(t). \end{aligned} \tag{7.18}$$

The expected numbers of visits to state j ($j = 0, 1, 2$) during $(0, t]$, starting from state -1, are

$$M_{-11}(t) = Q_{-11}(t) * [1 + M_{11}(t)], \tag{7.19}$$

$$M_{-1j}(t) = Q_{-11}(t) * M_{1j}(t), \qquad j = 0, 2. \tag{7.20}$$

Thus, combining (7.16)~(7.20), we have the LS transforms of the expected numbers:

$$m_{-11}(s) = \frac{q_{-11}(s)}{1 - h_{11}(s)}, \tag{7.21}$$

$$m_{-1j}(s) = \frac{q_{-11}(s)q_{1j}(s)}{1 - h_{11}(s)}, \qquad j = 0, 2. \tag{7.22}$$

Note that $m_{-10}(s)$ and $m_{-12}(s)$ are the LS transforms of expected numbers of system recovery and system failure during $(0, t]$, respectively.

Further, if $F(t)$ and $G(t)$ are nonlattice and their means are finite, there exist the limits $M_j \equiv \lim_{t\to\infty} M_{ij}(t)/t$ $(i = -1, 1; j = 0, 1, 2)$ of expected numbers of visits to state j per unit of time in the steady-state, which are independent of an initial state i [63]. Thus, by applying the Tauberian theorem in (7.21) and (7.22), respectively,

$$M_1 = \frac{1}{l_{11}}, \tag{7.23}$$

$$M_j = \frac{q_{1j}(0)}{l_{11}}, \qquad j = 0, 2. \tag{7.24}$$

A quantity of some interest is the total expected number of units which have failed (or have been repaired) during $(0, t]$. Let $M_{if}(t)$ (or $M_{ir}(t)$) be the expected number of failed units (or repaired units) during $(0, t]$, if the process starts from state i $(i = -1, 1)$. Then,

$$\begin{aligned} M_{1f}(t) &= Q_{12}(t) - Q_{11}^{(2)}(t) + H_{11}(t) * [1 + M_{1f}(t)] \\ &= F(t)\overline{G}(t) + H_{11}(t) * [1 + M_{1f}(t)], \end{aligned} \tag{7.25}$$

$$\begin{aligned} M_{1r}(t) &= Q_{10}(t) - Q_{11}^{(0)}(t) + H_{11}(t) * [1 + M_{1r}(t)] \\ &= \overline{F}(t)G(t) + H_{11}(t) * [1 + M_{1r}(t)]. \end{aligned} \tag{7.26}$$

If the process starts from state -1, then

$$M_{-1f}(t) = Q_{-11}(t) * [1 + M_{1f}(t)], \tag{7.27}$$

$$M_{-1r}(t) = Q_{-11}(t) * M_{1r}(t). \tag{7.28}$$

Thus, combining them, we have

$$m_{-1f}(s) = \frac{q_{-11}(s)[1 + q_{12}(s) - q_{11}^{(2)}(s)]}{1 - h_{11}(s)}, \tag{7.29}$$

$$m_{-1r}(s) = \frac{q_{-11}(s)[q_{10}(s) + q_{11}^{(2)}(s)]}{1 - h_{11}(s)}. \tag{7.30}$$

7.2.4 Transition probabilities

Denote as $P_{ij}(t)$ the probability that the process is in state j $(j = 0, 1, 2)$ at time t, starting from state i $(i = -1, 1)$ at time 0. Then, we have

$$P_{10}(t) = Q_{10}(t) - Q_{11}^{(0)}(t) + H_{11}(t) * P_{10}(t)$$
$$= \overline{F}(t)G(t) + H_{11}(t) * P_{10}(t), \tag{7.31}$$
$$P_{11}(t) = 1 - Q_{10}(t) - Q_{12}(t) + H_{11}(t) * P_{11}(t)$$
$$= \overline{F}(t)\overline{G}(t) + H_{11}(t) * P_{11}(t), \tag{7.32}$$
$$P_{12}(t) = Q_{12}(t) - Q_{11}^{(2)}(t) + H_{11}(t) * P_{12}(t)$$
$$= F(t)\overline{G}(t) + H_{11}(t) * P_{12}(t), \tag{7.33}$$

which can be similarly interpreted as described in the preceding sections.

Further, if the process starts from state -1, then

$$P_{-10}(t) = 1 - Q_{-11}(t) + Q_{-11}(t) * P_{10}(t)$$
$$= \overline{F}(t) + Q_{-11}(t) * P_{10}(t), \tag{7.34}$$
$$P_{-1j}(t) = Q_{-11}(t) * P_{1j}(t), \qquad j = 0, 2. \tag{7.35}$$

Thus, the LS transforms of $P_{-1j}(t)$ $(j = 0, 1, 2)$ are

$$p_{-10}(s) = 1 - q_{-11}(s) + \frac{q_{-11}(s)[q_{10}(s) - q_{11}^{(0)}(s)]}{1 - h_{11}(s)}, \tag{7.36}$$

$$p_{-11}(s) = \frac{q_{-11}(s)[1 - q_{10}(s) - q_{12}(s)]}{1 - h_{11}(s)}, \tag{7.37}$$

$$p_{-12}(s) = \frac{q_{-11}(s)[q_{12}(s) - q_{11}^{(2)}(s)]}{1 - h_{11}(s)}, \tag{7.38}$$

where it is evident that

$$P_{-10}(t) + P_{-11}(t) + P_{-12}(t) = 1. \tag{7.39}$$

If $F(t)$ and $G(t)$ are non-lattice and their means are finite, there exist the limiting probabilities $P_j \equiv \lim_{t\to\infty} P_{ij}(t)$ $(i = -1, 1; j = 0, 1, 2)$ which are independent of an initial state i. Thus, by applying the Tauberian theorem in (7.36)$\sim$(7.38),

$$P_0 = 1 - \frac{1}{\mu l_{11}}, \tag{7.40}$$

$$P_1 = -1 + \frac{1/\lambda + 1/\mu}{l_{11}}, \tag{7.41}$$

$$P_2 = 1 - \frac{1}{\lambda l_{11}}. \tag{7.42}$$

It is noted that $A(t) \equiv P_{-10}(t) + P_{-11}(t)$ represents the pointwise availability of the system at time t, and $\overline{A}(t) \equiv P_{-12}(t)$ represents the pointwise unavailability at time t, given that the two units are initially good. It is also noted that $A \equiv P_0 + P_1$ represents the steady-state (or limiting interval) availability and $\overline{A} \equiv P_2$ represents the steady-state unavailability.

7.3 Preventive Maintenance of Two-Unit Systems

Preventive maintenance (pm) is of great interest in reliability theory, and many contributions to pm policies have been made in Chapter 5. For a two-unit standby system, Rozhdestvenskiy and Fanarzhi [64] considered two pm policies and obtained one method of approximating the optimum pm policy which minimizes the mean time to system failure. Osaki and Asakura [18] showed that the mean time to system failure of the system with pm is greater than that with only repair maintenance under the suitable conditions.

Berg [65], [66], [67] considered the replacement policy for a two-unit system where, at failure points of one unit, the other unit is replaced if its age exceeds a control limit. Almeida and Souza [68] obtained the maintenance strategy for a two-unit standby system with waiting time for repair. Further, Gupta and Kumar [69], and Pullen and Thomas [70] examined the opportunistic replacement policy for two-unit systems.

Consider a two-unit standby redundant system with repair and pm maintenances. We derive optimum pm policies which maximize the mean time to system failure and the availability. We obtain such quantities above by using the technique of Markov renewal processes used in Section 7.2, and discuss optimum policies under suitable conditions.

7.3.1 Model and analysis

Consider the same two-unit standby system as in Section 7.2, where two units are statistically identical: it is assumed that an operating unit has a failure time distribution $F(t)$ with finite mean $1/\lambda$ and a failed unit has a repair time distribution $G_1(t)$ with finite mean $1/\mu_1$. The other assumptions are the same as in Section 7.2.

We consider the following pm policy for an operating unit. When an operating unit works for a specified time T without failure, we stop the operation of an operating unit for pm. It is assumed that the time to pm completion has a general distribution $G_2(t)$ with finite mean $1/\mu_2$, and a preventively maintained unit is as good as new upon pm completion. Further, we make the following two assumptions:

1. Pm of an operating unit is done only if the other unit is in standby.
2. An operating unit that has forfeited pm because of assumption 1 undergoes pm just upon repair or pm completion of the other unit.

Under the above assumptions, we derive the mean time to system failure and the availability as follows [26]: the mean time to system failure is

$$l(T) = \frac{\left[1/\gamma_1 + \int_0^T (\theta_1 + G_1(t))\overline{F}(t)dt\right]\int_0^T (\theta_2 + G_2(t))dF(t) + \left[1/\gamma_2 + \int_0^T (\theta_2 + G_2(t))\overline{F}(t)dt\right]\left[1 - \int_0^T (\theta_1 + G_1(t))dF(t)\right]}{\theta_1 \int_0^T (\theta_2 + G_2(t))dF(t) + \theta_2\left[1 - \int_0^T (\theta_1 + G_1(t))dF(t)\right]}, \tag{7.43}$$

where

$$\theta_i \equiv \int_0^\infty \overline{G_i}(t)dF(t), \qquad i = 1, 2,$$

$$1/\gamma_i \equiv \int_0^\infty \overline{F}(t)\overline{G_i}(t)dt, \qquad i = 1, 2.$$

Consider the following two special cases of pm policies. The first one is that where an operating unit undergoes pm upon repair or pm completion. That is, a repairman is always busy for repair or pm. In this case, setting $T = 0$ in (7.43), we have

$$l(0) = \frac{1}{\theta_2\gamma_2}. \tag{7.44}$$

The second is the case where no pm is done. In this case, setting $T = \infty$, we have

$$l(\infty) = \frac{1}{\lambda} + \frac{1}{\lambda\theta_1}, \tag{7.45}$$

which is equal to (7.14).

The steady-state availability $A(T)$ is

$$A(T) = \frac{\left[1/\gamma_1 + \int_0^T \overline{F}(t)G_1(t)dt\right]\left[1 - \int_T^\infty G_2(t)dF(t)\right] + \left[1/\gamma_2 + \int_0^T \overline{F}(t)G_2(t)dt\right]\int_T^\infty G_1(t)dF(t)}{\left[1/\mu_1 + \int_0^T \overline{F}(t)G_1(t)dt\right]\left[1 - \int_T^\infty G_2(t)dF(t)\right] + \left[1/\mu_2 + \int_0^T \overline{F}(t)G_2(t)dt\right]\int_T^\infty G_1(t)dF(t)}. \tag{7.46}$$

When an operating unit undergoes pm upon repair or pm completion, *i.e.*, $T = 0$, the availability is

$$A(0) = \frac{\theta_2/\gamma_1 + (1 - \theta_1)/\gamma_2}{\theta_2/\mu_1 + (1 - \theta_1)/\mu_2}. \tag{7.47}$$

Next, when no pm is done, *i.e.*, $T = \infty$, the availability is

$$A(\infty) = \frac{1/\lambda}{1/\lambda + 1/\mu_1 - 1/\gamma_1}, \tag{7.48}$$

which is equal to the sum of (7.40) and (7.41).

7.3.2 Optimum preventive maintenance policies

Of our interest is to find optimum scheduled times T^* which maximize the mean time $l(T)$ in (7.43) and the availability $A(T)$ in (7.46). It is assumed that $G_1(t) < G_2(t)$ for $0 < t < \infty$. That is, the probability that repair is completed up to time t is less than the probability that pm is completed up to time t. It is also assumed that there exists a density $f(t)$ of $F(t)$. Let $r(t) \equiv f(t)/\overline{F}(t)$ be the failure rate.

We have the following optimum pm policy which maximizes the mean time to system failure:

Theorem 7.3.1 It is assumed that $G_1(t) < G_2(t)$ for $0 < t < \infty$, and the failure rate $r(t)$ is continuous and strictly increasing where $r(\infty) \equiv \lim_{t\to\infty} r(t)$.

(i) If $r(\infty) > M$, $\theta_1\gamma_1 > \theta_2\gamma_2$ and $r(0) < m$, or $r(\infty) > M$ and $\theta_1\gamma_1 \leq \theta_2\gamma_2$, then there exists a finite and unique optimum scheduled time T^* $(0 < T^* < \infty)$, which satisfies

$$r(T)\left[\int_0^T \overline{F}(t)K(t)dt + \int_0^\infty \overline{F}(t)\overline{K}(t)dt\right] - \int_0^T K(t)dF(t) = \frac{\theta_1 \int_0^\infty \overline{K}(t)dF(t) + \theta_2 \int_0^\infty K(t)dF(t)}{\theta_1 - \theta_2}, \tag{7.49}$$

and the resulting mean time is

$$l(T^*) = \frac{1}{\lambda} + \frac{1}{\lambda\theta_1} + \frac{1}{\theta_1}\left[\frac{\int_{T^*}^\infty (\theta_1 + G_1(t))dF(t)}{r(T^*)} - \int_{T^*}^\infty (\theta_1 + G_1(t))\overline{F}(t)dt\right]. \tag{7.50}$$

(ii) If $r(\infty) \leq M$, then the optimum scheduled time is $T^* = \infty$, *i.e.*, no pm is done, and the mean time is given in (7.45).

(iii) If $\theta_1\gamma_1 > \theta_2\gamma_2$ and $r(0) \geq m$, then the optimum scheduled time is $T^* = 0$, *i.e.*, pm is done just upon repair or pm completion, and the mean time is given in (7.44), where

$$K(t) \equiv \frac{\theta_1 G_2(t) - \theta_2 G_1(t)}{\theta_1 - \theta_2},$$

$$m \equiv \frac{\theta_2 \gamma_1 \gamma_2}{\theta_1 \gamma_1 - \theta_2 \gamma_2},$$

$$M \equiv \frac{\lambda \theta_1}{\theta_1 - \theta_2}.$$

Proof. First note that $K(t) > 0$ for $0 < t < \infty$ and

$$\int_0^\infty \overline{F}(t)\overline{K}(t)dt = \frac{\theta_1/\gamma_2 - \theta_2/\gamma_1}{\theta_1 - \theta_2}.$$

It is further noted that $\theta_1 > \theta_2$ and $\gamma_2 > \gamma_1$ from the assumption $G_1(t) < G_2(t)$ for $0 < t < \infty$.

To find an optimum time T^* which maximizes $A(T)$ in (7.46), differentiating $A(T)$ with respect to T and setting it equal to zero, we have

$$r(T)\left\{\int_0^T [\theta_1 G_2(t) - \theta_2 G_1(t)]\overline{F}(t)dt + \theta_1/\gamma_2 - \theta_2/\gamma_1\right\}$$

$$-\int_0^T [\theta_1 G_2(t) - \theta_2 G_1(t)]dF(t) = \int_0^\infty [\theta_1 \overline{G_2}(t) + \theta_2 G_1(t)]dF(t). \quad (7.51)$$

Accepting the definition of $K(t)$ and adjusting both side of (7.51), we have (7.49).

Define that

$$L(T) \equiv r(T)\left[\int_0^T \overline{F}(t)K(t)dt + \int_0^\infty \overline{F}(t)\overline{K}(t)dt\right] - \int_0^T K(t)dF(t),$$

then, we easily have

$$L(0) = r(0)\int_0^\infty \overline{F}(t)\overline{K}(t)dt,$$

$$L(\infty) = \frac{r(\infty)}{\lambda} - \int_0^\infty K(t)dF(t),$$

$$L'(T) = r'(T)\left[\int_0^T \overline{F}(t)K(t)dt + \int_0^\infty \overline{F}(t)\overline{K}(t)dt\right].$$

Therefore, if $\theta_1\gamma_1 > \theta_2\gamma_2$, then $L(T)$ is continuous, positive for $T > 0$, and is strictly increasing since $r(t)$ is continuous and strictly increasing. Next, accepting the definition that

$$D \equiv \frac{\theta_1 \int_0^\infty \overline{K}(t)dF(t) + \theta_2 \int_0^\infty K(t)dF(t)}{\theta_1 - \theta_2}.$$

Then, $D > 0$, and if further $r(0) < m$ and $r(\infty) > M$, then $L(0) < D < L(\infty)$. From the monotonicity and continuity of $L(T)$, there exists a finite

and unique T^* $(0 < T^* < \infty)$ which satisfies (7.51). In this case, the resulting mean time is easily given in (7.50).

If $r(0) \geq m$, then $L(0) \geq D$. Thus, $l'(T) < 0$ for any positive T, which implies that the optimum time is $T^* = 0$, since $l(T)$ is a strictly decreasing function of T.

If $r(\infty) \leq M$, then $L(\infty) \leq D$, *i.e.*, $L(T) < D$ for any finite T, which implies that the optimum time is $T^* = \infty$, *i.e.*, no pm is done.

On the other hand, if $\theta_1\gamma_1 < \theta_2\gamma_2$, then $L(0) < 0$, $L'(0) < 0$ and $L'(\infty) > 0$. Further, it is easily seen that there exists a unique solution of $L'(T) = 0$ for $0 < T < \infty$. Thus, $L(T)$ is a unimodal function and strictly increasing in the interval between the solution of $L'(T) = 0$ and infinity. If $\theta_1\gamma_1 = \theta_2\gamma_2$, then $L(0) = 0$ and $L(T)$ is strictly increasing. In either case, if $r(\infty) > M$, there exists a finite and unique T^* $(0 < T^* < \infty)$ which satisfies (7.49). If $r(\infty) \leq M$, then $L(T) < D$ for any finite T, and hence, the optimum time is $T^* = \infty$. ■

Next, we seek an optimum pm time T^* which maximizes the availability $A(T)$ in (7.46), and have the optimum pm policy:

Theorem 7.3.2 It is assumed that $G_1(t) < G_2(t)$ for $0 < t < \infty$, and the failure rate is continuous and strictly increasing.

(i) If $r(\infty) > M_2$, $\beta_1\mu_1 > \beta_2\mu_2$ and $r(0) < m_2$, or $r(\infty) > M_2$ and $\beta_1\mu_1 \leq \beta_2\mu_2$, then there exists a finite and unique optimum scheduled time T^* $(0 < T^* < \infty)$ which satisfies

$$r(T)\left[\int_0^T \overline{F}(t)K_2(t)dt + \int_0^\infty \overline{K_2}(t)dt\right] - \int_0^T K_2(t)dF(t) = \frac{\beta_1 \int_0^\infty \overline{K_2}(t)dF(t) + \beta_2 \int_0^\infty K_2(t)dF(t)}{\beta_1 - \beta_2}, \tag{7.52}$$

and the resulting availability is

$$A(T^*) = \frac{1/\lambda + \int_{T^*}^\infty G_1(t)dF(t)/r(T^*) - \int_{T^*}^\infty \overline{F}(t)G_1(t)dt}{1/\mu_1 - 1/\gamma_1 + 1/\lambda + \int_{T^*}^\infty G_1(t)dF(t)/r(T^*) - \int_{T^*}^\infty \overline{F}(t)G_1(t)dt}. \tag{7.53}$$

(ii) If $r(\infty) \leq M_2$, then the optimum scheduled time is $T^* = \infty$, *i.e.*, no pm is done, and the availability is given in (7.48).

(iii) If $\beta_1\mu_1 > \beta_2\mu_2$ and $r(0) \geq m_2$, then the optimum scheduled time is $T^* = 0$, and the availability is given in (7.47), where

$$\beta_i \equiv \int_0^\infty F(t)\overline{G_i}(t)dt, \qquad i = 1, 2,$$

$$K_2(t) \equiv \frac{\beta_1 G_2(t) - \beta_2 G_1(t)}{\beta_1 - \beta_2},$$

$$m_2 \equiv \frac{\beta_1\theta_2 + \beta_2(1-\theta_1)}{\beta_1/\mu_2 - \beta_2/\mu_1},$$

$$M_2 \equiv \frac{\lambda\beta_1}{\beta_1 - \beta_2}.$$

Proof. Differentiating $A(T)$ in (7.46) with respect to T and setting it equal to zero, we have

$$r(T)\left\{\int_0^T [\beta_1 G_2(t) - \beta_2 G_1(t)]\overline{F}(t)dt + \beta_1/\mu_2 - \beta_2/\mu_1\right\}$$

$$-\int_0^T [\beta_1 G_2(t) - \beta_2 G_1(t)]dF(t) = \int_0^\infty [\beta_1\overline{G_2}(t) + \beta_2 G_1(t)]dF(t), \qquad (7.54)$$

which implies (7.52) after some calculations using $K_2(t)$. In a similar method of proving Theorem 7.3.1, we can prove this theorem. ■

It has been shown that the problem of maximizing the availability is formally coincident with that of minimizing the expected cost [32].

7.3.3 Replacement of a two-unit parallel system

We consider the following three age replacement policies for a two-unit parallel system (see Section 7.4.1), where it is replaced at (1) time T or both failures of two units, whichever occurs first, (2) time T or failure of either of the two units, whichever occurs first, and (3) time T or failure of either of the two units, whichever occurs last. In case (3), the system is also replaced when both units have failed before time T.

The two units are mutually independent and have an identical failure time distribution $F(t)$. We consider one cycle from $t = 0$ to the replacement: let c_i $(i = 0, 1, 2)$ be the replacement cost when i units have failed. Then, the expected cost per one cycle is, for case (1),

$$C_1(T) = \frac{c_2[F(T)]^2 + 2c_1F(T)\overline{F}(T) + c_0[\overline{F}(T)]^2}{\int_0^T \{1 - [F(t)]^2\}dt}, \qquad (7.55)$$

for case (2),

$$C_2(T) = \frac{c_1\{1 - [\overline{F}(T)]^2\} + c_0[\overline{F}(T)]^2}{\int_0^T [\overline{F}(t)]^2 dt}, \qquad (7.56)$$

and for case (3),

$$C_3(T) = \frac{c_2[F(T)]^2 + c_1\{1 - [F(T)]^2\}}{\int_0^T \{1 - [F(t)]^2\}dt + \int_T^\infty [\overline{F}(t)]^2 dt}. \tag{7.57}$$

We can discuss optimum replacement times which minimize the expected costs $C_i(T)$ $(i = 1, 2, 3)$. Further, if costs c_i are given, we compare three expected costs and decide which policy is the best among three models.

7.4 Other Two-Unit Systems

We show other typical two-unit redundant systems; (1) two-unit parallel system, (2) two-unit priority standby system, and (3) two-unit standby system with imperfect switchover. We are interested in the following three reliability measures for each model: (i) the distribution of the time to system failure, (ii) the expected number of occurrences of system failure during $(0, t]$, and (iii) the pointwise unavailability at time t. To analyze the systems, we use a unique modification of regeneration point techniques in Markov renewal processes. The detained derivations are omitted and the results are shown directly.

7.4.1 Two-unit parallel system

Consider a system of two identical units, where the system can perform its functioning by one of two units: two identical units begin to operate at time 0. If one of two operating units fails, the system continues to operate only on the other unit, and a repair of the failed unit is done. A system failure occurs when both units are down simultaneously. Several contributions to such a two-unit parallel system have been made: Epstein and Hosford [1] discussed the system with both exponential failure and repair times. Gaver [2] extended the model with exponential failure and general repair times, and derived the steady-state availability by using the supplementary variable techniques. Gnedenko [7] proved two limiting theorems under the assumption that the repair time is significantly less than the time to system failure.

In this section, we consider a two-unit parallel system with exponential failure and general repair times, which includes a standby system. It would be impossible to analyze the system under the assumption that all distributions are arbitrary, since there is no regeneration point at which the system makes a transition into any state.

Assume that an operating unit has an exponential distribution and a failed unit has a general distribution $G(t)$ with finite mean $1/\mu$. A failed unit recovers its functioning upon repair completion and begins to operate again immediately. If both units are down simultaneously, there will be a

queue until a repairman is free. This situation means that a system failure has occurred. Further, from the exponential assumption of failure times, the probability that either of two units fails during $(t, t + \triangle t]$, given that two units are operating at time t, is $\lambda_0 \triangle t + o(\triangle t)$, and the probability that one unit fails during $(t, t + \triangle t]$, given that one unit is operating at time t, is $\lambda_1 \triangle t + o(\triangle t)$.

Two units begin to operate at time 0, and it can be said that a system failure occurs when both units are down simultaneously. Then, the LS transform of the distribution of the time to system failure is

$$h(s) = \frac{\lambda_0 \lambda_1}{s + \lambda_1} \frac{1 - g(s + \lambda_1)}{s + \lambda_0[1 - g(s + \lambda_1)]}, \tag{7.58}$$

and its mean time is

$$l = \frac{1}{\lambda_1} + \frac{1}{\lambda_0[1 - g(\lambda_1)]}. \tag{7.59}$$

The LS transform of the expected number of system failures during $(0, t]$ is

$$m(s) = \frac{\lambda_0}{s + \lambda_0} \frac{\lambda_1}{s + \lambda_1} \frac{1 - g(s + \lambda_1)}{1 - g(s) + sg(s + \lambda_1)/(s + \lambda_0)}, \tag{7.60}$$

and its expected number of system failures per unit of time in the steady-state, *i.e.*, $M \equiv \lim_{t \to \infty} M(t)/t$, is

$$M = \frac{1 - g(\lambda_1)}{1/\mu + g(\lambda_1)/\lambda_0}. \tag{7.61}$$

Finally, the LS transform of the pointwise unavailability is

$$\overline{a}(s) = \frac{\lambda_0}{s + \lambda_0} \frac{1 - g(s) - s[1 - g(s + \lambda_1)]/(s + \lambda_1)}{1 - g(s) + sg(s + \lambda_1)/(s + \lambda_0)}, \tag{7.62}$$

and the steady-state unavailability is

$$\overline{A} = \frac{1/\mu - [1 - g(\lambda_1)]/\lambda_1}{1/\mu + g(\lambda_1)/\lambda_0}. \tag{7.63}$$

The model discussed so far includes several interesting redundant models as special cases, as follows:

(1) $\lambda_0 = \lambda_1 = \lambda$

The model corresponds to a two-unit standby system with exponential failure and repair times, which was discussed in Section 7.2 when $F(t) = 1 - e^{-\lambda t}$.

The mean time to system failure is

$$l = \frac{1}{\lambda} + \frac{1}{\lambda[1 - g(\lambda)]}, \tag{7.64}$$

the expected number of system failures is

$$M = \frac{1 - g(\lambda)}{1/\mu + g(\lambda)/\lambda}, \tag{7.65}$$

and the unavailability is

$$\overline{A} = \frac{1/\mu - [1 - g(\lambda)]/\lambda}{1/\mu + g(\lambda)/\lambda}. \tag{7.66}$$

(2) $\lambda_0 = 2\lambda$ **and** $\lambda_1 = \lambda$
The model corresponds to a two-unit parallel system with identical units. Gaver [2] obtained the LS transform of the distribution of the time to system failure and its mean time. He further obtained the steady-state availability by using the supplementary variable technique.
(3) $\lambda_0 = \lambda + \lambda'$ **and** $\lambda_1 = \lambda$
The model corresponds to a two-unit standby system with standby failure, *i.e.*, each unit has the failure rate λ when it is operating and the failure rate λ' when it is in standby. Gnedenko *et al.* [60] obtained the LS transform of the distribution of the time to system failure and its mean time, and studied their asymptotic behaviors.
(4) $\lambda_0 = n\lambda$ **and** $\lambda_1 = (n-1)\lambda$ $(n \geq 2)$
The model is a *2-out-of-n* system $(n \geq 2)$, which is composed of n parallel units and a single repairman. A system failure occurs when 2 out of n units are down simultaneously. Downton [71] discussed, in general, a *k-out-of-n* system $(n \geq k)$. Further, Ramanarayanan [72], Kumar [73], Linton [74], Schneeweiss [75] and Bruning [76] considered the modified models of two-unit systems which form a part of *k-out-of-n* systems.

7.4.2 Two-unit priority standby system

Consider a two-unit standby system of a priority and a non-priority units: if the priority unit fails then it undergoes repair immediately even if the non-priority unit is under repair. If repair of the priority unit is completed, then it begins to operate again immediately even if the non-priority unit is operating. If repair of non-priority unit is completed, then it is immediately ready for system operation as standby. Osaki [14] first considered this model, and Buzacott [22] extended it and derived the steady-state availability.

Assume that the time to failure and the time to repair completion of the priority unit have general distributions $F_1(t)$ with finite mean $1/\lambda_1$ and $G_1(t)$ with finite mean $1/\mu_1$, respectively. Also assume that the time to failure and the time to repair completion of the non-priority unit have the respective exponential distributions $(1 - e^{-\lambda_2 t})$ and $(1 - e^{-\mu_2 t})$. The other assumptions are the same as described in Section 7.2.

It can be said that a system failure occurs when both units are down simultaneously. Then, the LS transform of the distribution of the time to system failure is

$$h(s) = \frac{f_1(s)[\lambda_2/(s+\lambda_2)][1 - g_1(s+\lambda_2)]}{1 - f_1(s)g_1(s+\lambda_2)}, \tag{7.67}$$

and its mean time is

$$l = \frac{1}{\lambda_2} + \frac{1}{\lambda_1[1 - g_1(\lambda_2)]}. \tag{7.68}$$

The LS transform of the expected number of system failures during $(0, t]$ is

$$m(s) = \frac{f_1(s)\{[\lambda_2/(s+\lambda_2)][1 - g_1(s+\lambda_2)][1 - f_1(s+\mu_2)g_1(s)] + f_1(s+\mu_2)[g_1(s) - g_1(s+\lambda_2)]\}}{[1 - f_1(s)g_1(s)][1 - f_1(s+\mu_2)g_1(s+\lambda_2)]}, \tag{7.69}$$

and its expected number in the steady-state is

$$M = \frac{1 - g_1(\lambda_2)}{(1/\lambda_1 + 1/\mu_1)[1 - f_1(\mu_2)g_1(\lambda_2)]}. \tag{7.70}$$

The LS transform of the pointwise unavailability is

$$\overline{a}(s) = \frac{f_1(s)[1 - g_1(s)]}{1 - f_1(s)g_1(s)} - \frac{f_1(s)[s/(s+\lambda_2)][1 - g_1(s+\lambda_2)][1 - f_1(s+\mu_2)g_1(s)]}{[1 - f_1(s)g_1(s)][1 - f_1(s+\mu_2)g_1(s+\lambda_2)]}, \tag{7.71}$$

and its steady-state unavailability is

$$\overline{A} = \frac{1/\mu_1}{1/\lambda_1 + 1/\mu_1} - \frac{[1 - f_1(\mu_2)][1 - g_1(\lambda_2)]}{\lambda_2(1/\lambda_1 + 1/\mu_1)[1 - f_1(\mu_2)g_1(\lambda_2)]}. \tag{7.72}$$

Further, Dimitrov [77] considered the model where the non-priority unit fails when its total operating time exceeds a specified time ζ, *i.e.*, a system failure occurs when the total repair time of the priority unit is ζ. Mine and Kawai [78] and Singh *et al.* [79] considered two-unit systems with priority repair.

7.4.3 Two-unit standby system with imperfect switchover

In the preceding sections, we have assumed that each switchover is perfect. However, it is true that the switchover device (switch) has an important role for a two-unit standby system. In this section, we discuss a two-unit standby system taking account of failure of the switch.

Suppose that the switch assumes up and down states repeatedly, which is independent of the behavior of failure and repair times of each unit. The switch can work only when a unit is changed from standby state to operating state.

Consider a two-unit standby system in Section 7.2 taking account of imperfect switchover, *i.e.*, the failure and repair times of each unit have general distributions $F(t)$ and $G(t)$ with means $1/\lambda$ and $1/\mu$, respectively. The switch assumes up and down states repeatedly, with the two exponential distributions having means of $1/\lambda_s$ and $1/\mu_s$, respectively. That is, the probability that the switch is up at time t, given that it was up at time 0, is

$$P(t) = \frac{\mu_s}{\lambda_s + \mu_s} + \frac{\lambda_s}{\lambda_s + \mu_s} e^{-(\lambda_s + \mu_s)t}, \tag{7.73}$$

(see, *e.g.*, Barlow and Proschan [80], p. 78).

The switch can be used only when a unit is changed from standby state to operating state. Thus, if the switch is down (or faulty) when it is needed, the system is down until it becomes up.

Then, the LS transform of the distribution of the time to system failure, starting that one unit begins to operate, the other unit is in standby and the switch is up, is

$$h(s) = f(s) - \frac{[1 - f(s)] \int_0^\infty e^{-st} P(t) dF(t)}{1 - \int_0^\infty e^{-st} P(t) G(t) dF(t)}, \tag{7.74}$$

and its mean time is

$$l = \frac{1}{\lambda} \left[1 + \frac{\int_0^\infty P(t) dF(t)}{1 - \int_0^\infty P(t) G(t) dF(t)} \right]. \tag{7.75}$$

The LS transform of the expected number of system failures during $(0, t]$ is

$$\begin{aligned} m(s) = {} & \int_0^\infty e^{-st} \overline{P}(t) dF(t) \\ & + \frac{\{f(s) - [s/(s+\mu_s)] \int_0^\infty e^{-st} \overline{P}(t) dF(t)\} \times \left[f(s) - \int_0^\infty e^{-st} P(t) G(t) dF(t) \right]}{1 - \int_0^\infty e^{-st} P(t) G(t) dF(t) - \int_0^\infty e^{-st} F(t) dG(t) - [\mu_s/(s+\mu_s)] \int_0^\infty e^{-st} \overline{P}(t) G(t) dF(t)}, \end{aligned} \tag{7.76}$$

and its expected number in the steady-state is

$$M = \frac{1 - \int_0^\infty P(t) G(t) dF(t)}{1/\lambda + 1/\mu - 1/\gamma + (1/\mu_s) \int_0^\infty \overline{P}(t) G(t) dF(t)}, \tag{7.77}$$

where $\overline{P} \equiv 1 - P$ and $1/\gamma \equiv \int_0^\infty \overline{F}(t) \overline{G}(t) dt$.

The LS transform of the pointwise availability is

$$a(s) = 1 - f(s) + \frac{[1-f(s)]\left\{\int_0^\infty e^{-st}P(t)dF(t) + [\mu_s/(s+\mu_s)]\int_0^\infty e^{-st}\overline{P}(t)dF(t)\right\}}{1 - \int_0^\infty e^{-st}P(t)G(t)dF(t) - \int_0^\infty e^{-st}F(t)dG(t) - [\mu_s/(s+\mu_s)]\int_0^\infty e^{-st}\overline{P}(t)G(t)dF(t)}, \tag{7.78}$$

and its steady-state availability is

$$A = \frac{1/\lambda}{1/\lambda + 1/\mu - 1/\gamma + (1/\mu_s)\int_0^\infty \overline{P}(t)G(t)dF(t)}. \tag{7.79}$$

The results obtained here are coincident with those in Section 7.2 when $P(t) = 1$ for $t \geq 0$. Osaki [16] obtained the distribution of the time to system failure when $F(t) = 1 - e^{-\lambda t}$. Further, Garkavi and Gogolovskiy [19] and Nakagawa and Osaki [30] considered the modified model where the failure of the switch can be detected only when it is used. Nakagawa [81] considered two types of switching failures.

Omar and Nasr [82], Srinivasan [9], Osaki [12] and Kalpakam and Shahul-Hameed [83] considered a two-unit standby system with non-instantaneous switchover. That is, the time required for bringing the standby unit into the operating state, which is called the switchover time, is introduced.

Mazumdar [21] considered a two-unit standby system in which failures are revealed by inspections. That is, a failure in the standby state is revealed only by a specific test which is made at periodic intervals. After that, Nielsen and Runge [84], Gopalan [85], Gupta [86], Singh [87] and Veklerov [88] suggested the modified models with several switching failures and analyzed them.

7.4.4 Other models

1) Two-unit parallel fuel charging system

Buzacott [89] discussed a fuel charge/discharge system for a nuclear reactor, which consists of two fuel-charging machines operating in parallel. When both machines have been down for more than T, the reactor becomes subcritical and shuts down spontaneously. Calabro [90] called such time T *allowed down time*. This model is an extension of a two-unit parallel system discussed in Section 7.4.1. Nakagawa and Osaki [91] derived the distribution of the time to the first reactor shutdown and the pointwise unavailability.

2) Two-unit parallel system with bivariate exponential failure law

Consider a two-unit parallel system with dissimilar units whose failure law is the bivariate exponential distribution. Let X_1 and X_2 denote the lifetimes

of units 1 and 2, respectively. Then, the joint survival distribution of X_1 and X_2 is

$$\Pr\{X_1 > x_1 \text{ and } X_2 > x_2\} = exp\{-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_{12} max(x_1, x_2)\}, \tag{7.80}$$

where the failure law of only unit i $(i = 1, 2)$ is exponential with rate $\lambda_i + \lambda_{12}$ and the simultaneous failure law of both units is so with rate λ_{12}. Harris [20] discussed such a system and obtained the distribution of the time to system failure. Okumoto [92] and Pijnenburg *et al.*[93] derived the steady-state availability.

Further, Dhillon [94] considered a parallel system with common-cause failures. Nahman and Mijušković [95], Lešanovský [96] and Alsammarae [97] analyzed the reliability of two transmission lines with common-cause failures. Murthy and Nguyen [98] considered a two-unit system with failure interaction and obtained the expected cost.

3) Intermittently used system

Consider an intermittently used system which is alternately operative and inoperative, but is used intermittently. Gaver [2] introduced the so-called *disappointment time*, which is the times to system failure during a usage period, or to occurrence of a need during system inoperative period, whichever occurs first. Osaki [99] and Nakagawa *et al.* [100] obtained the distribution of the disappointment time. Further, Srinivasan [5] and Kapil *et al.* [101] discussed a two-unit standby system which is intermittently used.

References

1. Epstein, B. and Hosford, J. (1960), "Reliability of some two unit redundant systems," *Proceedings* 6^{th} *National Symposium in Reliability and Quality Control*, 469-476
2. Jr. Gaver, D. P. (1963), "Time to failure and availability of paralleled systems with repair," *IEEE Transactions on Reliability*, **R-12**, 30-38
3. Jr. Gaver, D. P. (1964), "Failure time for a redundant repairable system of two dissimilar elements," *IEEE Transactions on Reliability*, **R-13**, 14-22
4. Gnedenko, B. V. (1964), "Idle duplication," *Engineering Cybernetics*, **2**, 1-9
5. Srinivasan, V. S. (1966), "The effect of standby redundancy in system's failure with repair maintenance," *Operations Research*, **14**, 1024-1036
6. Liebowitz, B. H. (1966), "Reliability considerations for a two element redundant system with generalized repair times," *Operations Research*, **14**, 233-246
7. Gnedenko, B. V. (1964), "Duplication with repair," *Engineering Cybernetics*, **2**, 102-108
8. Solovyev, A. D. (1964), "Asymptotic distribution of lifetime of a duplicated element," *Engineering Cybernetics*, **2**, 109-111
9. Srinivasan, V. S. (1968), "A standby redundant model with noninstanteneous switchover," *IEEE Transactions on Reliability*, **R-17**, 175-178
10. Mine, H. and Osaki, S. (1969), "On failure time distributions for systems of dissimilar units," *IEEE Transactions on Reliability*, **R-18**, 165-16
11. Osaki, S. (1970), "A note on a two-unit standby redundant system ," *Journal of the Operations Research Society of Japan*, **12**, 43-51
12. Osaki, S. (1970),"System reliability analysis by Markov renewal processes," *Journal of the Operations Research Society of Japan*, **12**, 127-188
13. Osaki, S. (1970),"Renewal theoretic aspects of two-unit redundant systems," *IEEE Transactions on Reliability*, **R-19**, 105-110
14. Osaki, S. (1970), "Reliability analysis of a two-unit standby redundant system with priority," *Canadian Operational Research Society Journal*, **8**, 60-62
15. Osaki, S. (1971), "A note on a two-unit standby-redundant system with imperfect switchover," *Revue Française d'Informatique et de Recherche Opérationnelle*, **2**, 103-109
16. Osaki, S. (1972), "On a two-unit standby-redundant system with imperfect switchover," *IEEE Transactions on Reliability*, **R-21**, 20-24
17. Osaki, S. (1972), "Reliability analysis of a two-unit standby-redundant system with preventive maintenance," *IEEE Transactions on Reliability*, **R-21**, 24-29
18. Osaki, S. and Asakura, T. (1970), "A two-unit standby redundant system with preventive maintenance ," *Journal of Applied Probability*, **7**, 641-648
19. Garkavi, A. L. and Gogolovskiy, V. B. (1963), "Calculation of the mean time of reliable operation in duplicated equipment with automatic switching and reserve repair," *Engineering Cybernetics*, **1**, 23-32

20. Harris, R. (1968), "Reliability applications of a bivariate exponential distribution," *Operations Research*, **16**, 18-27
21. Mazumdar, M. (1970), "Reliability of two-unit redundant repairable systems when failures are revealed by inspections," *SIAM Journal of Applied Mathematics*, **19**, 637-647
22. Buzacott, J. A. (1971), "Availability of priority standby redundant systems," *IEEE Transactions on Reliability*, **R-20**, 60-63
23. Linton, D. G. and Braswell, R. N. (1973), "Laplace transforms for the two-unit cold-standby redundant system," *IEEE Transactions on Reliability*, **R-22**, 105-108
24. Osaki, S. and Nakagawa, T. (1971), "On a two-unit standby redundant system with standby failure," *Operations Research*, **19**, 510-523
25. Nakagawa, T. and Osaki, S. (1974), "Stochastic behaviour of a two-unit standby redundant system," *INFOR*, **12**, 66-70
26. Nakagawa, T. and Osaki, S. (1974), "Optimum preventive maintenance policies for a 2-unit redundant system," *IEEE Transactions on Reliability*, **R-23**, 86-91
27. Nakagawa, T. and Osaki, S. (1974), "Stochastic behavior of a two-dissimilar-unit standby redundant system with repair maintenance," *Microelectronics and Reliability*, **13**, 143-148
28. Nakagawa, T. and Osaki, S. (1974), "Optimum preventive maintenance policies maximizing the mean time to the first system failure for a two-unit standby redundant system," *Journal of Optimiziation Theory and Applications*, **14**, 115-129
29. Nakagawa, T. and Osaki, S. (1975), "Stochastic behaviour of a two-unit priority standby redundant system with repair," *Microelectronics and Reliability*, **14**, 309-313
30. Nakagawa, T. and Osaki, S. (1975), "Stochastic behavior of 2-unit standby redundant system with imperfect switchover," *IEEE Transactions on Reliability*, **R-24**, 143-146
31. Nakagawa, T. and Osaki, S. (1975), "Stochastic behavior of two-unit paralleled redundant system with repair maintenance," *Microelectronics and Reliability*, **14**, 457-461
32. Nakagawa, T. and Osaki, S. (1976), "A summary of optimum preventive maintenance policies for a two-unit standby redundant system," *Zeitshrift für Operations Research*, **20**, 171-187
33. Yamada, S. and Osaki, S. (1980), "Reliability evaluation of a 2-unit unrepairable system," *Microelectronics and Reliability*, **20**, 589-597
34. Mine, H. and Nakagawa, T. (1976), "Stochastic behavior of two-unit redundant systems which operate at discrete times," *Microelectronics and Reliability*, **15**, 551-554
35. Kodama, M. and Deguchi, H. (1974), "Reliability considerations for a 2-unit redundant system with erlang-failure and general repair distributions," *IEEE Transactions on Reliability*, **R-23**, 75-81
36. Kodama, M., Nakamichi, H. and Takamatsu, S. (1976),"Analysis of 7 models for the 2-dissimilar-unit, warm standby, redundant system," *IEEE Transactions on Reliability*, **R-25**, 273-280
37. Adachi, K. and Kodama, M. (1980), "Availability analysis of two-unit warm standby system with inspection time," *Microelectronics and Reliability*, **20**, 449-456
38. Ohashi, M. and Nishida, T. (1980), "A two-unit paralleled systems with general distributions," *Journal of the Operations Research Society of Japan*, **23**, 313-325
39. Parathasarathy, P. R. (1979), "Cost analysis for 2-unit systems," *IEEE Transactions on Reliability*, **R-28**, 268-269

40. Wiens, D. (1981), "Analysis of a hot-standby system with 2 identical-dependent units and a general erlang failure time distribution," *IEEE Transactions on Reliability*, **R-30**, 386
41. Jack, N. (1986), "Analysis of a repairable 2-unit parallel redundant system with dependent failures," *IEEE Transactions on Reliability*, **R-35**, 444-446
42. Wells, C. E. (1987), "Reliability measures for a regenerative system viewed over a random horizon," *IEEE Transactions on Reliability*, **R-36**, 124-128
43. Vanderperre, E. J. (1998), "On the reliability of Gaver's parallel system sustained by a cold standby unit and attended by two repairmen," *Journal of the Operations Research Society of Japan*, **41**, 171-180
44. Mine, H. and Kawai, H. (1974), "An optimal maintenance policy for a 2-unit parallel system with degraded states," *IEEE Transactions on Reliability*, **R-23**, 81-86
45. Kumar, A., Chopra, M. G. and Kapoor, V. B. (1981), "A computer algorithm for optimal maintenance of standby redundant systems," *IEEE Transactions on Reliability*, **R-30**, 28-29
46. Nakagawa, T. and Osaki, S. (1979), "Bibliography for reliability and availability of stochastic systems," *IEEE Transactions on Reliability*, **R-25**, 284-287
47. Kumar, A. and Agarwal, M. (1980), "A review of standby redundant systems," *IEEE Transactions on Reliability*, **R-29**, 290-294
48. Yearout, R. D., Reddy, P. and Grosh, D. L. (1986),"Standby redundancy in reliability-A review," *IEEE Transactions on Reliability*, **R-35**, 285-292
49. Laprie, J. C., Costes, A. and Landrault, C. (1981), "Parametric analysis of 2-unit redundant computer systems with corrective and preventive maintenance," *IEEE Transactions on Reliability*, **R-30**, 139-144
50. De, B. B. and Krakan, H. B. (1981), "Fault-tolerance in a multiprocessor, digital switching system," *IEEE Transactions on Reliability*, **R-30**, 246-252
51. Trivedi, K. S. and Geist, R. M. (1983), "Decomposition in reliability analysis of fault-tolerant systems," *IEEE Transactions on Reliability*, **R-32**, 463-468
52. Ng, S. W. (1986), "Reliability & availability of duplex systems: Some simple models," *IEEE Transactions on Reliability*, **R-35**, 295-300
53. Hu, M. and Mouftah, H. T. (1987), "Fault-tolerant system using 3-value logic circuits," *IEEE Transactions on Reliability*, **R-36**, 227-231
54. Ibe, O. C., Howe, R. C. and Trivedi, K. S. (1989), "Approximate availability analysis of VAXcluster systems," *IEEE Transactions on Reliability*, **38**, 146-152
55. Walker, B. K., Wereley, N. M., Luppold, R. H. and Gai, E. (1989), "Effects of redundancy management on reliability modeling," *IEEE Transactions on Reliability*, **38**, 475-481
56. Reibman, A. L. (1990), "Modeling the effect of reliability on performance," *IEEE Transactions on Reliability*, **39**, 314-320
57. Choi, C. Y., Johnson, B. W. and Profeta V, J. A. (1997), "Safety issues in the comparative analysis of dependable architectures," *IEEE Transactions on Reliability*, **46**, 314-320
58. Gai, E., Harrison, J. V. and Luppold, R. H. (1983), "Reliability analysis of a dual-redundant engine controller," *IEEE Transactions on Reliability*, **R-32**, 14-20
59. Tapiero, C. S. and Hsu, L. F. (1986), "Randomized quality control of a 2-station machining process with blocking," *IEEE Transactions on Reliability*, **R-35**, 455-458
60. Gnedenko, B. V., Belyayev, Y. K. and Solovyev, A. D. (1969), Mathematical Methods of Reliability Theory, Academic Press, New York
61. Pyke, R. (1961), "Markov renewal processes: Definitions and preliminary properties ," *Annals of Mathematical Statistics*, **32**, 1231-1242

62. Pyke, R. (1961), "Markov renewal processes with finitely many states ," *Annals of Mathematical Statistics*, **32**, 1243-1259
63. Smith, W. L. (1958), "Renewal theory and its ramifications," *Journal Royal Statistics Society, Series B,* **20**, 243-302
64. Rozhdestvenskiy, D. V. and Fanarzhi, G. N. (1970), "Reliability of a duplicated system with renewal and preventive maintenance," *Engineering Cybernetics*, **8**, 475-479
65. Berg, M. (1976), "Optimal replacement policies for two-unit machines with increasing running cost I," *Stochastic Processes and their Applications*, **4**, 89-106
66. Berg, M. (1977), "Optimal replacement policies for two-unit machines with running costs: II," *Stochastic Processes and their Applications*, **5**, 315-322
67. Berg, M. (1978), "General trigger-off replacement procedures for two-unit system," *Naval Research Logistics Quarterly*, **25**, 15-29
68. Teixeira de Almeida, A. and Campello de Souza, F.M. (1993), "Decision theory in maintenance strategy for a 2-unit redundant standby system," *IEEE Transactions on Reliability*, **R-42**, 401-407
69. Gupta, P. P. and Kumar, A. (1981), "Operational availability of a complex system with two types of failure under different repair preemptions," *IEEE Transactions on Reliability*, **R-30**, 484-485
70. Pullen, K. W. and Thomas, M. U. (1986), "Evaluation of an opportunistic replacement policy for a 2-unit system," *IEEE Transactions on Reliability*, **R-35**, 320-324
71. Downton, F. (1966), "The reliability of multiplex systems with repair," *Journal of Royal Statistical Society, Series B*, **28**, 459-476
72. Ramanarayanan, R. (1976), "Availability of the 2-out-of-n:F system," *IEEE Transactions on Reliability*, **R-25**, 43-44
73. Kumar, A. (1977), "Steady-state profit in several 1-out-of-2:G systems," *IEEE Transactions on Reliability*, **R-26**, 366-369
74. Linton, D. G. (1981), "Life distributions and degradation for a 2-out-of-n:F system," *IEEE Transactions on Reliability*, **R-30**, 82-84
75. Schneeweiss, W. G. (1995), "Mean time to first failure of repairable systems with one cold spare," *IEEE Transactions on Reliability*, **R-44**, 567-574
76. Bruning, K. L. (1996), "Determining the discrete-time reliability of a repairable 2-out-of-(N+1):F system," *IEEE Transactions on Reliability*, **45**, 150-155
77. Dimitrov, M. TS. (1972), "A limit theorem for a duplicated system with unrenewable redundancy," *Engineering Cybernetics*, **10**, 816-819
78. Mine, H. and Kawai, H. (1979), "Repair priority effect on availability of a 2-unit system," *IEEE Transactions on Reliability*, **R-28**, 325-326
79. Singh, S. K., Singh, R. P. and Shukla, S. (1991), "Cost-benefit analysis of a 2-unit priority-standby system with patience-time for repair," *IEEE Transactions on Reliability*, **R-40**, 11-14
80. Barlow, R. E. and Proschan, F. (1965), Mathematical Theory of Reliability. John Wiley & Sons, New York
81. Nakagawa, T. (1977), "A 2-unit repairable redundant system with switching failure," *IEEE Transactions on Reliability*, **R-26**, 128-130
82. Omar, A. and Nasr, Y. (1966), "Nonloaded duplexing taking switching time into account," *Engineering Cybernetics*, **4**, 310-313
83. Kalpakam, S. and Shahul-Hameed, M. A. (1980), "General 2-unit redundant system with random delays," *IEEE Transactions on Reliability*, **R-29**, 86-87
84. Nielsen, D. and Runge, B. (1974), "Unreliability of a standby system with repair and imperfect switching," *IEEE Transactions on Reliability*, **R-23**, 19-24

85. Gopalan, M. N. (1975), "Availability and reliability of 1-server 2-unit system with imperfect switch," *IEEE Transactions on Reliability*, **R-24**, 218-219
86. Gupta, R. K. (1978), "A standby redundant complex system with imperfect switch," *IEEE Transactions on Reliability*, **R-27**, 298-300
87. Singh, J. (1980), "Effect of switch failure on 2 redundant systems," *IEEE Transactions on Reliability*, **R-29**, 82-83
88. Veklerov, E. (1987), "Reliability of redundant systems with unreliable switches," *IEEE Transactions on Reliability*, **R-36**, 470-472
89. Buzacott, J. A. (1973), "Reliability analysis of a nuclear reactor fuel charging system," *IEEE Transactions on Reliability*, **R-22**, 88-91
90. Calabro, S. R. (1962), Reliability Principles and Practices. McGraw-Hill, New York
91. Nakagawa, T. and Osaki, S. (1975), "Stochastic behavior of a 2-unit parallel fuel charging system," *IEEE Transactions on Reliability*, **R-24**, 302-304
92. Okumoto, K. (1981), "Availability of a 2-component dependent system," *IEEE Transactions on Reliability*, **R-30**, 205
93. Pijnenburg, M., Ravichandran, N. and Regterschot, G. (1993), "Stochastic analysis of a dependent parallel system," *European Journal of Operational Research*, **68**, 90-104
94. Dhillon, B. S. (1981), "Repairable-system models," *IEEE Transactions on Reliability*, **R-30**, 492-493
95. Nahman, J. and Mijušković, N. (1985), "Reliability modeling of multiple overhead transmission lines," *IEEE Transactions on Reliability*, **R-34**, 281-285
96. Lešanovský, A. (1988), "Multistate Markov models for systems with dependent units," *IEEE Transactions on Reliability*, **37**, 505-511
97. Alsammarae, A. J. (1989), "Modeling dependent failures for the availability of extra high voltage transmission lines," *IEEE Transactions on Reliability*, **38**, 236-241
98. Murthy, D. N. P. and Nguyen, D. G. (1985), "Study of two-component system with failure interaction," *Naval Research Logistics Quarterly*, **32**, 239-247
99. Osaki, S. (1972), "An intermittently used system with preventive maintenance," *Journal of the Operations Research Society of Japan*, **15**, 102-111
100. Nakagawa, T., Goel, A. L. and Osaki, S. (1975), "Stochastic behavior of an intermittently used system," *Revue Française d'Informatique et de Recherche Opérationnelle*, **2**, 101-112
101. Kapil, D. V. S., Kapur, P. K. and Kapoor, K. R. (1980), "Intermittently used 2-unit redundant system with PM," *IEEE Transactions on Reliability*, **R-29**, 277-278

8. Optimal Maintenance Problems for Markovian Deteriorating Systems

Hajime Kawai, Junji Koyanagi
Department of Social Systems Engineering,
Tottori University
Tottori 680–0945, Japan
and
Masamitsu Ohnishi
Graduate School of Economics,
Osaka University
Osaka 560-0043, Japan

Summary.

This chapter deals with optimal maintenance problems for Markovian deteriorating systems. The function of the system deteriorates with time, and the grade of deterioration is classified as one of $s+2$ discrete states, $0, 1, \ldots, s, s+1$, in the order of increasing deterioration. State 0 is a good state, *i.e.*, the system is like new, the states $1, \ldots, s$ are deterioration states and the state $s+1$ is a failure state. In a normal operation, these states are assumed to constitute a discrete or continuous time Markovian process with an absorbing state $s+1$. In Section 8.1, we first introduce a basic replacement problem for a discrete time Markovian deteriorating system. In Section 8.2, we discuss an optimal inspection and replacement problem for the system in Section 8.1. In Section 8.3, we consider an optimal inspection and replacement problem under incomplete system observation. In Section 8.4, we treat a continuous time Markovian deteriorating system and discuss an optimal inspection and replacement problem. In Section 8.5, we deal with a maintenance problem in queueing system and discuss an optimal maintenance policy based on both the queue length and the server state.

Keywords: Markovian deteriorating system, Markovian decision process, control limit rule, semi-Markov decision process, partially observable Markov decision process, switch curve structure, totally positive of order 2

8.1 A Basic Optimal Replacement Problem for a Discrete Time Markovian Deteriorating System

The system is observed at each time and then we can choose one of two actions, that is,

action 0: we continue to operate the system without replacement,

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

action 1 : we replace the system with a new one.

When we choose action 0 for the system in state i, then the system moves to state j with probability p_{ij} at the next time and operating cost L_i is incurred. When we choose action 1 for the system in state i, then the system becomes new at the next time and replacement cost C_i is incurred. Here, for simplicity of discussion, we assume that it takes one unit of time for replacement. For such a system, our problem is to choose a policy *i.e.*, to choose an action for each state, which minimizes total discounted expected cost in an infinite time horizon. Such a policy is called an optimal policy. We denote a discount factor by β, $0 < \beta < 1$.

8.1.1 Some conditions on transition probabilities and cost structure

For transition probability p_{ij} and costs L_i, C_i, we introduce the following conditions. In this chapter, the term 'increasing' means 'nondecreasing.'

(C.1) For any k, $\sum_{j=k}^{s+1} p_{ij}$ is increasing in i.
(C.2) L_i and C_i are increasing in i.
(C.3) $L_i - C_i$ are increasing in i.

(C.1) is shown to be equivalent to the following condition (C.4).

(C.4) For any increasing function f_i, $\sum_{j=0}^{s+1} p_{ij} f_j$ is increasing in i.

(C.1) means that as the system deteriorates, it is more likely to make a transition to higher states. (C.2) means that as the system deteriorates, it is more costly to operate or to replace. (C.3) means that the merit of replacement becomes bigger as the system deteriorates.

The Markovian deteriorating system, an optimal replacement problem and the conditions in this section were proposed and studied by Derman [4].

8.1.2 Formulation by Markovian decision process (MDP)

Our problem is formulated by a Markovian Decision Process with state space $\mathcal{S} + 1 = \{0, 1, \ldots, s, s+1\}$ where each number corresponds to the state of the system, action space $\{0, 1\}$ and immediate costs are L_i and C_i. We let v_i denote the total discounted expected cost incurred when the system starts with state i and an optimal policy is employed. Here in after, $\sum_j$ denotes that the sum is taken over all the values which j can take. Then, v_i obeys the following equations:

$$v_i = \min\{L_i + \beta \sum_j p_{ij} v_j,\ C_i + \beta v_0\}, \qquad i \in \mathcal{S} + 1. \tag{8.1}$$

The first term in the parenthesis corresponds to action 0 and the second term, to action 1. Corresponding to the above equation, we consider the function $(n = 0, 1, \ldots)$ defined inductively as

$$\begin{aligned} v_i^0 &= 0, \\ v_i^{n+1} &= \min\{L_i + \beta \sum_j p_{ij} v_j^n,\ C_i + \beta v_0^n\}, \\ & i \in \mathcal{S} + 1, \qquad n = 1, 2, \ldots, . \end{aligned} \tag{8.2}$$

v_i^n is interpreted as the optimal expected cost of the n period version of our problem. By the standard argument of the theory of contraction mapping, it is guaranteed that v_i^n converges to v_i, $i \in \mathcal{S} + 1$.

8.1.3 Optimality of control limit rule

In this section we examine the structure of an optimal replacement policy. We first give the following lemma, which has an important role in our discussion.

Lemma 8.1.1 The optimal cost function v_i is increasing in i.

Proof. The proof is done through a mathematical induction method in n in (8.2). First v_i^1 is increasing in i under (C.2). Assuming that v_i^n is increasing in i, then $\sum_j p_{ij} v_j^n$ is shown to be increasing in i by (C.4). Furthermore, v_i^n converges to v_i. Hence, together with condition (C.2), v_i is increasing. ■

Using this lemma, we have the following theorem with respect to the structure of an optimal policy. We let D_i denote an optimal action at state i.

Theorem 8.1.1 There exists the state K such that

$$D_i = \begin{cases} 0, & \text{for } 0 \le i \le K - 1, \\ 1, & \text{for } K \le i \le s + 1, \end{cases}$$

where $0 \le K \le s + 2$.

Proof. Subtracting the second term corresponding to action 1 from the first term corresponding to action 0 in (8.1), then we have

$$L_i - C_i + \beta \sum_j p_{ij} v_j - \beta v_0.$$

According to the condition (C.3) and (C.4) together with lemma 8.1.1, the above function increases in i. This implies that this theorem holds. ■

Such a type of policy is called the "Control Limit Rule" and state K is called the control limit state.

8.2 An Optimal Inspection and Replacement Problem

In this section, we discuss an optimal inspection and replacement problem for a discrete time Markovian deteriorating system. In the basic optimal replacement problem, it is assumed that the state of the system can be identified at any time without any cost. It is, however, costly to identify the state through inspection. In this section, we take account of such a situation and discuss an optimal inspection and replacement problem. In this problem, we assume that each inspection time is negligibly small and inspection cost A is incurred.

8.2.1 Transition probability

We investigate some properties of the transition probability of the states of the system without replacement. First, we introduce the following conditions instead of the condition (C.1).

(C.5) The system cannot recover its function without replacement, that is,

$$p_{ij} = 0, \text{ for } i > j.$$

(C.6) A function $f(x, y)$ of two variables over linearly ordered sets is said to be totally positive for order 2 (TP_2), if

$$f(x_1, y_1)f(x_2, y_2) \geq f(x_2, y_1)f(x_1, y_2) \quad \text{for any } x_1 \leq x_2,\ y_1 \leq y_2.$$

One-step transition probability p_{ij} is (TP_2), that is

$$p_{im}p_{jn} \geq p_{in}p_{jm} \quad \text{for any } i \leq j,\ m \leq n,\ i, j \in \mathcal{S} + 1.$$

It is easily seen that (C.6) implies (C.1). That is, (C.6) is a stronger condition than (C.1).

If we denote $P_{ij}(t)$ as t-step transition probability from state i to state j, then it has the the following properties. For proof, see Karlin [8].

(P.1) For any k and t, $\sum_{j=k}^{s+1} P_{ij}(t)$ is increasing in i.
(P.2) For any t, $P_{ij}(t)$ is TP_2 in i, j.
(P.3) For any i, $P_{ij}(t)$ is TP_2 in j, t.

Here, we give a so-called variation diminishing property of TP_2 function, which is stated for our transition probability and in our problem.

(P.4) If f_i changes its sign at most once in i and any change that occurs is from negative, then $\sum_j P_{ij}(t)f_j$ changes its sign at most once in t for each i and in i for each t, and if a change occurs, it is from negative. These properties have an important role in our discussion.

8.2.2 Formulation by semi-Markov decision process (SMDP)

Our inspection and replacement problem is formulated by a semi-Markov decision process. First, we let E_i denote the time when the state of the system is identified to be in state i by inspection or replacement. At each E_i, we can choose one of the following actions.

action $I(t)$ $(1 \leq t \leq \infty)$: We do not replace the system, and the next inspection is planned to be made t unit time after, where $I(\infty)$ means that we do neither inspection nor replacement of the system.
action R : We replace the system with a new one.

Our problem is to determine the action at each E_i which minimizes the total discounted expected cost in an infinite time horizon. We let v_i denote the total discounted cost when the system starts with E_i and an optimal policy is adopted thereafter. Furthermore, we let $H_i(t)$ denote the total discounted cost when the system starts from E_i with action $I(t)$ and an optimal policy is adopted after the next inspection time. Then $H_i(t)$ is given by

$$H_i(t) = L_i(t) + \beta^t A + \beta^t \sum_j P_{ij}(t) v_j, \tag{8.3}$$

where $L_i(t)$ is the discounted expected operating cost during time interval t when the system starts from E_i, and is given by,

$$\begin{aligned} L_i(0) &= 0, \\ L_i(t) &= L_i + \beta \sum_j p_{ij} L_j(t-1). \end{aligned} \tag{8.4}$$

We let R_i denote total discounted cost when the system starts from E_i with action R and an optimal policy is adopted after the system becomes new. R_i is given by

$$R_i = C_i + \beta v_0. \tag{8.5}$$

Then, optimal cost v_i obeys the following equation.

$$v_i = \min\{\min_{1 \le t \le \infty} H_i(t),\ R_i\}, \qquad i \in \mathcal{S} + 1. \tag{8.6}$$

Corresponding to the above equation, we consider the following recursive equation.

$$\begin{aligned} v_i^0 &= 0, \\ v_i^{n+1} &= \min\{\min_{1 \le t \le \infty} H_i^{n+1}(t),\ R_i^{n+1}\}, \\ H_i^{n+1}(t) &= L_i(t) + \beta^t A + \beta^t \sum_j P_{ij}(t) v_j^n, \\ R_i^{n+1} &= C_i + \beta v_0^n, \quad i \in \mathcal{S} + 1, \qquad n = 0, 1, 2, \ldots . \end{aligned} \tag{8.7}$$

8.2.3 Structure of optimal inspection and replacement policy

We investigate the structure of an optimal policy. First, we have the following lemmas.

Lemma 8.2.1 Expected operating cost $L_i(t)$ is increasing in i.

Proof. Proof is easily desived from (C.2), (P.1) and (8.8). ■

Lemma 8.2.2 The optimal function v_i is increasing in i.

Proof. Using the recursive (8.7), lemma 8.2.1, and (P.1), proof is done through mathematical induction method in n. ■

The following theorem means that an optimal policy exists in a control limit rule for replacement. We let D_i denote an optimal action at E_i. For example, when we write $D_i = I(t_i)$, then an optimal action at E_i is that we do not replace the system and the next inspection is made t_i unit time after.

Theorem 8.2.1 There exist K such that

$$D_i = \begin{cases} I(t_i), & \text{for } 0 \le i < K-1, \\ R, & \text{for } K \le i \le s+1, \end{cases} \tag{8.8}$$

where $0 \le K \le s+2$.

Proof. Subtracting R_i from $H_i(t)$, we have

$$H_i(t) - R_i = L_i - C_i - \beta v_0 + \beta^t A + \beta \sum_j p_{ij} L_j(t-1) + \beta^t \sum_j P_{ij}(t) v_j. \tag{8.9}$$

From (C.3), (P.1), lemmas 8.2.1 and 8.2.2, it is shown that $H_i(t) - R_i$ is increasing in i, which implies that a control limit rule is optimal. ■

Next, we investigate the optimal inspection time interval t_i, where $i = 0, 1, \ldots, K-1$. Let

$$B_i(t) = H_i(t+1) - H_i(t). \tag{8.10}$$

Using (8.3) and (8.4), we have

$$H_i(t) = L_i + \beta \sum_j p_{ij} H_j(t-1) \tag{8.11}$$

Hence, $B_i(t)$ can be written as

$$B_i(t) = \beta^t \sum_j P_{ij}(t) b_j, \tag{8.12}$$

where

$$b_i = L_i + \beta A + \beta \sum_j p_{ij} v_j - A - v_i. \tag{8.13}$$

For the sign of b_i, we have the following lemma.

Lemma 8.2.3 b_i changes its sign at most once, and when a change occurs it is from negative.

Proof. For $0 \le i \le K-1$, using (8.11), we have

$$v_i = \begin{cases} H_i(1) = L_i + \beta A + \beta \sum_j p_{ij} v_j, & \text{if } t_i = 1, \\ H_i(t_i) = L_i + \beta \sum_j p_{ij} H_j(t_i - 1) & \text{if } t_i \ge 2, \\ \qquad \ge L_i + \beta \sum_j p_{ij} v_j. \end{cases} \tag{8.14}$$

From (8.13) and (8.14), b_i is negative. For $K \le i \le s+1$, since $D_i = R$, we have

$$b_i = -(1-\beta)A + L_i - C_i + \beta \sum_j p_{ij} v_j - \beta v_0.$$

The right hand side increases in i, which means that the sign of b_i changes at most once and when change occurs it is from negative to positive. This completes the proof. ■

For the behavior of $B_i(t)$ in i and t, we have the following lemma.

Lemma 8.2.4 $B_i(t)$ changes its sign in t at most once and when change occurs it is from negative. Moreover, if $B_i(t) \geq 0$, then $B_{i+1}(t) \geq 0$.

Proof. Using properties (P.2) (P.3) (P.4), (8.12) and lemma 8.2.3, the proof is done. ■

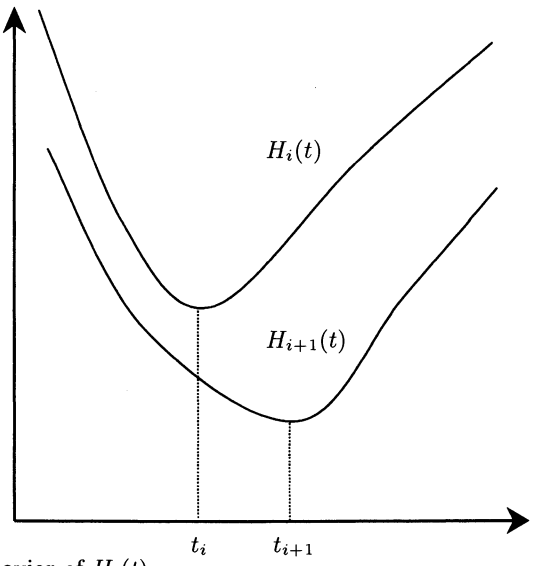


Fig. 8.1. The behavior of $H_i(t)$

From this lemma, we have the following theorem for the structure of an optimal inspection and replacement policy.

Theorem 8.2.2 There exists an optimal policy of the form,

$$D_i = \begin{cases} I(t_i), & 0 \leq i \leq K-1, \\ R, & K \leq i \leq s+1, \end{cases}$$

and $t_0 \geq t_1 \geq \cdots \geq t_{K-1}$.

8.3 An Optimal Inspection and Replacement Policy with Incomplete Information

In this section, we discuss an optimal inspection and replacement problem in circumstances where the system is monitored by some mechanism which gives

some information about the state of the system but does not necessarily tell the true state of the system. We assume that the outcome of the monitoring mechanism is classified into levels $0, 1, \ldots, m$. We let $\mathcal{M}$ denote the set of the outcome $\{0, 1, \ldots, m\}$.

8.3.1 Some notations and conditions

The relation between the true state of the system and the outcome of the monitor is assumed to be described by the following conditional probability.

$$q_{i\theta} = \mathrm{P}\ (\text{the outcome is } \theta \text{ given that the system is in state } i).$$

The cost structure of the system and the notations are same as in the previous section, that is,

L_i : operating cost of the system in state i,
C_i : replacement cost of the system in state i,
A : inspection cost.

Here, we assume that replacement and inspection take one unit of time. For convenience of expression, we define the following notations:

$$P = (p_{ij})_{i,j \in \mathcal{S}+1} : (s+2) \times (s+2) \text{ matrix},$$
$$Q = (q_{i\theta})_{i \in \mathcal{S}+1, \theta \in \mathcal{M}} : (s+2) \times (m+1) \text{ matrix},$$
$$L = (L_0, L_1, \ldots, L_{s+1})^T,$$
$$C = (C_0, C_1, \ldots, C_{s+1})^T,$$
$$F = \{(f_0, f_1, \ldots, f_{s+1})^T : \ f_i \text{ is increasing in } i\}.$$

Our inspection and replacement problem is considered under the following conditions, using the above notations.

(C.2) $L \in F$, $C \in F$
(C.3) $L - C \in F$
(C.5) P is upper triangular.
(C.6) Transition matrix P is TP_2.

These conditions have been introduced in the previous two sections. In addition to the above conditions, we introduce the following condition, which relates the true states of the system and the outcomes of the monitor.

(C.7) Q is TP_2, that is

$$q_{i\theta} q_{j\theta'} - q_{i\theta'} q_{j\theta} \geq 0 \text{ for any } i \leq j,\ \theta \leq \theta'.$$

This condition means that higher deterioration of the system gives higher outcome levels from the monitor probabilistically.

8.3.2 Formulation by partially observable Markov decision process (POMDP)

Our problem is to find the optimal action at each time which minimizes the total discounted expected cost in an infinite time horizon. Appealing to the standard theory of POMDP studied by Eckles [5], Sondik [22] and others, our problem is formulated by a Markov decision process in which the state

space is probability distribution over $\mathcal{S}+1$. We denote the state space in MDP formulation by

$$X = \{x = (x_0, x_1, \ldots, x_{s+1}) : x_i \geq 0,\ i \in \mathcal{S}+1,\ \sum_i x_i = 1\},$$

where x_i is the probability that the system is in state i.

At each time, we can choose one of the following three actions based on the state probability $x \in X$.

action 0: we continue to operate the system with monitoring,
action 1: we inspect the system (this action means that we can exactly catch the true state of the system),
action 2: we replace the system with a new one.

When an action is selected at state x (probability distribution over $\mathcal{S}+1$), we have the following state transition at the next time.

(0) When action 0 (continue to operate) is selected, the probability that the outcome at the next time is θ is

$$P(\theta|x) = \sum_i \sum_j x_i p_{ij} q_{j\theta},$$

which is the $(\theta+1)$-th component of xPQ. If the outcome is θ, then the probability that the system is in state j is given by

$$T_j(x,\theta) = \frac{\sum_i x_i p_{ij} q_{j\theta}}{\sum_i \sum_j x_i p_{ij} q_{j\theta}}, \tag{8.15}$$

which is derived by using Bayes' formula. We let

$$T(x,\theta) = (T_0(x,\theta), T_1(x,\theta), \ldots, T_{s+1}(x,\theta)),$$

which is the next state in MDP formulation.

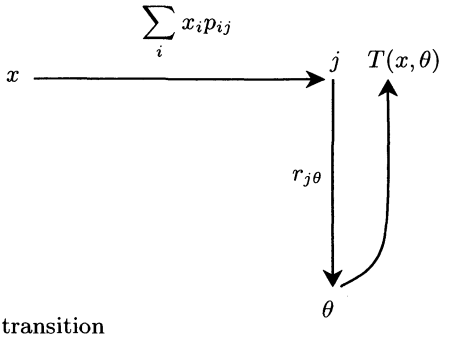


Fig. 8.2. State transition

(1) When action 1 is selected, the system is known to be in state j with probability $(xP)_j$, which is the $(j+1)$-th component of xP, that is

$$(xP)_j = \sum_i x_i p_{ij}. \tag{8.16}$$

When the state of the system is exactly identified to be in state j, the state in MDP is e_j, where

$$e_j = (0_0, 0_1, \ldots, 0_{j-1}, 1_j, 0_{j+1}, \ldots, 0_{s+1}).$$

(2) When action 2 is selected, the next state becomes e_0, since the system is replaced with a new one.

We let $v(x)$ denote the optimal total discounted expected cost in an infinite time horizon when the initial state is x. Then $v(x)$ obeys the following equation.

$$v(x) = \min\{W(x),\ I(x),\ R(x)\}, \tag{8.17}$$

where $W(x)$, $I(x)$ and $R(x)$ correspond to action 0, 1 and 2, respectively, and are

$$\begin{aligned} W(x) &= xL + \beta \sum_\theta P(\theta|x) v(T(x,\theta)), \\ I(x) &= xL + A + \beta \sum_j (xP)_j v(e_j), \\ R(x) &= xC + \beta v(e_0). \end{aligned} \tag{8.18}$$

Corresponding to the above equation, we have the following recursive equation.

$$v^{n+1}(x) = \min\{W^{n+1}(x),\ I^{n+1}(x),\ R^{n+1}(x)\},$$

where

$$\begin{aligned} v^0(x) &= 0, \\ W^{n+1}(x) &= xL + \beta \sum_\theta P(\theta|x) v^n(T(x,\theta)), \\ I^{n+1}(x) &= xL + A + \beta \sum_j (xP)_j v^n(e_j), \\ R^{n+1}(x) &= xC + \beta v^n(e_0). \end{aligned} \tag{8.19}$$

From the theory of contraction mapping, $v^n(x)$ is seen to converge on the optimal cost function $v(x)$ uniformly on X.

8.3.3 Some properties of TP_2 order

In this section, we introduce two partial orders on X and discuss their properties.

Definition 8.3.1 $x \overset{S}{<} x'$ if and only if $\sum_{i=k}^{s+1} x_i \leq \sum_{i=k}^{s+1} x'_i$ for any k, $k = 0, 1, \ldots, s+1$.

Definition 8.3.2 $x \overset{T}{<} x'$ if and only if $x_j x'_i \leq x_i x'_j$ for any $0 \leq i \leq j \leq s+1$.

It is easily seen that the binary relations $\overset{S}{<}$ and $\overset{T}{<}$ are partial orders on X. Under the conditions (C.5), (C.6) and (C.7), we have the following lemmas . For the proof of these lemmas, see Ohnishi [17].

Lemma 8.3.1 $x \overset{S}{<} x'$ if and only if $xf \leq x'f$ for any $f \in F$.

Lemma 8.3.2 If $x \overset{T}{<} x'$, then $x \overset{S}{<} x'$.

Lemma 8.3.3 If $x \overset{T}{<} x'$, then

$$xP \overset{T}{<} x'P.$$

Lemma 8.3.4 If $x \overset{T}{<} x'$, then

$$xPQ \overset{T}{<} x'PQ.$$

This implies that

$$xPQg \leq x'PQg$$
$$\text{for any } g \in \{(g(0), g(1), \ldots, g(m))^T : \ g(0) \leq g(1) \leq \cdots \leq g(m)\}.$$

Note that this binary relation $\overset{T}{<}$ is defined on the set of the probability distribution on $\mathcal{M}$.

Lemma 8.3.5 If $x \overset{T}{<} x'$, then

$$T(x, \theta) \overset{T}{<} T(x', \theta).$$

for any $\theta \in \mathcal{M}$.

Lemma 8.3.6 If $\theta < \theta'$, then

$$T(x, \theta) \overset{T}{<} T(x, \theta').$$

Lemma 8.3.7 Let a real-valued function $f(x, \theta)$ of two variables $x \in X$ and $\theta \in \mathcal{M}$ satisfy the following two properties:

(1) $f(x, \theta)$ is increasing in θ for any x.

(2) If $x \overset{T}{<} x'$, then $f(x, \theta) \leq f(x', \theta)$ for any θ.

Then, if $x \overset{T}{<} x'$,

$$\sum_{\theta} P(\theta|x) f(x,\theta) \leq \sum_{\theta} P(\theta|x') f(x',\theta).$$

Lemma 8.3.8 If $x \overset{T}{<} x'$, then

$$x \overset{T}{<} (1-\alpha)x + \alpha x' \overset{T}{<} x'$$

for any $0 \leq \alpha \leq 1$.

8.3.4 Some properties of optimal function

In this section, we discuss some properties of the optimal function $v(x)$. First we have the following lemma.

Lemma 8.3.9 If $x \overset{T}{<} x'$, then $v(x) \leq v(x')$.

Proof. The proof is done through mathematical induction method in n, using (8.19).

$$W^1(x) \leq W^1(x'),\ I^1(x) \leq I^1(x') \text{ and } R^1(x) \leq R^1(x')$$

are easily seen to hold from the (C.2) and lemma 8.3.1. We assume that

$$v^n(x) \leq v^n(x').$$

Then, from lemma 8.3.5, we have

$$v^n(T(x,\theta)) \leq v^n(T(x',\theta)) \text{ for any } \theta \in \mathcal{M}.$$

Hence , using lemma 8.3.4, we have

$$\sum_{\theta} P(\theta|x) v^n(T(x,\theta)) \leq \sum_{\theta} P(\theta|x') v^n(T(x',\theta)) \text{ for any } \theta \in \mathcal{M}. \tag{8.20}$$

which means that

$$W^{n+1}(x) \leq W^{n+1}(x').$$

It is easily seen that $e_i \overset{T}{\leq} e_j$ for $i \leq j$. Using lemmas 8.3.1, 8.3.2 and 8.3.3, we have

$$\sum_{j} (xP)_j v^n(e_j) \leq \sum_{j} (x'P)_j v^n(e_j), \tag{8.21}$$

which means that

$$I^{n+1}(x) \leq I^{n+1}(x').$$

Hence, from (8.19), we have

$$v^{n+1}(x) \leq v^{n+1}(x'),$$

which means that

$$v(x) \leq v(x')$$

since $v^n(x)$ converges to $v(x)$. ■

Lemma 8.3.10 $v(x)$ is a concave function.

Proof. Proof is done through a mathematical induction method in n, using (8.19). $W^1(x)$, $I^1(x)$ and $R^1(x)$ are linear in x, which means that they are concave functions. We assume that $v^n(x)$ is concave in x. We let

$$u(x) = (\sum_i x_i p_{i0} q_{0\theta}, \sum_i x_i p_{i1} q_{1\theta}, \ldots, \sum_i x_i p_{is} q_{s\theta}),$$

$$U(x) = \sum_i \sum_j x_i p_{ij} q_{j\theta}.$$

Then,

$$P(\theta|x)v^n(T(x,\theta)) = U(x)v^n\left(\frac{u(x)}{U(x)}\right).$$

Here , note that $u(x)$ and $U(x)$ are linear.

For arbitrary x, x' and α $(0 \leq \alpha \leq 1)$, we let

$$x_\alpha = (1-\alpha)x + \alpha x'.$$

Then, we have

$$\begin{aligned}
&U(x_\alpha)v^n\left(\frac{u(x_\alpha)}{U(x_\alpha)}\right) \\
&= U(x_\alpha)v^n\left(\frac{(1-\alpha)U(x)}{U(x_\alpha)}\frac{u(x)}{U(x)} + \frac{\alpha U(x')}{U(x_\alpha)}\frac{u(x')}{U(x')}\right) \\
&\geq U(x_\alpha)\left[\frac{(1-\alpha)U(x)}{U(x_\alpha)}v^n\left(\frac{u(x)}{U(x)}\right) + \frac{\alpha U(x')}{U(x_\alpha)}v^n\left(\frac{u(x')}{U(x')}\right)\right] \\
&= (1-\alpha)U(x)v^n\left(\frac{u(x)}{U(x)}\right) + \alpha U(x')v^n\left(\frac{u(x')}{U(x')}\right)
\end{aligned}$$

which means that $W^{n+1}(x)$ is concave. $I^{n+1}(x)$ and $R^{n+1}(x)$ are linear, that is, they are concave. Hence $v^{n+1}(x)$ is a concave function. Furthermore, $v^{n+1}(x)$ converges on the optimal function $v(x)$ uniformly, which means that $v(x)$ is a concave function. This completes the proof. ■

If $x \overset{T}{<} x'$ implies $h(x) \leq h(x')$, then we say that $h(x)$ is increasing with respect to the partial order $\overset{T}{<}$. Then, according to the same discussion as on the proof of lemmas 8.3.9 and 8.3.10, we have the following lemma.

Lemma 8.3.11 $W(x)$, $I(x)$ and $R(x)$ are increasing with respect to $\overset{T}{<}$ and are concave functions.

8.3.5 Structure of optimal inspection and replacement policy

In this section, we investigate the structural properties of an optimal inspection and replacement policy with incomplete information.

Lemma 8.3.12 $W(x) - R(x)$ and $I(x) - R(x)$ are increasing with respect to the partial order $\overset{T}{<}$ and are concave functions.

Proof. From (8.18), we have

$$W(x) - R(x) = x(L - C) + \beta \sum_{\theta} P(\theta|x)v(T(x,\theta)) - \beta v(e_0),$$

and

$$I(x) - R(x) = x(L - C) + A + \beta \sum_{j} (xP)_j v(e_j) - \beta v(e_0).$$

From (C.3) and (8.20) and (8.21), the result follows. ■

Lemma 8.3.13 $W(x) - I(x)$ is a concave function.

Proof. Since $W(x)$ is a concave function and $I(x)$ is a linear function, the result follows. ■

Using lemmas 8.3.12 and 8.3.13, we obtain the following structure of an optimal policy.

Theorem 8.3.1 We let $D(x)$ denote the optimal action at $x \in X$. For any x and x' ($x \overset{T}{<} x'$), α_1, α_2 and α_3 ($0 \leq \alpha_1 \leq \alpha_2 \leq \alpha_3 \leq 1$) exist such that

$$D(x_\alpha) = \begin{cases} 0, & \text{for } 0 \leq \alpha \leq \alpha_1, \\ 1, & \text{for } \alpha_1 \leq \alpha \leq \alpha_2, \\ 0, & \text{for } \alpha_2 \leq \alpha \leq \alpha_3, \\ 2, & \text{for } \alpha_3 \leq \alpha \leq 1, \end{cases}$$

where $x_\alpha = (1 - \alpha)x + \alpha x'$.

This theorem states that there exists an optimal policy such that the line segment $\{x_\alpha : x_\alpha = (1-\alpha)x + \alpha x',\ 0 \leq \alpha \leq 1, x \overset{T}{<} x'\}$ is divided into at most four regions and a control limit rule for replacement holds.

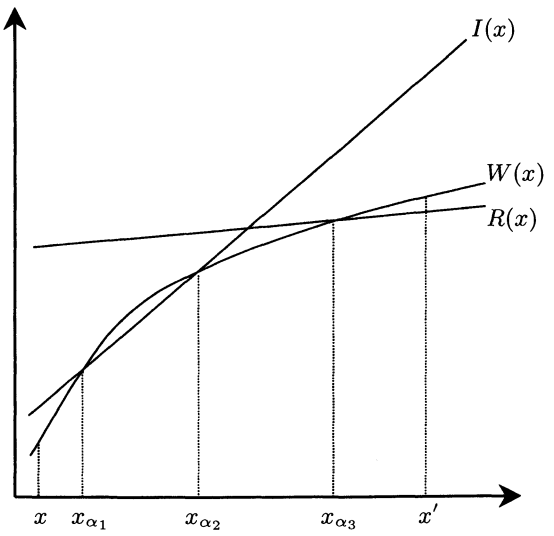


Fig. 8.3. The behavior of $W(x)$, $I(x)$ and $R(x)$

8.4 An Optimal Inspection and Replacement Problem of a Continuous Time Markovian Deteriorating System

8.4.1 A continuous time Markovian deteriorating system

The system considered here has the following properties.

1) The level of deterioration is assumed to be quantified in many states, $0, 1, \ldots, s, s+1$ in order of increasing deterioration, where state 0 is a good state, *i.e.* the system is like new, states $1, 2, \ldots, s$ are deterioration states, and state $s+1$ is a failed state. In a normal operation, these states constitute a continuous time Markov process with an absorbing state $s+1$. We let $\mathcal{S} = \{0, 1, \ldots, s\}$.
2) From state i, a random transition is possible only to state $i+1$ or $s+1$, that is, one-step deterioration or catastrophic failure can occur.
3) The failure of the system can be detected at any time and then the system is replaced with a new one, where replacement time is assumed to be negligible.
4) The states $0, 1, \ldots, s$ cannot be identified without inspection, where inspection time is assumed to be negligible.
5) When the system is observed to be in state $i \in \mathcal{S}$ by inspection or replacement, we can take one of the following actions.

   R : the system is preventively replaced with a new one,
   $I(t)$: we do not replace, and the next inspection is planned to be made t unit time after.
   $I(\infty)$: we do not replace and operate the system without inspection until it fails.

6) We let E_i, $i \in \mathcal{S}$ denote the time instant at which inspection or replacement has been completed and the system is in state i. Then our problem is to find an action at each E_i to minimize the total discounted expected cost.
7) For the cost structure, we treat the following simple case.

a : inspection cost,
b : preventive replacement cost, *i.e.*, the cost for replacement of the system in state $i \in \mathcal{S}$,
c : corrective replacement cost, *i.e.*, the cost of a failed system.

We need the cost $a + b$ for each preventive replacement. We assume that

$$c > a + b.$$

8.4.2 Transition probability

We consider the state transition probability of the system with neither inspection nor replacement, which describes the behavior of the system between two successive inspections and/or replacement. We first introduce the following notations.

β_i : transition rate from state i to state $i+1$, $\beta_s = 0$,
α_i : transition rate from state i to state $i+1$.

We assume that

$$\alpha_i \leq \alpha_j \quad \text{for } i < j,$$

$P_{ij}(t)$: Pr{ the system is in state j at time t | the system is in state i at time 0,}
$F_i(t)$: $P_{i,s+1}(t)$,
$\overline{F}_i(t)$: $1 - F_i(t) = \sum_{j \in \mathcal{S}} P_{ij}(t)$,

$P_{ij}(t)$ and $F_i(t)$ satisfy the following equations.

$$\begin{aligned} \frac{dP_{ij}(t)}{dt} &= -\lambda_i P_{ij}(t) + \beta_i P_{i+1,j}(t) \\ &= -P_{ij}(t)\lambda_j + P_{i,j-1}(t)\beta_{j-1}, \end{aligned} \tag{8.22}$$

$$\begin{aligned} f_i(t) = \frac{dF_i(t)}{dt} &= \lambda_i \overline{F}_i(t) - \beta_i \overline{F}_{i+1}(t) \\ &= \sum_{j \in \mathcal{S}} P_{ij}(t)\alpha_j. \end{aligned} \tag{8.23}$$

With respect to $P_{ij}(t)$, we have the following lemmas.

Lemma 8.4.1 $P_{ij}(t)$ is TP_2 in $i, j \in \mathcal{S}$.

Lemma 8.4.2 $P_{ij}(t)$ is TP_2 in $j \in \mathcal{S}$, t.

Lemma 8.4.3 $\overline{F}_i(t) \geq \overline{F}_j(t)$ for $i < j$.

Proof. For the proofs of these lemmas, see Mine [16]. ■

8.4.3 Formulation by semi-Markov decision process

Our optimal inspection and replacement problem is formulated by semi-Markov decision process. The following notations are introduced. We let v_i the optimal total discounted expected cost when the system starts with E_i, $i \in \mathcal{S}$. We introduce $\alpha > 0$ as a discount factor.

By an elementary probabilistic consideration, v_i is expressed as

$$\begin{aligned} v_i &= \min\{\min_{0<t\le\infty} H_i(t),\ b + v_0\}, \\ H_i(t) &= ae^{-\alpha t}\overline{F}_i(t) + e^{-\alpha t}\sum_{j\in\mathcal{S}} P_{ij}(t)v_j \\ &\quad + (c + v_0)\int_0^t e^{-\alpha x} f_i(x)dx, \end{aligned} \tag{8.24}$$

where $H_i(t)$ $(0 < t < \infty)$, $H(\infty)$ and $b + v_0$ correspond to action $I(t)$, $I(\infty)$ and R, respectively.

It is easily shown that (8.24) is equivalent to the following equation.

$$\begin{aligned} v_i &= \min\{\min_{0<t\le\infty} G_i(t),\ b + v_0\}, \\ G_i(t) &= \frac{1}{1 - e^{-\alpha t}P_{ii}(t)}\left\{ae^{-\alpha t}\overline{F}_i(t) + e^{-\alpha t}\sum_{j=i+1}^{s} P_{ij}(t)v_j\right. \\ &\quad \left. + (c + v_0)\int_0^t e^{-\alpha x} f_i(x)dx\right\}. \end{aligned} \tag{8.25}$$

8.4.4 Structure of optimal policy

We discuss the property of an optimal policy. First, we have the following theorem. We let D_i denote an optimal action at E_i.

Theorem 8.4.1 $D_s = I(\infty)$ or R.

Proof. From (8.25), we have

$$G_s(t) = \frac{ae^{-(\alpha+\lambda_s)t}}{1 - e^{-(\alpha+\lambda_s)t}} + \frac{(c + v_0)\lambda_s}{\alpha + \lambda_s}.$$

It is clear, then, that $G_s(t)$ is decreasing in t, which completes the proof. ■

Theorem 8.4.2 When $D_s = I(\infty)$, then $D_i = I(\infty)$ for $i \in \mathcal{S}$.

Proof. From (8.25), we have

$$G_i(\infty) = (c + v_0)\left(1 - \alpha\int_0^\infty e^{-\alpha x}\overline{F}_i(x)dx\right),$$

which is decreasing in i. Hence, $G_s(\infty) \le b + v_0$ means that $G_i(\infty) \le b + v_0$, that is $D_i \ne R$. We assume that $D_j = I(\infty)$ for $j \ge i + 1$, then we have

$$\begin{aligned} &[G_i(t) - G_i(\infty)](1 - e^{-\alpha t}P_{ii}(t)) \\ &= ae^{-\alpha t}\overline{F}_i(t) + e^{-\alpha t}\sum_{j\in\mathcal{S}} P_{ij}(t)G_j(\infty) \\ &\quad + (c + v_0)\int_0^t e^{-\alpha x} f_i(x)dx - G_i(\infty) \\ &= e^{-\alpha t}\overline{F}_i(t) \ge 0, \end{aligned}$$

which means that $D_i \neq I(t)$, $0 < t < \infty$. ■

Theorem 8.4.3 When $D_s = R$, optimality of a control limit rule for replacement holds.

Proof. It is sufficient to show that $D_i = R$ and $D_{j+1} = D_{j+2} = \cdots = D_s = R$, $j > i$ implies that $D_j = R$.

$$
\begin{aligned}
&[G_j(t) - (b + v_0)](1 - e^{-\alpha t}P_{jj}(t)) - [G_i(t) - (b + v_0)](1 - e^{-\alpha t}P_{ii}(t)) \\
&\geq ae^{-\alpha t}\overline{F}_j(t) + (b + v_0)e^{-\alpha t}\overline{F}_j(t) + (c + v_0)\int_0^t e^{-\alpha x} f_j(x)dx \\
&\quad - \left\{ae^{-\alpha t}\overline{F}_i(t) + (b + v_0)e^{-\alpha t}\overline{F}_i(t) + (c + v_0)\int_0^t e^{-\alpha x} f_i(x)dx\right\} \\
&= (c - a - b)e^{-\alpha t}\overline{F}_i(t) + \alpha(c + v_0)\int_0^t e^{-\alpha x}\overline{F}_i(x)dx \\
&\quad - \left\{(c - a - b)e^{-\alpha t}\overline{F}_j(t) + \alpha(c + v_0)\int_0^t e^{-\alpha x}\overline{F}_j(x)dx\right\} \geq 0,
\end{aligned}
$$

since $c > a + b$ and $\overline{F}_i(x) \geq \overline{F}_j(x)$ for $i < j$. ■

In the following, we consider the case where

$$
D_i = \begin{cases} I(t_i), & \text{for } 0 \leq i \leq K - 1, \\ R, & \text{for } K \leq i \leq s, \end{cases}
$$

where $0 \leq K \leq s$.

By using (8.22) and (8.23), we have that

$$
\begin{aligned}
\frac{dH_i(t)}{dt} &= -(\alpha + \lambda_i)H_i(t) + \beta_i H_{i+1}(t) + \alpha_i(c + v_0) \\
&= e^{-\alpha t}\sum_{j \in \mathcal{S}} P_{ij}(t)h_j,
\end{aligned}
$$

where

$$
h_i = -(\alpha + \lambda_i)v_i + \beta_i v_{i+1} + \alpha_i(c + v_0) - (\alpha + \alpha_i)a.
$$

For $i = 0, 1, \ldots, K - 1$, we have

$$
\begin{aligned}
0 = \frac{dH_i(t_i)}{dt} &= -(\alpha + \lambda_i)v_i + \beta_i H_{i+1}(t_i) + \alpha_i(c + v_0) \\
&\geq -(\alpha + \lambda_i)v_i + \beta_i v_{i+1} + \alpha_i(c + v_0).
\end{aligned}
$$

Hence, $h_i < 0$.

For $i = K, \ldots, s$, we have

$$
h_i = (c - a - b)\alpha_i - (a + b + v_0)\alpha
$$

which increases in i. Hence h_i, $i \in \mathcal{S}$ changes its sign at most once, and if a sign change occurs it is from negative. Since $P_{ij}(t)$ is TP_2 in i, j $(i, j \in \mathcal{S})$ and in $j \in \mathcal{S}$, t, $\sum_{j \in \mathcal{S}} P_{ij}(t)h_j$ changes its sign at most once in t for each i and in i for each t by variation diminishing property of TP_2 function. From the

above discussion, we have the following theorem for the structural property of an optimal policy.

Theorem 8.4.4 There exists an optimal policy with the form

$$D_i = \begin{cases} I(t_i), & \text{for } 0 \le i \le K-1, \\ R, & \text{for } K \le i \le s, \end{cases} \tag{8.26}$$

and $\infty \ge t_0 \ge t_1 \cdots \ge t_{K-1}$.

8.5 An Optimal Maintenance Problem for a Queueing System

This section studies an optimal maintenance policy for a queueing system. Though service to the customer is one of the main purposes of a queueing system, failure of the server makes it impossible to serve the customers. To achieve the maximum service or minimum loss of the customers, we need to consider an optimal maintenance policy for queueing systems. For this purpose, we must consider the processes of the queue length and deterioration, which means that the state space becomes two-dimensional. In a problem with two-dimensional state space, the switch curve structure is often discussed, as the control limit rule is discussed in a problem with one-dimensional state space. In this section, we introduce and analyze a maintenance problem for an $M/G/1$ queueing system.

8.5.1 Model description

We consider an $M/G/1$ queueing system with arrival rate $\lambda(>0)$ and service distribution function $G(t)$. We assume the existence of density function $g(x)$. The server has $s+2$ states $\{0, 1, \ldots, s, s+1\}$. As explained in the previous sections, the state 0 is a good state, *i.e.*, the system is like new, the states $1, \ldots, s$ are deterioration states, the state $s+1$ is a failure state, and the sets $\mathcal{S}+1$ and $\mathcal{S}$ are also defined in a similar way. The server state transition forms a continuous time Markovian process with transition rate γ_{ij}. We can assume $\sum_j \gamma_{ij} = \Gamma$ for all i by uniformization [21]. The queue length process and the server state process are assumed to be independent. When the server fails, the system stops and cannot serve the customer. To prevent a failure and to recover from a failure, we have a preventive maintenance action and a corrective maintenance action, respectively. At decision epochs, we can start the maintenance. The decision epochs are the departure time, the transition time of the server state when the system is empty, and the arrival time to the empty system. For convenience, the failure epoch is also considered as a decision epoch, though only one action (corrective maintenance) is allowed. In other words, the decision epochs are the time when either the queue length or the server state changes and no customer is in service. To perform the maintenance, we lose the customers who are in the system at the start of the maintenance and who arrive while the maintenance. The distribution functions of the preventive and the corrective maintenance time are denoted by $H_1(x)$ and $H_2(x)$, respectively. By maintenance the server becomes new and the

system becomes empty, because all the customers are lost during maintenance. We consider that a cost of one for each lost customer is incurred and minimize the total expected discounted cost with discount factor α.

The system state is expressed by the pair of the queue length i ($i \geq 0$) and the server state $k \in \mathcal{S} + 1$.
In state (i, k) ($k \in \mathcal{S}$), we can choose one of the following two actions

action 1: the preventive maintenance,
action 2: continuing the service.

In $(i, s+1)$ we must choose

action 3: the corrective maintenance.

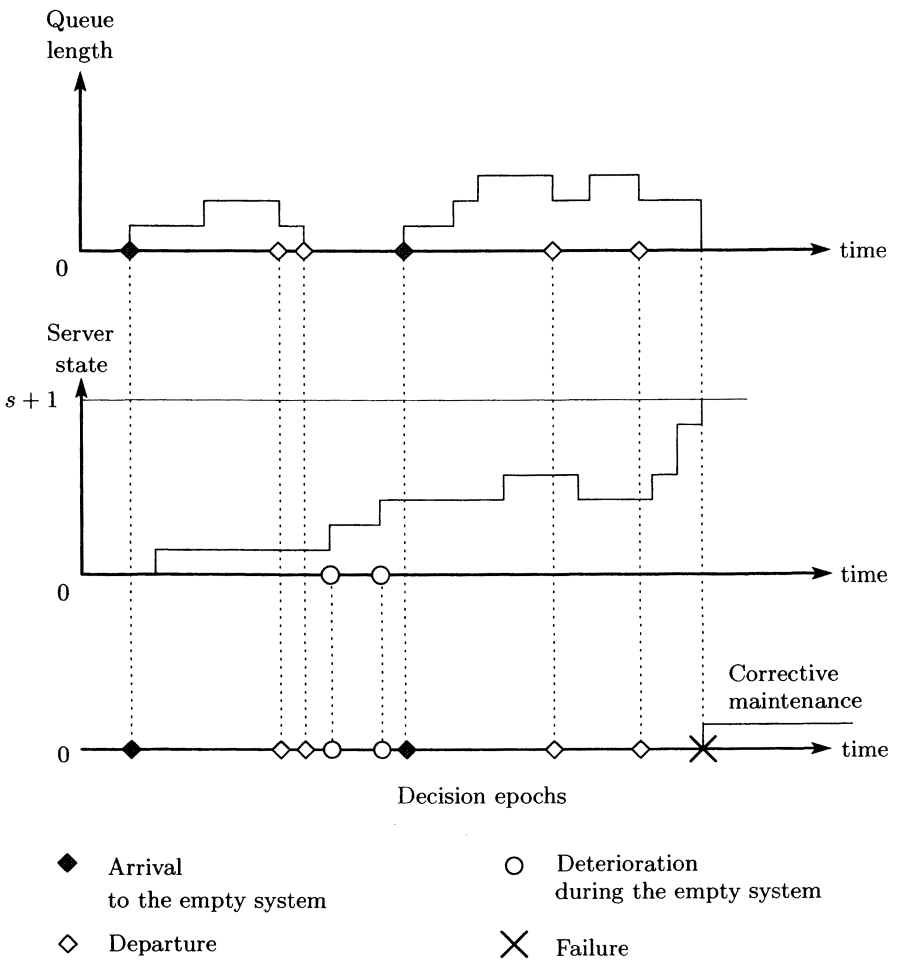


Fig. 8.4. The process of queue length, server state and decision epoch

The transition probabilities are expressed by using following probabilities.

$A_i(x)$: The probability that i customers arrive at the queue during a time x. Since the Poisson arrival is assumed,

$$A_i(x) = \frac{(\lambda x)^i}{i!} e^{-\lambda x}.$$

$Q_{kl}(x)$: The probability that the server state is l after time x (without the failure), given the initial state k.

By this notation, we can express the transition probability from state (i,k) to (j,l) within time t for each action.

(1) When action 1 is selected, the next state becomes $(0,0)$ and the transition time distribution is $H_1(t)$.

(2) When action 2 is selected, the probabilities to state (j,l) are

$$\begin{cases} \displaystyle\int_0^t A_{j-i+1}(x) Q_{kl}(x) g(x) dx & \text{if } i \geq 1 \text{ and } l \in \mathcal{S}, \\ \displaystyle\int_0^t A_{j-i}(x) \overline{G}(x) dQ_{k,s+1}(x) & \text{if } i \geq 1 \text{ and } l = s+1, \\ \displaystyle\int_0^t \gamma_{kl} e^{-(\Gamma+\lambda)x} dx & \text{if } i = 0 \text{ and } j = 0, \\ \displaystyle\int_0^t \lambda e^{-(\Gamma+\lambda)x} dx & \text{if } i = 0,\ j = 1 \text{ and } k = l. \end{cases} \tag{8.27}$$

(3) When action 3 is selected, the next state becomes $(0,0)$ and the transition time distribution is $H_2(t)$.

We consider the cost functions which specify the total expected discounted cost until the next transition for each pair of state and action.

Here we consider the total expected discounted number of arrivals in the time interval $(0,t]$. We can easily obtain

$$N(t) = \frac{\lambda}{\alpha}(1 - e^{-\alpha t}). \tag{8.28}$$

The function $N(t)$ is also the total expected discounted cost during the maintenance when time t is needed for the maintenance, because all arrivals are lost during the maintenance.

(1) When action 1 is selected in state (i,k), the expected cost until the next transition is

$$i + \int_0^\infty N(t) dH_1(t) = i + \lambda h_1, \tag{8.29}$$

where $h_1 = \displaystyle\int_0^\infty e^{-\alpha t} \overline{H}_1(t) dt$. ($h_2$ is also defined in the same way.)

(2) When action 2 is selected in state (i,k), the expected cost until the next transition is 0.

(3) When action 3 is selected in state $(i, s+1)$, the expected cost until the next transition is

$$i + \int_0^\infty N(t) dH_2(t) = i + \lambda h_2. \tag{8.30}$$

8.5.2 Formulation by semi-Markov decision process

For the optimality equation, we define the following functions with respect to (i,k):

$M(i,k)$: the cost function when we perform preventive maintenance at the decision epoch and operate optimally thereafter,
$W(i,k)$: the cost function when we continue the service at the decision epoch and operate optimally thereafter,
$V(i,k)$: the optimal value function for state (i,k),
$D(i,k)$: the optimal action for state (i,k).

Through the standard use of semi-Markov decision process, we obtain the following equations for this problem.

$$
\begin{aligned}
M(i,k) &= i + \lambda h_1 + (1-\alpha h_1)V(0,0), \\
W(0,k) &= \frac{1}{\alpha+\lambda+\Gamma}\left[\sum_{l=0}^{s+1}\gamma_{kl}V(0,l) + \lambda V(1,k)\right], \\
W(i,k) &= \sum_{j=0}^{\infty}\sum_{l=0}^{s}\int_0^{\infty} e^{-\alpha x}A_j(x)Q_{kl}(x)V(i+j-1,l)g(x)dx \qquad (8.31) \\
&\quad + \sum_{j=0}^{\infty}\int_0^{\infty} e^{-\alpha x}A_j(x)\overline{G}(x)V(i+j,s+1)dQ_{i,s+1}(x), \quad i \geq 1, \\
V(i,k) &= \min[M(i,k),\ W(i,k)], \qquad k \in \mathcal{S}, \\
V(i,s+1) &= i + \lambda h_2 + (1-\alpha h_2)V(0,0).
\end{aligned}
$$

Since $M(i,k)$ is independent of k, $M(i,k)$ is denoted by $M(i)$ in the rest of this section.

8.5.3 Properties of value function

We introduce the following conditions.

(C.8) For any m, $\sum_{l=m}^{s+1}\gamma_{kl}$ is increasing in k.
(C.9) $\overline{H}_2(x) \geq \overline{H}_1(x)$ for all x.
(C.10) $\lambda \leq g(x)/\overline{G}(x)$ for all x.

(C.8) means that as the system deteriorates, it is more likely to make a transition to higher states. (C.9) indicates that the time for corrective maintenance is stochastically longer than that for preventive maintenance. From (C.9), $h_1 \leq h_2$ holds. (C.10) indicates that the service rate is always larger than the arrival rate. From (C.8), the following lemma is obtained [23].

Lemma 8.5.1 For any m, $\sum_{l=m}^{s+1} Q_{kl}(x)$ is increasing in k.

Under these conditions, we have the following lemma through the value iteration similar to (8.2) and the mathematical induction method in n.

Lemma 8.5.2

1. $W(i,k)$ and $V(i,k)$ are increasing with respect to k.
2. $W(i+1,k) - W(i,k) \leq 1$.

Proof. For the proof, see Koyanagi [13]. ■

8.5.4 Structure of optimal policy

We obtain the following theorem for the structure of the optimal policy by lemma 8.5.2.

Theorem 8.5.1 If $D(i,l) = 2$, then $D(j,k) = 2$ for $l \geq k$ and $i \leq j$.

Proof. First, we note that $W(i,l) \leq M(i)$ holds when $D(i,l) = 2$. Then we can prove the following inequality

$$\begin{aligned} W(j,k) &\leq W(j,l) \leq W(i,l) + j - i \\ &\leq M(i) + j - i = M(j). \end{aligned}$$

Thus, $D(j,k) = 2$. The first and the second inequalities hold with the first and the second properties of lemma 8.5.2, respectively. ■

This theorem indicates the switch curve structure of the optimal policy. The switch curve structure indicates that a two-dimensional state space is divided by an increasing function and the optimal action changes across the function.

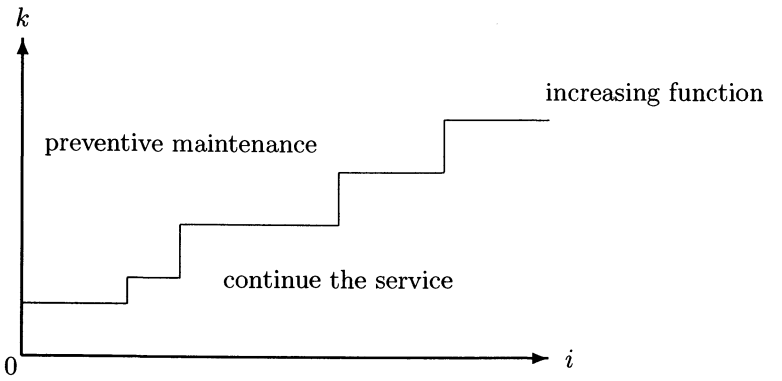


Fig. 8.5. The switch curve structure of the optimal policy

References

1. Awi, F. and So, K. C. (1990), "Optimal maintenance policies for single-server queueing systems subject to breakdowns," *Operations Research*, **38**, 330–343
2. Barlow, R. E. and Proschan, F. (1965), Mathematical Theory of Reliability. John Wiley & Sons, New York
3. Birge, J. *et al.* (1990), "Single-machine scheduling subject to stochastic breakdowns," *Naval Research Logistics Quarterly*, **37**, 661–677
4. Derman, C. (1970), Finite State Markovian Decision Processes. Academic Press, New York
5. Eckles, J. E. (1968), "Optimum maintenance with incomplete information," *Operations Research*, **16**, 1067–1085
6. Glasserman, P. and Yao, D. (1994), "Monotone optimal control of permutable GSMPs," *Mathematics of Operations Research*, **19**, 449–476
7. Hajek, B. (1984), "Optimal control of two interacting service stations," *IEEE Transactions on Automatic Control*, **29**, 491–499
8. Karlin, J. (1968), Total Positivity. **Vol.1**, Stanford University Press, Stanford, California
9. Kawai, H. and Koyanagi, J. (1992), "An optimal maintenance policy of a discrete time Markovian deterioration system," *Computer & Mathematics with Applications*, 103–108
10. Koyanagi, J. and Kawai, H. (1998), "An optimal maintenance policy for a server with decreasing arrival rate and non-cancelable customer," *Journal of Quality in Maintenance Engineering*, **4**, 299–311
11. Koyanagi, J. and Kawai, H. (1995), "An assignment problem for a parallel queueing system with two heterogeneous servers," *Computer and Mathematical Modelling*, **22**, 173–181
12. Koyanagi, J. and Kawai, H. (1997), "An optimal maintenance policy for a queueing system server under periodic observation," *Reliability, Quality and Safety Engineering*, **4**, 357-367
13. Koyanagi, J. and Kawai, H. (1997), "An optimal maintenance policy for a deteriorating server of an $M/G/1$ queueing system," Stochastic Modelling in Innovative Manufacturing (Christer, A. H., Osaki, S. and Thomas, L. C. eds.). Lecture Notes in Economics and Mathematical Systems,**445**, 215-224
14. Monahan, G. E. (1982), "A survey of partially observable Markov decision processes, theory, models and algorithm," *Management Science*, **28**, 1-16
15. Mine, H. and Kawai, H. (1975), "An optimal inspection and replacement," *IEEE Transactions on Reliability*, **24**, 305-309
16. Mine, H. and Kawai, H. (1982), "An optimal inspection and maintenance policy of a deteriorating system," *Journal of Operations Research Society of Japan*, **25**, 1-15

17. Ohnishi, M., Kawai, H. and Mine, H. (1986), "An optimal inspection and replacement policy under incomplete state information," *European Journal of Operational Research*, **27**, 117-128
18. Ohnishi, M., Kawai, H. and Mine, H. (1986), "An optimal inspection and replacement policy for a deteriorating system," *Journal of Applied Probability*, **23**, 973-988
19. Rosenfield, D. (1976), "Markovian deterioration with uncertain information," *Operations Research*, **24**, 141-155
20. Ross, S. (1970), Applied Probability Models with Optimization Application. Holden-Day, San Francisco, California
21. Serfozo, R. (1979), "An equivalence between discrete and continuous time Markov decision processes," *Operations Research*, **27**, 616–620
22. Sondik, E. (1978), "The optimal control of partially observable Markov processes over the infinite horizon, discounted cost," *Operations Research*, **26**, 282-304
23. Stoyan, D. (1983), Comparison Methods for Queues and Other Stochastic Models. John Wiley & Sons, New York
24. Veatch, M. and Wein, L. (1992), "Monotone control of queueing networks," *Queueing System*, **12**, 391–408
25. Walrand, J. (1988), An Introduction to Queueing Networks. Prentice Hall, New Jersey
26. Wessels, J. (1977), "Markov programming by successive approximations with respect to weighted supremum Norms," *Journal of Mathematical Analysis and Applications*, **58**, 326–335
27. White, C. (1978), "Optimal inspection and repair of a production process subject to deterioration," *Journal of the Operational Research Society*, **29**, 235-243
28. Wolff, R. W. (1989), Stochastic Modeling and the Theory of Queues. Prentice Hall, New Jersey

9. Transient Analysis of Semi–Markov Reliability Models – A Tutorial Review with Emphasis on Discrete–Parameter Approaches

Attila Csenki
School of Computing and Mathematics,
University of Bradford
Bradford BD7 1DP, Great Britain

Summary.

Semi–Markov models can be usefully employed for the analysis of various reliability and performance characteristics of technical systems. We consider semi–Markov dependability models of systems whose finite state space $\mathcal{S}$ is partitioned into the set of up states $\mathcal{U}$, the set of repairable down states $\mathcal{D}$, and a state ω, standing for irrecoverable system failure: $\mathcal{S} = \mathcal{U} \cup \mathcal{D} \cup \{\omega\}$. A great number of reliability and performance measures for such models are examined in this chapter. It is discussed how these measures can be obtained as solutions of certain systems of integral equations if the modelling process has a continuous time parameter. The discrete parameter framework is discussed separately in more detail, since, even though it is subsumed within the continuous parameter case, it deserves special attention for three reasons. First, in some applications, discrete–time modelling is appropriate. Second, discrete–time models can be used for approximately analysing continuous–time models. Finally, for discrete–time models, the computational solution is technically less involved and it will thus be more accessible for even the mathematically less sophisticated analyst. We also present a framework within which many of the measures will be seen to admit of a formal closed form solution.

Keywords: Reliability, maintenance, semi–Markov processes

AMS Subject Classifications: 60K05, 60K15, 60K20, 90B25

9.1 Introduction

For the evaluation of reliability and performance characteristics of real-life industrial or computer systems, it is fairly common to assume a Markovian (*e.g.*, [12], [60], [99]) or semi–Markovian framework (*e.g.*, [3], [20], [59], [71], [87]). Whereas knowledge of the analysis techniques for Markov systems is rather universal and the topic is well covered in the textbook literature (*e.g.*, [8], [53], [56], [65], [67], [72], [96], [104], [110]), the reliability textbooks also dealing with the semi–Markov context at any length are much fewer in number (*e.g.*, [9], [91], [92], [101]). The analysis of semi–Markov models in various applications has for decades been an active research area (as evidenced, for example, by

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

the books [29], [58], [63], [84], [89], [109]). It appears though that a systematic collection of results pertaining to the analysis of semi–Markov reliability models has not been undertaken. It is our intention here to present a unified framework within which all known reliability/performance characteristics of semi–Markov models can be arrived at in a systematic fashion. The work reported here has partly been inspired by the author's study of Birolini's books [13], [14] which, with the material in this paper, will attain, so it is hoped, a certain degree of completion.

This chapter is organised as follows. In Section 9.2, we present the class of models which will be examined later. These are essentially semi–Markov models whose finite state space is partitioned into 'working,' 'repair,' and 'failure' states. In Section 9.3, we collect and interrelate the most commonly used reliability and dependability measures for the class of systems under consideration. In Section 9.4, the methods of analysis are expounded by exemplifying them on a selected set of the measures from Section 9.3. The renewal argument in its various guises will be seen to lead to systems of integral equations, or, in the discrete–parameter case, to recurrence relations. In Section 9.5, we collect and discuss (without proofs) the results pertaining to some further reliability measures not covered in Section 9.4. In Section 9.6, we discuss numerical solution techniques and describe in particular how to approximate continuous–time models by discrete–parameter models for a computational solution of the former. The chapter finishes with Section 9.7, where some recent developments are referred to and possible futher work is indicated.

9.2 Modelling Framework

Two classes of models will be considered here: models of irreparable, and models of repairable systems. The models belonging to the first category are semi–Markov processes Y whose finite state space $\mathcal{S}$ is partitioned as $\mathcal{S} = \mathcal{U} \cup \mathcal{D} \cup \Omega$, where $\mathcal{U}$ stands for the set of 'up' (*i.e.*, working) states, $\mathcal{D}$ stands for the set of repairable 'down' states, and the states in Ω stand for irrecoverable system failure. Y is thus absorbing with Ω being the set of all absorbing states. Ω is assumed to contain one single element if the analysis does not involve considering the system's failure modes; in this case, $\Omega = \{\omega\}$, say. Models of repairable systems, on the other hand, will have their finite state space partitioned as $\mathcal{S} = \mathcal{U} \cup \mathcal{D}$, where, as before, $\mathcal{U}$ and $\mathcal{D}$ stand for the set of working and repairable down states, respectively. In this latter case, the system will alternate between $\mathcal{U}$ and $\mathcal{D}$ indefinitely and it is usually assumed that the modelling process Y is irreducible. (It should be noted though that irreducibility is a somewhat stronger assumption than what is needed for modelling the repairability condition.)

We can also distinguish between continuous– and discrete–parameter models. Let us first describe the notational framework and the assumptions for the continuous–parameter case. Here, the parameter set is $[0, +\infty)$ and t is the time parameter, *i.e.*, $Y = \big(Y_t : t \in [0, +\infty)\big)$. The sequence of states visited by Y is governed by the transition probabilities $p_{s,s'}, s, s' \in S, s \neq s'$, giving rise to $\mathbf{P} = (p_{s,s'} : s, s' \in S)$, the transition probability matrix of the embedded Markov chain $X = (X_i : i = 0, 1, \cdots)$. (The diagonal entries of $\mathbf{P}$ are zero by definition.) $F_{s,s'}$ stands for the cumulative distribution function (cdf) of the holding time in $s \in S$ given that the next state to be visited by Y is $s' \in S, s' \neq s$. The matrix $\mathbf{P}$ and the conditional holding time distributions

allow a complete probabilistic description of the behaviour of Y, provided the initial condition at $t = 0$ is known. It will be assumed that Y starts in some $s \in S$ at time zero, formally denoted by the event $\{Y_0 = s, Y_{0-} \neq s\}$. It is also assumed that no instantaneous transitions may occur, *i.e.* $F_{s,s'}(0) = 0$ for all $s, s' \in S$, $s \neq s'$. The semi–Markov kernel $\mathbf{Q}(t) = \big(q_{s,s'}(t) : s, s' \in S\big)$, defined by $q_{s,s'}(t) = p_{s,s'} F_{s,s'}(t)$, $t \geq 0$ [66] or [92] is a matrix–valued function on $[0, +\infty)$ whose entries are cdfs of finite measures on $[0, +\infty)$; it will be used to represent our results concisely. (Again, the diagonal entries of $\mathbf{Q}$ are identical to zero by definition.)

A great deal of material is available in the literature on discrete–parameter models in the reliability/performance area. Whereas most of it is concerned with models which are, or which can be made, Markovian (see, *e.g.*, [17] and [116] and the references cited therein), the semi–Markov case has received much less attention [55]. Discrete–parameter models are important for two reasons. First, in some modelling contexts, the system's age is best measured not in units of time but in units of some other, integral quantity; for example, the age of an assembly plant (or some component thereof) may be measured by the number of units processed; or, the mechanical age of a vehicle (though usually not its monetary value) is a function of the vehicle's mileage. This latter example is indicative of the other role in which discrete–parameter models are used: it is in a practical sense not meaningful to quote a vehicle's exact mileage; for, say, maintenance purposes, it suffices to record it to the nearest 1000 miles. Thus, what we are given is a discrete–parameter approximation to a continuous–parameter process. It is indeed an established technique for numerical analysis to replace a continuous–time dependability model by an approximating discrete–parameter model; in [51] and [52], for example, a continuous–time Markov model of a computer system is numerically analyzed by considering a discrete–parameter Markov chain approximating the original model at the time instances $0, \delta, 2\delta, \cdots$ with some 'small' $\delta > 0$. This discrete–time approximation idea is, incidentally, useful also in other contexts than the computational. For example, in [40] and [24], closed form expressions for certain quantities concerning continuous–time Markov processes could be derived by relating them to their discrete–parameter counterpart. Also, in [58], p. 842, the Kolmogorov differential equations are arrived at via their discrete–parameter counterpart. (This classical work by Howard [58], Chapter 10, is still the best reference for discrete–time semi–Markov models.) Mode and Pickens in [84] and [85] have demonstrated the utility of discrete–parameter semi–Markov models in the context of demography. In [85], it is argued that discrete–parameter modelling has several advantages: it may be more appropriate from a modelling point of view; it is readily amenable to computer implementation, *i.e.* it obviates the need for the discretisation step involved in the numerical solution of continuous–time models; finally, the modeller, not always a mathematician, may be more comfortable working with discrete–parameter entities. Even though discrete–parameter semi–Markov processes can be subsumed within the continuous–parameter case, it is for the above reasons that they will be commented upon separately. In addition, the construction of approximate discrete–parameter models for a given continuous–time model will also be described. The notational frameworks for the discrete– and continuous–parameter cases are similar. Now, the parameter set is assumed to be $\Delta = \{\delta_0, \delta_1, \delta_2, \cdots\}$ with $\delta_0 < \delta_1 < \delta_2 < \cdots$, and the system may be observed at, for example, equally spaced time instances starting at 0, in which case $\delta_i = i\delta$, $i \geq 0$. As before, the embedded Markov chain $X = (X_i : i = 0, 1, \cdots)$ records the sequence of states visited by Y. $\mathbf{P}$ is the

transition probability matrix of X. Now, it will be more appropriate to work with the probability mass functions (rather than the cdfs) of the conditional holding times: $f_{s,s'}(t)$ stands for probability of Y spending t time instants in $s \in S$ given that the next state to be visited is $s' \in S$, $s' \neq s$. By definition, $f_{s,s'}(0) = 0$. In lieu of the semi–Markov kernel, we shall use the matrix–valued function $\mathbf{H}(t)$, defined by $\mathbf{H}(t) = \big(h_{s,s'}(t) : s, s' \in S\big)$, $t = 0, 1, \cdots$, where

$$h_{s,s'} = \begin{cases} p_{s,s'} f_{s,s'}, & \text{for } s \neq s', \\ 0, & \text{for } s = s'. \end{cases}$$

Notice that $\mathbf{H}$ can be thought of as the density with respect to the counting measure of the semi–Markov kernel of Y. Corresponding to the continuous–parameter case, it will be assumed that Y starts a new visit in some $s \in S$ at time δ_0, formally expressed by the event $\{Y_{\delta_0} = s, Y_{\delta_1} \neq s\}$.

To conclude this section, let us add some more notation. $\mathbf{I}$ and $\mathbf{0}$ stand respectively for the identity matrix and the zero matrix; $\mathbf{1}$ stands for the column vector of ones. (The size of these will be clear from the context.) An obvious subscript notation will be used to denote submatrices of a given matrix or a group of entries of a given vector; for example, the $\mathcal{U}$–$\mathcal{D}$–entries of $\mathbf{Q}$ will be denoted by $\mathbf{Q}_{\mathcal{U}\mathcal{D}}$, *i.e.* $\mathbf{Q}_{\mathcal{U}\mathcal{D}} = (q_{u,d} : u \in \mathcal{U}, d \in \mathcal{D})$. Finally, $I_{\{\cdots\}}$ denotes the indicator function of the subscript event $\{\cdots\}$.

9.3 Dependability Measures

Practical, application–driven circumstances will determine the reliability measures chosen by the reliability engineer for the assessment of any given system. In what follows, we discuss a class of measures which are, from a mathematical point of view, closely related to each other in that for semi–Markov models they are all amenable to analysis by the same technique, the renewal argument. The core set of the measures addressed below is listed in Birolini's books [13], [14], where the focus of attention is on two types of systems: first, the one–unit system modelled by the alternating renewal process (this is the special case when the state space S comprises two elements, $S = \{u\} \cup \{d\}$); and, second, the many–unit system modelled by a Markov process.

We are now going to introduce the most commonly used reliability measures. The *reliability* is defined as the probability of the system being in the set of up states $\mathcal{U}$ throughout the time interval $[0, t]$, $t \geq 0$,

$$R(t) = \mathbb{P}(Y_v \in \mathcal{U} \text{ for all } v \in [0, t]).$$

The *point availability* is defined as the probability that the system is in the set of up states $\mathcal{U}$ at the time instant t, $t \geq 0$ [13], [14],

$$PA(t) = \mathbb{P}(Y_t \in \mathcal{U}).$$

The *interval reliability* is defined as the probability that the system is in the set of up states $\mathcal{U}$ throughout the time interval $[t, t + x]$, with $x, t \geq 0$ [13], [14], [30], [34],

$$IR(x, t) = \mathbb{P}(Y_v \in \mathcal{U} \text{ for all } v \in [t, t + x]).$$

For $t = 0$, the above is the reliability: $IR(x, 0) = R(x)$; for $x = 0$, it equals the point availability: $IR(0, t) = PA(t)$. The next two reliability measures are instances of what Baxter in [6] and [7] termed 'compound availability measures'.

The first of them is the *joint availability*. It is defined as the probability of the system being in the set of up states $\mathcal{U}$ at both time instants t and $t+x$ with $t, x \geq 0$,

$$JA(x,t) = \mathbb{P}(Y_t, Y_{t+x} \in \mathcal{U}).$$

The *joint interval reliability* is defined as the probability that the system is in the set of up states $\mathcal{U}$ throughout both $[t_1, t_1+x_1]$ and $[t_2, t_2+x_2]$ with $x_1, t_1, x_2, t_2 \geq 0$ [6], [13], [14],

$$JIR(x_1, x_2, t_1, t_2) =$$
$$\mathbb{P}(Y_v \in \mathcal{U} \text{ for all } v \in [t_1, t_1+x_1] \cup [t_2, t_2+x_2]).$$

If the intervals $[t_1, t_1+x_1]$ and $[t_2, t_2+x_2]$ overlap, the joint interval reliability can be written in terms of the interval reliability as

$$JIR(x_1, x_2, t_1, t_2) =$$
$$IR(max\{x_1+t_1, x_2+t_2\} - min\{t_1, t_2\}, min\{t_1, t_2\}).$$

We note that the measures JA and JIR (and many of the other compound availability measures) have been discussed in the literature for various special systems [64], [88]. All the reliability measures listed thus far are seen to be instances of the *set reliability*,

$$SR(T) = \mathbb{P}(Y_t \in \mathcal{U} \text{ for all } t \in T),$$

where T is some non–empty subset of $[0, +\infty)$ [35]. The *cumulative work until final breakdown* is defined as

$$C = \int_0^{+\infty} I_{\{Y_t \in \mathcal{U}\}} dt.$$

This performance measure is applicable to systems the state space of which comprises up states, repairable down states and one or more irrecoverable failure states; the modelling semi–Markov process is thus absorbing. The *cumulative operational time* is defined as

$$CO(t) = \int_0^t I_{\{Y_v \in \mathcal{U}\}} dv.$$

This is the total time spent by the system in the set up states $\mathcal{U}$ during the finite time interval $[0, t]$, $t \geq 0$; see, *e.g.* [28], [43], [100]. Notice that for absorbing models it is $CO(+\infty) = C$, the cumulative work until final breakdown. Another performance measure expressible in terms of the cumulative operational time $CO(t)$ is the *interval availability* [37], [110]

$$IA(t, t+x) = \frac{1}{x} \int_t^{t+x} I_{\{Y_v \in \mathcal{U}\}} dv = \frac{CO(t+x) - CO(t)}{x};$$

this is the proportion of time spent by the system in $\mathcal{U}$ during $[t, t+x]$, where $x > 0$, $t \geq 0$. For the definition of the next dependability measure, we need the notion of the *mission time*, $m(t)$; this is the earliest calendar time by which a mission of length t is completed, *i.e.*,

$$m(t) = min\{v \geq 0 : CO(v) \geq t\}.$$

The above is well defined since the sample paths of $CO(.)$ are continuous and non–decreasing with $CO(t) \to +\infty$ a.s. for $t \to +\infty$. Now, the *mission availability* [13], [14], [36], is defined as the probability that during the completion

of a mission of length $t_0 > 0$, no down–period lasts for longer than $t_f > 0$ time units,

$$MA(t_0, t_f) =$$
$$\mathbb{P}(\text{no visit of } Y \text{ in } \mathcal{D} \text{ during } [0, m(t_0)] \text{ lasts for longer than } t_f).$$

This dependability measure is of importance in applications where system down periods of up to a certain length can be tolerated, for example, in the nuclear and food industries; a special system of this kind was examined in [54]. We note in passing that there seems to be some disagreement in the literature about what should be termed 'mission availability'. In [65], p. 139, 'mission availability' is used to denote what is the mean of the interval availability, and what in [101] is termed the 'fractional duration in the set of up states $\mathcal{U}$'; we shall use the latter terminology. A measure similar to the mission availability is the *work–mission availability* [13], [14], [39],

$$WMA(t_0, t_f) = \mathbb{P}\left(\int_0^{m(t_0)} I_{\{Y_t \in \mathcal{D}\}} dt \leq t_f\right).$$

It is the probability that during the completion of a mission of length $t_0 > 0$, in total no more than $t_f > 0$ units of time are spent in the set of down states $\mathcal{D}$. From

$$\int_0^{m(t_0)} I_{\{Y_t \in \mathcal{D}\}} dt = m(t_0) - \int_0^{m(t_0)} I_{\{Y_t \in \mathcal{U}\}} dt = m(t_0) - t_0,$$

we see that the work–mission availability can be expressed in terms of the mission time m as $WMA(t_0, t_f) = \mathbb{P}(m(t_0) - t_0 \leq t_f)$. It can also be expressed in terms of the cdf of the cumulative operational time CO. To show this, notice that CO and m are interrelated by $CO(t) \geq v \Leftrightarrow m(v) \leq t$. Thus,

$$\begin{aligned} WMA(t_0, t_f) &= \mathbb{P}(m(t_0) \leq t_f + t_0) \\ &= \mathbb{P}(CO(t_f + t_0 \geq t_0) \\ &= 1 - \mathbb{P}(CO(t_f + t_0 < t_0 - 0). \end{aligned} \tag{9.1}$$

For a repairable system, the modelling irreducible semi–Markov process Y alternates between $\mathcal{U}$ and $\mathcal{D}$ indefinitely. Let the length of the ith visit to $\mathcal{B} \in \{\mathcal{U}, \mathcal{D}\}$ be denoted by $T_{\mathcal{B},i}$, $i = 1, 2, \cdots$. The (joint) distribution of these *sequences of working and repair periods*, also called '*sojourn times*' and denoted respectively by $\{T_{\mathcal{U},i} : i = 1, 2, \cdots\}$ and $\{T_{\mathcal{D},i} : i = 1, 2, \cdots\}$, is a dependability measure of the system. Notice in particular that the reliability can be expressed in terms of the first sojourn time of Y in $\mathcal{U}$ as $R(t) = \mathbb{P}(T_{\mathcal{U},1} > t)$. A closely related measure is the *length of the first m working (and repair) periods*, $TS_{\mathcal{B},m} = T_{\mathcal{B},1} + \cdots + T_{\mathcal{B},m}$, $m \geq 1$, $\mathcal{B} \in \{\mathcal{U}, \mathcal{D}\}$. The latter two dependability measures are, of course, also declared for the absorbing model. Then, we have for $\mathcal{B} \in \{\mathcal{U}, \mathcal{D}\}$,

$$T_{\mathcal{B},i} = 0 \Leftrightarrow Y \text{ visits } \mathcal{B} \text{ less than } i \text{ times before absorption.}$$

The *number of repair periods during the finite time interval* $[0, t]$, denoted by $M_{\mathcal{D}}(t)$, can be thought of as the (undiscounted) cumulative repair cost during $[0, t]$ if every repair incident is associated with the same unit cost. $M_{\mathcal{D}}(t)$, $t > 0$, can be expressed in terms of the sequences of sojourn times as follows:

$$\begin{aligned} &\{M_{\mathcal{D}}(t) = m\} = \\ &\quad \{Y_0 \in \mathcal{U}, TS_{\mathcal{U},m} + TS_{\mathcal{D},m-1} < t \leq TS_{\mathcal{U},m+1} + TS_{\mathcal{D},m}\} \\ &\quad \cup \{Y_0 \in \mathcal{D}, TS_{\mathcal{U},m-1} + TS_{\mathcal{D},m-1} < t \leq TS_{\mathcal{U},m} + TS_{\mathcal{D},m}\}. \end{aligned} \tag{9.2}$$

We note in passing that a similar reasoning can also be used to express $CO(t)$, the cumulative operational time during $[0,t]$, in terms of the sojourn times of Y in $\mathcal{U}$ and $\mathcal{D}$ [98]. It is thus seen that knowledge of the operational and repair periods gives access to a number of other system reliability measures of interest. Whereas $M_{\mathcal{D}}(t)$ is defined for both irreducible and absorbing models, the a. s. limit of $M_{\mathcal{D}}(t)$ for $t \to +\infty$, that is the *total number of repair periods until final breakdown*, is a meaningful measure for absorbing Y only. Even though this quantity is within the scope of this review, it will not be given further consideration here, since it is expressible in terms of Y's embedded Markov chain and hence it can be analysed within the Markovian framework ([26], [29], Ch. 3). It is also customary to consider the expected values of some of the above dependability characteristics which are random variables. For example, if $Y_0 \in \mathcal{U}$, $\mathbb{E}(T_{\mathcal{U},1})$ is known as the *mean time to failure* (MTTF). Another example for a measure of this kind is the *fractional duration in the set of up states* $\mathcal{U}$, defined in [101] as the expected value of the interval availability, $FD_{\mathcal{U}}(t,t+x) = \mathbb{E}(IA(t,t+x))$. It is easily seen that the fractional duration can be represented in terms of the point availability as

$$FD_{\mathcal{U}}(t,t+x) = \frac{1}{x}\int_t^{t+x} PA(v)dv.$$

In the steady state analysis, mean values are used as $t \to +\infty$. For instance, the *limiting availability* [92] is defined as the limit of the point availability: $lim_{t\to+\infty} PA(t)$. (Note, however, that steady state analysis does not concern us here).

The integral equations to be considered for the above dependability measures will require Y's initial state to be specified. To this end, we shall use a subscript notation. For example,

$$R_s(t) = \mathbb{P}(Y_v \in \mathcal{U} \text{ for all } v \in [0,t] \mid Y_0 = s, Y_{0-} \neq s)$$

stands for the reliability function of Y at time t given that Y starts in $s \in S$ at time zero; these then form the entries of the reliability (column) vector $\mathbf{R}(t) = (R_s(t) : s \in \mathcal{S})$. A similar notation applies to the vector form of the other dependability measures.

It should be noted that an alternative formulation for two of the dependability measures is afforded by the framework of semi-Markov reward processes. A semi–Markov reward process (*e.g.*, [58], Ch. 13, [20], [97]) is obtained if we complement our semi–Markov process Y (now called the structure–state process) with a vector $\boldsymbol{\rho}$ of real reward rates, $\boldsymbol{\rho} = (\rho(s) : s \in \mathcal{S})$. $\rho(s)$ is interpreted as the rate at which reward is accumulated while Y resides in $s \in \mathcal{S}$. It is easily seen that by defining $\rho(s) = I_{\mathcal{U}}(s)$, the reward accumulated during $[0,t]$ equals $CO(t)$. Transition–based rewards (as opposed to reward rates) are used to explore event counts [44]. Then, Y is endowed with a real array $\mathbf{A} = (a_{s_1,s_2} : s_1, s_2 \in \mathcal{S})$ of 'lump rewards': the reward a_{s_1,s_2} accrues every time Y makes a transition from s_1 to s_2. Now, if we are interested in, for example, $M_{\mathcal{D}}(t)$, the number of repair periods during $[0,t]$, knowing that Y has started in the set of up states $\mathcal{U}$ at time 0, the array $\mathbf{A}$ of rewards will be thus defined as $a_{s_1,s_2} = I_{\mathcal{U}}(s_1)I_{\mathcal{D}}(s_2)$, and the reward accumulated during $[0,t]$ now equals $M_{\mathcal{D}}(t)$. The special case of Markov reward models is well covered in literature; we would like to single out the review papers [44] and [113] and the references therein.

9.4 Methods of Analysis

Having collected the most commonly used dependability measures in the previous section, we now embark upon their analysis. The purpose of this section is to exemplify the techniques of analysis by applying them to a selected set of dependability characteristics; the remaining results will be discussed (without proof) in the next section.

9.4.1 Continuous–parameter models

We start with the first sojourn times $T_{\mathcal{U},1}$ and $T_{\mathcal{D},1}$, since they will be seen later to be of fundamental importance. For $\mathcal{B} \in \{\mathcal{U}, \mathcal{D}\}$, let the events $\mathcal{E}(b)$, $b \in \mathcal{B}$, be defined by

$$\mathcal{E}(b) = \{\text{first entry of } Y \text{ into } \mathcal{B} \text{ is via } b \text{ and } Y \text{ starts at } t = 0 \text{ in } \mathcal{S}\backslash\mathcal{B}\}.$$

If, as we shall assume without loss of generality, the sample paths of Y are right–continuous, then a more formal definition of $\mathcal{E}(b)$ via the first sojourn times is

$$\mathcal{E}(b) = \{Y_{T_{\mathcal{S}\backslash\mathcal{B},1}} = b \text{ and } Y_0 \in \mathcal{S}\backslash\mathcal{B}, Y_{0-} \in \mathcal{B}\}.$$

We shall need two families of finite measures, $\{\kappa_{u,d} : u \in \mathcal{U}, d \in \mathcal{D}\}$ and $\{\kappa_{d,u} : u \in \mathcal{U}, d \in \mathcal{D}\}$, defined for (measurable) $A \subseteq [0, +\infty)$ by

$$\kappa_{d,u} = \mathbb{P}(\{T_{\mathcal{D},1} \in A\} \cap \mathcal{E}(u) \mid Y_0 = d, Y_{0-} \neq d), \tag{9.3}$$

$$\kappa_{u,d} = \mathbb{P}(\{T_{\mathcal{U},1} \in A\} \cap \mathcal{E}(d) \mid Y_0 = u, Y_{0-} \neq u). \tag{9.4}$$

The cdfs of $\kappa_{u,d}$ and $\kappa_{d,u}$ will be denoted by $k_{u,d}$ and $k_{d,u}$ respectively, *i.e.*, we put $k_{u,d}(t) = \kappa_{u,d}([0,t])$ and $k_{d,u}(t) = \kappa_{d,u}([0,t])$; these are then assembled to form the matrices $\mathbf{K}_{\mathcal{U}\mathcal{D}}$ and $\mathbf{K}_{\mathcal{D}\mathcal{U}}$. These two matrices will give access to a large number of dependability measures from Section 9.3. For example, it is easily seen that the *reliability* can be expressed as

$$\mathbf{R}_{\mathcal{U}}(t) = \mathbf{1} - \mathbf{K}_{\mathcal{U}\mathcal{D}}(t), \ \mathbf{R}_{\mathcal{D}}(t) = \mathbf{0}.$$

Similarly, the *point availability*, though not directly expressible in terms of the matrices $\mathbf{K}_{\mathcal{U}\mathcal{D}}(t)$ and $\mathbf{K}_{\mathcal{D}\mathcal{U}}(t)$, satisfies the following system of integral equations

$$\mathbf{PA}_{\mathcal{U}}(t) = \int_{[0,t]} \mathbf{K}_{\mathcal{U}\mathcal{D}}(dw)\mathbf{PA}_{\mathcal{D}}(t-w) + \mathbf{1} - \mathbf{K}_{\mathcal{U}\mathcal{D}}(t)\mathbf{1}, \tag{9.5}$$

$$\mathbf{PA}_{\mathcal{D}}(t) = \int_{[0,t]} \mathbf{K}_{\mathcal{D}\mathcal{U}}(dw)\mathbf{PA}_{\mathcal{U}}(t-w) + \mathbf{1} - \mathbf{K}_{\mathcal{D}\mathcal{U}}(t)\mathbf{1}. \tag{9.6}$$

The matrices $\mathbf{K}_{\mathcal{U}\mathcal{D}}(t)$ and $\mathbf{K}_{\mathcal{D}\mathcal{U}}(t)$ are themselves solutions of some integral equations; these are

$$\mathbf{K}_{\mathcal{U}\mathcal{D}}(t) = \int_{[0,t]} \mathbf{Q}_{\mathcal{U}\mathcal{U}}(dw)\mathbf{K}_{\mathcal{U}\mathcal{D}}(t-w) + \mathbf{Q}_{\mathcal{U}\mathcal{D}}(t), \tag{9.7}$$

$$\mathbf{K}_{\mathcal{D}\mathcal{U}}(t) = \int_{[0,t]} \mathbf{Q}_{\mathcal{D}\mathcal{D}}(dw)\mathbf{K}_{\mathcal{D}\mathcal{U}}(t-w) + \mathbf{Q}_{\mathcal{D}\mathcal{U}}(t). \tag{9.8}$$

Some explanation is now in order on what kind of integral is meant in (9.5)–(9.8). Since $\mathbf{K}_{\mathcal{U}\mathcal{D}}$ and $\mathbf{K}_{\mathcal{D}\mathcal{U}}$ are matrices of cdfs of finite measures on $[0, +\infty)$,

and considering that the same also holds for $\mathbf{Q}_{\mathcal{U}\mathcal{U}}$ and $\mathbf{Q}_{\mathcal{D}\mathcal{D}}$, the integrals in (9.5)–(9.8) can be thought of as Stieltjes integrals. Alternatively, they can be interpreted in the more general measure–theoretic framework.

Before proceeding further, we want to outline the proofs of (9.7), (9.8) and (9.5), (9.6), these being typical samples of proofs in the present framework.

Proof of (9.7) and (9.8). For $u \in \mathcal{U}$, $d \in \mathcal{D}$ and $t \geq 0$, the following holds according to the *renewal argument*

$$
\begin{aligned}
k_{u,d}(t) &= \mathbb{P}(\{T_{\mathcal{U},1} \leq t\} \cap \mathcal{E}(d) \mid Y \text{ starts in } u) \\
&= \sum_{\substack{s \in \mathcal{S} \\ s \neq u}} \mathbb{P}(\{T_{\mathcal{U},1} \leq t\} \cap \mathcal{E}(d) \mid X_0 = u, X_1 = s) p_{u,s} \\
&= \sum_{\substack{s \in \mathcal{U} \\ s \neq u}} \mathbb{P}(\{T_{\mathcal{U},1} \leq t\} \cap \mathcal{E}(d) \mid X_0 = u, X_1 = s) p_{u,s} \\
&\quad + \mathbb{P}(\{T_{\mathcal{U},1} \leq t\} \cap \mathcal{E}(d) \mid X_0 = u, X_1 = d) p_{u,d} \\
&= \sum_{\substack{s \in \mathcal{U} \\ s \neq u}} p_{u,s} \int_{[0,t]} \mathbb{P}(\{T_{\mathcal{U},1} \leq t - w\} \cap \mathcal{E}(d) \mid Y \text{ starts in } s) \\
&\qquad\qquad F_{u,s}(dw) + p_{u,d} F_{u,d}(t) \\
&= \sum_{\substack{s \in \mathcal{U} \\ s \neq u}} \int_{[0,t]} k_{s,d}(t - w) q_{u,s}(dw) + q_{u,d}(t) \\
&= \sum_{s \in \mathcal{U}} \int_{[0,t]} k_{s,d}(t - w) q_{u,s}(dw) + q_{u,d}(t).
\end{aligned}
$$

(The last step is by $q_{u,u} \equiv 0$.) Thus, $(k_{u,d}(t) : u \in \mathcal{U}, d \in \mathcal{D})$ satisfies (9.7). Equation (9.8) is shown along the same lines by interchanging the roles of $\mathcal{U}$ and $\mathcal{D}$.

■

Note that (9.7) and (9.8) are also available from [89], p. 45, where convolution equations are systematically explored for transition times between specific subsets of the state space of a finite semi–Markov process.

Proof of (9.5) and (9.6). For $u \in \mathcal{U}$, $d \in \mathcal{D}$ and $t \geq 0$, we have, again by the renewal argument,

$$
\begin{aligned}
PA_u(t) &= \mathbb{P}(Y_t \in \mathcal{U}, T_{\mathcal{U},1} \le t \mid Y \text{ starts in } u) \\
&\quad + \mathbb{P}(T_{\mathcal{U},1} > t \mid Y \text{ starts in } u) \\
&= \sum_{d \in \mathcal{D}} \mathbb{P}(\{Y_t \in \mathcal{U}, T_{\mathcal{U},1} \le t\} \cap \mathcal{E}(d) \mid Y \text{ starts in } u) \\
&\quad + 1 - \sum_{d \in \mathcal{D}} \mathbb{P}(T_{\mathcal{U},1} \le t\} \cap \mathcal{E}(d) \mid Y \text{ starts in } u) \\
&= \sum_{d \in \mathcal{D}} \int_{[0,t]} \mathbb{P}(Y_{t-w} \in \mathcal{U} \mid Y \text{ starts in } u) \\
&\qquad \mathbb{P}(\{T_{\mathcal{U},1} \in dw\} \cap \mathcal{E}(d) \mid Y \text{ starts in } u) + 1 - k_{u,d}(t) \\
&= \sum_{d \in \mathcal{D}} \int_{[0,t]} PA_u(t-w) k_{u,d}(dw) + 1 - k_{u,d}(t). \qquad (9.9)
\end{aligned}
$$

The matrix form of (9.9) is (9.5). Equation (9.6) is obtained from

$$
\begin{aligned}
PA_d(t) &= \mathbb{P}(Y_t \in \mathcal{U}, T_{\mathcal{D},1} \le t \mid Y \text{ starts in } d) \\
&= \sum_{u \in \mathcal{U}} \mathbb{P}(\{Y_t \in \mathcal{U}, T_{\mathcal{D},1} \le t\} \cap \mathcal{E}(u) \mid Y \text{ starts in } d) \\
&= \sum_{u \in \mathcal{U}} \int_{[0,t]} \mathbb{P}(Y_{t-w} \in \mathcal{U} \mid Y \text{ starts in } u) \\
&\qquad \mathbb{P}(\{T_{\mathcal{D},1} \in dw\} \cap \mathcal{E}(u) \mid Y \text{ starts in } d) \\
&= \sum_{u \in \mathcal{U}} \int_{[0,t]} PA_u(t-w) k_{u,d}(dw).
\end{aligned}
$$

■

There is a great deal of material available in the literature on the numerical solution of systems of integral equations such as (9.5)–(9.8); see, for example, the books [41], [42], [68] Ch. 4, and [75]. Most (but not all) of the systems considered here will be of the convolution type, which is encountered notably in renewal theory and more generally whenever the renewal argument is applicable. Boehme *et al.* [15] have recently discussed the use of the two–point trapezoidal rule for the numerical solution of integral equations of the convolution type in reliability theory. In one of the numerical examples discussed in Section 9.6 their method was used.

We shall now consider $TS_{\mathcal{U},m}$, the *total length of the first* $m \ge 1$ *working periods* of a repairable system; put $TS_{\mathcal{U},0} = 0$. Let $\mathbf{h}(.;m)$ denote the column vector of the cdfs of $TS_{\mathcal{U},m}$, $m \ge 0$, *i.e.*,

$$
\mathbf{h}(t;m) = (\mathbb{P}(TS_{\mathcal{U},m} \le t \mid Y_0 = s, Y_{0-} \ne s) : s \in \mathcal{S}), t \in [0, +\infty).
$$

Then, the following recurrence relation holds

$$
\mathbf{h}_{\mathcal{U}}(.;0) \equiv \mathbf{1}, \qquad (9.10)
$$

$$
\mathbf{h}_{\mathcal{U}}(t;m) =
$$

$$
\int_{[0,t]} \mathbf{K}_{\mathcal{U}\mathcal{D}}(dw)(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1} \mathbf{P}_{\mathcal{D}\mathcal{U}} \mathbf{h}_{\mathcal{U}}(t-w; m-1), \qquad m \ge 1. \qquad (9.11)
$$

Furthermore,

$$
\mathbf{h}_{\mathcal{D}}(t;m) = (\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1} \mathbf{P}_{\mathcal{D}\mathcal{U}} \mathbf{h}_{\mathcal{U}}(t;m), \qquad m \ge 0. \qquad (9.12)
$$

Equations (9.10)–(9.12) hold of course only if the system under consideration is repairable. (Otherwise, $(\mathbf{I}-\mathbf{P}_{\mathcal{D}\mathcal{D}})$ is not invertible since Y's embedded Markov chain is absorbing.)

Proof of (9.10)–(9.12). We start with the proof of (9.12). Equation (9.12) holds of course for $m = 0$ since $\mathbf{P}$ is a stochastic matrix and thus $(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\mathbf{1} = \mathbf{1}$. Equation (9.12) is seen to hold for $m \geq 1$ by

$$\begin{aligned} h_d(t;m) &= \sum_{u\in\mathcal{U}} \mathbb{P}(\{TS_{\mathcal{U},m} \leq t\} \cap \mathcal{E}(u) \mid Y \text{ starts in } d) \\ &= \sum_{u\in\mathcal{U}} h_u(t;m)\mathbb{P}(\mathcal{E}(u) \mid Y \text{ starts in } d), \end{aligned}$$

and

$$\begin{aligned} &(\mathbb{P}(\mathcal{E}(u) \mid Y \text{ starts in } d) : d \in \mathcal{D}, u \in \mathcal{U}) = \\ &\mathbf{P}_{\mathcal{D}\mathcal{U}} + \mathbf{P}_{\mathcal{D}\mathcal{D}}\mathbf{P}_{\mathcal{D}\mathcal{U}} + \mathbf{P}_{\mathcal{D}\mathcal{D}}^2\mathbf{P}_{\mathcal{D}\mathcal{U}} + \cdots = (\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}. \end{aligned} \tag{9.13}$$

Equation (9.10) holds by $TS_{\mathcal{U},m} \equiv 0$. For $m = 1$, (9.11) holds by

$$\begin{aligned} h_u(t;1) &= \sum_{d\in\mathcal{D}} \mathbb{P}(\{T_{\mathcal{U},1} \leq t\} \cap \mathcal{E}(d) \mid Y \text{ starts in } u) \\ &= \sum_{d\in\mathcal{D}} k_{u,d}(t) = \mathbf{K}_{\{u\},\mathcal{D}}(t)\mathbf{1} \\ &= \int_{[0,t]} \mathbf{K}_{\{u\}\mathcal{D}}(dw)(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\mathbf{h}_{\mathcal{U}}(t-w;0). \end{aligned} \tag{9.14}$$

(The last step in (9.14) follows from (9.10) and from the fact that $\mathbf{P}$ is a stochastic matrix.) By the renewal argument, we have for $m \geq 2$,

$$\begin{aligned} h_u(t;m) &= \mathbb{P}\,(TS_{\mathcal{U},m} \leq t \mid Y \text{ starts in } u) \\ &= \sum_{d\in\mathcal{D}} \mathbb{P}(\{T_{\mathcal{U},1} + \sum_{\ell=2}^{m} T_{\mathcal{U},\ell} \leq t\} \cap \mathcal{E}(d) \mid Y \text{ starts in } u) \\ &= \sum_{d\in\mathcal{D}} \int_{[0,t]} \mathbb{P}(TS_{\mathcal{U},m-1} \leq t - w \mid Y \text{ starts in } d) \\ &\qquad\qquad \mathbb{P}(\{T_{\mathcal{U},1} \in dw\} \cap \mathcal{E}(d) \mid Y \text{ starts in } u) \\ &= \sum_{d\in\mathcal{D}} \int_{[0,t]} h_d(t-w;m-1)\kappa_{u,d}(dw), \end{aligned}$$

which in matrix form is written as

$$\mathbf{h}_{\mathcal{U}}(t;m) = \int_{[0,t]} \mathbf{K}_{\mathcal{U}\mathcal{D}}(dw)\mathbf{h}_{\mathcal{D}}(t-w;m-1). \tag{9.15}$$

Equation (9.15) with (9.12) gives (9.11).

■

We have already shown how the length of the first working period, $T_{\mathcal{U},1}$, is related to the $\mathcal{U}$–entries of the reliability vector $\mathbf{R}(t)$ and how this can be expressed in terms of the measures defined in (9.3) and (9.4). These measures can be used to express the cdf of the mth *working period*, $T_{\mathcal{U},m}$, $m \geq 1$. For the *first* working period, which is identical to the cumulative working time for $m = 1$, we have, according to (9.10)–(9.12),

$$(\mathbb{P}(T_{\mathcal{U},1} \le t \mid Y \text{ starts in } u) : u \in \mathcal{U}) =$$
$$\int_{[0,t]} \mathbf{K}_{\mathcal{U}\mathcal{D}}(dw)(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\mathbf{1} = \mathbf{K}_{\mathcal{U}\mathcal{D}}\mathbf{1}. \tag{9.16}$$

For $m \ge 2$ and $u \in \mathcal{U}$, it is

$$\mathbb{P}(T_{\mathcal{U},m} \le t \mid Y \text{ starts in } u) =$$
$$\sum_{s\in\mathcal{D}} \mathbb{P}(T_{\mathcal{U},m-1} \le t \mid Y \text{ starts in } s)\mathbb{P}(\mathcal{E}(s) \mid Y \text{ starts in } u) =$$
$$\sum_{s_1\in\mathcal{D}} \sum_{s_2\in\mathcal{U}} \mathbb{P}(T_{\mathcal{U},m-1} \le t\} \mid Y \text{ starts in } s_2)$$
$$\mathbb{P}(\mathcal{E}(s_2) \mid Y \text{ starts in } s_1)\mathbb{P}(\mathcal{E}(s_1) \mid Y \text{ starts in } u),$$

from which it follows by (9.13) that

$$(\mathbb{P}(T_{\mathcal{U},m} \le t \mid Y \text{ starts in } u) : u \in \mathcal{U}) =$$
$$(\mathbf{I} - \mathbf{P}_{\mathcal{U}\mathcal{U}})^{-1}\mathbf{P}_{\mathcal{U}\mathcal{D}}(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}$$
$$(\mathbb{P}(T_{\mathcal{U},m-1} \le t \mid Y \text{ starts in } u) : u \in \mathcal{U}). \tag{9.17}$$

By induction, we get from (9.16) and (9.17),

$$(\mathbb{P}(T_{\mathcal{U},m} \le t \mid Y \text{ starts in } u) : u \in \mathcal{U}) =$$
$$\left((\mathbf{I} - \mathbf{P}_{\mathcal{U}\mathcal{U}})^{-1}\mathbf{P}_{\mathcal{U}\mathcal{D}}(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\right)^{m-1}\mathbf{K}_{\mathcal{U}\mathcal{D}}\mathbf{1}, \qquad m \ge 1. \tag{9.18}$$

Similarly, for $m \ge 1$, $d \in \mathcal{D}$, it is

$$\mathbb{P}(T_{\mathcal{U},m} \le t \mid Y \text{ starts in } d) =$$
$$\sum_{s\in\mathcal{U}} \mathbb{P}(T_{\mathcal{U},m} \le t \mid Y \text{ starts in } s)\mathbb{P}(\mathcal{E}(s) \mid Y \text{ starts in } d),$$

from which it follows, again by (9.13), that

$$(\mathbb{P}(T_{\mathcal{U},m} \le t \mid Y \text{ starts in } d) : d \in \mathcal{D}) =$$
$$(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\,(\mathbb{P}(T_{\mathcal{U},m} \le t \mid Y \text{ starts in } u) : u \in \mathcal{U}). \tag{9.19}$$

Equations (9.18) and (9.19) show that

$$(\mathbb{P}(T_{\mathcal{U},m} \le t \mid Y \text{ starts in } d) : d \in \mathcal{D}) = (\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}$$
$$\left((\mathbf{I} - \mathbf{P}_{\mathcal{U}\mathcal{U}})^{-1}\mathbf{P}_{\mathcal{U}\mathcal{D}}(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\right)^{m-1}\mathbf{K}_{\mathcal{U}\mathcal{D}}\mathbf{1}, \qquad m \ge 1. \tag{9.20}$$

Equations (9.18) and (9.20) show once again the instrumental role played by the measures in (9.3) and (9.4).

Rewritten in the form of one single integral equation, recurrence relations like (9.10)–(9.11) involve high–dimensional Stieltjes integrals which are awkward to evaluate numerically. In such cases, it may still be possible to solve the system in the Laplace transform domain and then obtain a computational solution by numerical Laplace transform inversion. Let $\mathbf{K}^*_{\mathcal{U}\mathcal{D}}$ and $\mathbf{K}^*_{\mathcal{D}\mathcal{U}}$ stand for the matrix of Laplace transforms of the measures defined in (9.3) and (9.4), *i.e.*,

$$\mathbf{K}^*_{\mathcal{A}\mathcal{B}}(\tau) = \int_{[0,+\infty)} \mathbf{K}_{\mathcal{A}\mathcal{B}}(dt)\exp(-\tau t);\ \mathcal{A},\mathcal{B} \in \{\mathcal{U},\mathcal{D}\}, \qquad \mathcal{A} \ne \mathcal{B}.$$

(The entries of the above matrices are of course the Laplace–Stieltjes transforms of the corresponding cdfs.) Furthermore, let $\mathbf{h}^*(\tau; m)$ stand for the vector

of Laplace transform of the random variable $TS_{\mathcal{U},m}$, *i.e.*, we put for $m \geq 0$ and $\tau \in \mathbb{C}$, $\Re(\tau) > 0$,

$$\mathbf{h}^*(\tau; m) = \big(\mathbb{E}\left(\exp(-\tau TS_{\mathcal{U},m}) \mid Y \text{ starts in } s\right) : s \in \mathcal{S}\big).$$

Then, (9.10)–(9.12) translates to

$$\begin{aligned}
&\mathbf{h}^*_{\mathcal{U}}(\tau; 0) \equiv \mathbf{1},\\
&\mathbf{h}^*_{\mathcal{U}}(\tau; m) = \mathbf{K}^*_{\mathcal{U}\mathcal{D}}(\tau)(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\mathbf{h}^*_{\mathcal{U}}(\tau; m-1), \qquad m \geq 1,\\
&\mathbf{h}^*_{\mathcal{D}}(\tau; m) = (\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\mathbf{h}^*_{\mathcal{U}}(\tau; m), \qquad m \geq 0.
\end{aligned}$$

Thus, for $m \geq 0$

$$\mathbf{h}^*_{\mathcal{U}}(\tau; m) = \big(\mathbf{K}^*_{\mathcal{U}\mathcal{D}}(\tau)(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\big)^m \mathbf{1}, \tag{9.21}$$

$$\mathbf{h}^*_{\mathcal{D}}(\tau; m) = (\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\big(\mathbf{K}^*_{\mathcal{U}\mathcal{D}}(\tau)(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\big)^m \mathbf{1}. \tag{9.22}$$

By (9.7), we have of course the matrix $\mathbf{K}^*_{\mathcal{U}\mathcal{D}}(\tau)$ expressed in terms of the Laplace transform of the kernel matrix $\mathbf{Q}$,

$$\mathbf{Q}^*(\tau) = \int_{[0,+\infty)} \mathbf{Q}(dt)\exp(-\tau t),$$

as

$$\mathbf{K}^*_{\mathcal{U}\mathcal{D}}(\tau) = (\mathbf{I} - \mathbf{Q}^*_{\mathcal{U}\mathcal{U}})^{-1}\mathbf{Q}^*_{\mathcal{U}\mathcal{D}}. \tag{9.23}$$

The system of equations (9.21)–(9.23) represent a closed form solution in the Laplace transform domain for the total length of the first m cumulative working periods. An exhaustive treatment of the sojourn time vector in the Laplace transform domain can be found in [22] and [29], Ch. 8.

Many methods exist for the numerical inversion of the Laplace transform; we would like to list the following references: [2], [10], [11], Ch. 6, [19], [21], [45], [46], [47] Chapter VII.6, [48], [49], [50], [61], [62], [76], [77], [78], [79], [102], [105], [106], [107], [108], Chapters 8 and 9, [111], and [115]. In the reliability/dependability setting, the methods from [102] and [105] have been employed in [18] and [71] respectively. In the author's own work [25], [29], Ch. 10, a method due to Weeks [115] has been employed; it is based upon a Laguerre polynomial representation of the original function whose Laguerre coefficients are inferred from the Laplace transform by using fast Fourier transform. (This method is also implemented in the widely used NAG Fortran library [90].) It must be added, however, that according to [11], Section 1.10, no single method can be devised which will perform numerical Laplace transform inversion to a given accuracy, since the inverse of the Laplace transform is unstable under small perturbations.

We have thus far outlined the numerical procedures available for the analysis of continuous–parameter semi–Markov reliability models. In most cases, it will also be possible to establish a formal closed form solution of the integral equations under consideration. For this, we shall need the notion of the Stieltjes convolution of matrix–valued functions on $[0, +\infty)$. Let $\mathbf{M}$ be a matrix of cdfs of finite measures on $[0, +\infty)$. Furthermore, let $\mathbf{N}$ be a (compatible) matrix of measurable functions on $[0, +\infty)$; for our purposes it suffices to assume that the entries of $\mathbf{N}$ are bounded functions. Then, $\mathbf{M} * \mathbf{N}$, the Stieltjes convolution of $\mathbf{M}$ and $\mathbf{N}$, is defined for $t \in [0, +\infty)$ by the Stieltjes integral

$$(\mathbf{M} * \mathbf{N})(t) = \int_{[0,t)} \mathbf{M}(dw)\mathbf{N}(t - w). \tag{9.24}$$

The (both right and left) neutral element of the Stieltjes convolution is $\mathbf{id} = \big(I_{[0,+\infty)}\mathbf{I} : \geq 0\big)$. For a square matrix of cdfs, $\mathbf{M}$ say, repeated, i–fold convolution is defined recursively by $\mathbf{M}^{*(0)} = \mathbf{id}$, $\mathbf{M}^{*(i+1)} = \mathbf{M}^{*(i)} * \mathbf{M}$. With this notation then, the formal solution of (9.5)–(9.8) is as follows

$$\mathbf{PA}_{\mathcal{U}} = \sum_{i=0}^{\infty} \big((\mathbf{K}_{\mathcal{UD}} * \mathbf{K}_{\mathcal{DU}})^{*(i+1)}\,\mathbf{1} - \mathbf{K}_{\mathcal{UD}} * (\mathbf{K}_{\mathcal{DU}} * \mathbf{K}_{\mathcal{UD}})^{*(i+1)}\,\mathbf{1}\big) + \mathbf{1} - \mathbf{K}_{\mathcal{UD}}\mathbf{1}, \tag{9.25}$$

$$\mathbf{PA}_{\mathcal{D}} = \sum_{i=0}^{\infty} \big((\mathbf{K}_{\mathcal{DU}} * \mathbf{K}_{\mathcal{UD}})^{*(i)} * \mathbf{K}_{\mathcal{DU}}\mathbf{1} - (\mathbf{K}_{\mathcal{DU}} * \mathbf{K}_{\mathcal{UD}})^{*(i+1)}\,\mathbf{1}\big), \tag{9.26}$$

where

$$\mathbf{K}_{\mathcal{UD}} = \sum_{i=0}^{\infty} \mathbf{Q}_{\mathcal{UU}}{}^{*(i)}\mathbf{Q}_{\mathcal{UD}}, \tag{9.27}$$

$$\mathbf{K}_{\mathcal{DU}} = \sum_{i=0}^{\infty} \mathbf{Q}_{\mathcal{DD}}{}^{*(i)}\mathbf{Q}_{\mathcal{DU}}. \tag{9.28}$$

Proof of (9.25)–(9.28). Write (9.7) as

$$\mathbf{K}_{\mathcal{UD}} = \mathbf{Q}_{\mathcal{UU}} * \mathbf{K}_{\mathcal{UD}} + \mathbf{Q}_{\mathcal{UD}},$$

from which, by induction, it follows for $n \geq 1$ that

$$\mathbf{K}_{\mathcal{UD}} = \mathbf{Q}_{\mathcal{UU}}{}^{*(n)} * \mathbf{K}_{\mathcal{UD}} + \sum_{i=0}^{n-1} \mathbf{Q}_{\mathcal{UU}}{}^{*(i)} * \mathbf{Q}_{\mathcal{UD}}. \tag{9.29}$$

Equation (9.27) follows from (9.29) by $n \to +\infty$. (9.28) is deduced from (9.8) in a similar fashion. To show (9.25) and (9.26), we write (9.5) and (9.6) as

$$\mathbf{PA}_{\mathcal{U}} = \mathbf{K}_{\mathcal{UD}} * \mathbf{PA}_{\mathcal{D}} + \mathbf{1} - \mathbf{K}_{\mathcal{UD}}\mathbf{1}, \tag{9.30}$$

$$\mathbf{PA}_{\mathcal{D}} = \mathbf{K}_{\mathcal{DU}} * \mathbf{PA}_{\mathcal{U}}. \tag{9.31}$$

Using (9.31), it follows upon premultiplication of (9.30) by $\mathbf{K}_{\mathcal{DU}}$ that

$$\mathbf{PA}_{\mathcal{D}} = \mathbf{K}_{\mathcal{DU}} * \mathbf{K}_{\mathcal{UD}} * \mathbf{PA}_{\mathcal{D}} + \mathbf{K}_{\mathcal{DU}}\mathbf{1} - \mathbf{K}_{\mathcal{DU}} * \mathbf{K}_{\mathcal{UD}}\mathbf{1}. \tag{9.32}$$

By induction, we get from (9.32) for $n \geq 1$,

$$\mathbf{PA}_{\mathcal{D}} = (\mathbf{K}_{\mathcal{DU}} * \mathbf{K}_{\mathcal{UD}})^{*(n)} * \mathbf{PA}_{\mathcal{D}} + \sum_{i=0}^{n-1} \big((\mathbf{K}_{\mathcal{DU}} * \mathbf{K}_{\mathcal{UD}})^{*(i)} * \mathbf{K}_{\mathcal{DU}}\mathbf{1} - (\mathbf{K}_{\mathcal{DU}} * \mathbf{K}_{\mathcal{UD}})^{*(i+1)}\,\mathbf{1}\big). \tag{9.33}$$

Equation (9.26) now follows from (9.33) by $n \to +\infty$. (9.25) readily follows from (9.26) and (9.30).

■

Let us conclude this section by observing that if Y is an alternating renewal process and thus the system being modelled comprises one repairable unit, it is $\mathcal{S} = \{u, d\}$, and $\mathbf{K}_{\mathcal{UD}} = F_{u,d}$ and $\mathbf{K}_{\mathcal{DU}} = F_{d,u}$ are respectively the time to failure and repair time cdf.

9.4.2 Discrete–parameter models

The results for each quantity in the discrete–parameter case will of course look very similar to the corresponding continuous–parameter dependability measure. However, in the discrete–parameter formulation, the schemes for the solution of the equations will arise in a natural fashion from the recurrence relations defining those equations. To illustrate the point, let us examine the mth working period, $T_{\mathcal{U},m}$. In analogy to (9.18) and (9.20), we now have the following: for $t = 0, 1, \cdots$ it is

$$(\mathbb{P}(T_{\mathcal{U},m} \leq t \mid Y \text{ starts in } u) : u \in \mathcal{U}) = \left((\mathbf{I} - \mathbf{P}_{\mathcal{U}\mathcal{U}})^{-1}\mathbf{P}_{\mathcal{U}\mathcal{D}}(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\right)^{m-1}\mathbf{L}_{\mathcal{U}\mathcal{D}}\mathbf{1}, \qquad m \geq 1, \tag{9.34}$$

$$(\mathbb{P}(T_{\mathcal{U},m} \leq t \mid Y \text{ starts in } d) : d \in \mathcal{D}) = (\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}} \left((\mathbf{I} - \mathbf{P}_{\mathcal{U}\mathcal{U}})^{-1}\mathbf{P}_{\mathcal{U}\mathcal{D}}(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\right)^{m-1}\mathbf{L}_{\mathcal{U}\mathcal{D}}\mathbf{1}, \qquad m \geq 1, \tag{9.35}$$

where $\mathbf{L}_{\mathcal{U}\mathcal{D}}$ is a matrix–valued function on $\{0, 1, \cdots\}$ satisfying the recurrence relation

$$\mathbf{L}_{\mathcal{U}\mathcal{D}}(t) = \sum_{i=0}^{t-1} \mathbf{H}_{\mathcal{U}\mathcal{U}}(t-i)\mathbf{L}_{\mathcal{U}\mathcal{D}}(i) + \sum_{i=0}^{t} \mathbf{H}_{\mathcal{U}\mathcal{D}}(i), \qquad t = 0, 1, \cdots. \tag{9.36}$$

The correspondence between (9.7) and (9.36) is obvious. It is also clear that (9.36) can be used directly to compute the matrices $\mathbf{L}_{\mathcal{U}\mathcal{D}}(0)$, $\mathbf{L}_{\mathcal{U}\mathcal{D}}(1)$, $\cdots$. Thus, the discrete–parameter framework leads us to a rather straightforward computational implementation, whereas in the continuous–parameter case we arrive at a system of integral equation; therefore, from a practitioner's point of view, the discrete–parameter model may well be the preferred option.

Proof of (9.34)–(9.36). In view of the fact that the discrete–parameter case can be thought of as a special case within the continuous–parameter framework, (9.34) and (9.35) are seen to hold by (9.18) and (9.20), respectively, if we put

$$\mathbf{L}_{\mathcal{U}\mathcal{D}}(i) = \mathbf{K}_{\mathcal{U}\mathcal{D}}(t) \text{ for } t \in [i, i+1), \qquad i = 0, 1, \cdots.$$

Using

$$q_{s_1,s_2}(i) = p_{s_1,s_2} \sum_{j=0}^{i} f_{s_1,s_2}(j) = \sum_{j=0}^{i} h_{s_1,s_2}(j) \qquad , s_1, s_2 \in \mathcal{S},$$

we can write (9.7) as

$$\begin{aligned}\mathbf{L}_{\mathcal{U}\mathcal{D}}(i) &= \sum_{j=0}^{i} \mathbf{H}_{\mathcal{U}\mathcal{U}}(j)\mathbf{L}_{\mathcal{U}\mathcal{D}}(i-j) + \sum_{j=0}^{i} \mathbf{H}_{\mathcal{U}\mathcal{D}}(j) \\ &= \sum_{j=0}^{i} \mathbf{H}_{\mathcal{U}\mathcal{U}}(i-j)\mathbf{L}_{\mathcal{U}\mathcal{D}}(j) + \sum_{j=0}^{i} \mathbf{H}_{\mathcal{U}\mathcal{D}}(j).\end{aligned}$$

This is identical to (9.36) since $\mathbf{H}_{\mathcal{U}\mathcal{U}}(0) = \mathbf{0}$.

■

We conclude this section by noting that all discrete–parameter results are obtainable independently from their continuous–parameter counterpart by reworking through their proofs via the renewal argument. As indicated earlier, this approach may indeed be more appropiate in a context where the problem domain calls for discrete–parameter modelling and if the modeller is less interested in mathematical niceties and, finally, if a computational implementation is the modeller's priority; see, [58], Ch. 10, [84] and [85]. There is also the possibility of analysing the convolution equations for discrete–parameter semi–Markov models by means of generating functions. Since this is well explained in [58] (there called 'geometric transform'), this topic will not be covered here.

9.5 Equations for the Dependability Measures

We have dealt with four of the measures from Section 9.3 already: the reliability $\mathbf{R}$, the point availability $\mathbf{PA}$, the length of the mth working period $T_{\mathcal{U},m}$, and the total length of the first m working periods $TS_{\mathcal{U},m}$. Let us now address the remaining measures in turn.

For the *interval reliability*, we have the following integral equations [34]

$$\begin{aligned}\mathbf{IR}_{\mathcal{U}}(x,t) = &\int_{[0,t]} \mathbf{K}_{\mathcal{UD}}(dw)\mathbf{IR}_{\mathcal{D}}(x,t-w) \\ &+\mathbf{1} - \mathbf{K}_{\mathcal{UD}}(t+x)\mathbf{1}. \qquad (9.37)\end{aligned}$$

$$\mathbf{IR}_{\mathcal{D}}(x,t) = \int_{[0,t]} \mathbf{K}_{\mathcal{DU}}(dw)\mathbf{IR}_{\mathcal{U}}(x,t-w). \qquad (9.38)$$

The *joint availability* satisfies the following system [31],

$$\begin{aligned}\mathbf{JA}_{\mathcal{U}}(x,t) = &\int_{[0,t]} \mathbf{K}_{\mathcal{UD}}(dw)\mathbf{JA}_{\mathcal{D}}(x,t-w) \\ &+\int_{(t,t+x]} \mathbf{K}_{\mathcal{UD}}(dw)\mathbf{PA}_{\mathcal{D}}(t+x-w) \\ &+\mathbf{1} - \mathbf{K}_{\mathcal{UD}}(t+x)\mathbf{1}, \qquad (9.39)\end{aligned}$$

$$\mathbf{JA}_{\mathcal{D}}(x,t) = \int_{[0,t]} \mathbf{K}_{\mathcal{DU}}(dw)\mathbf{JA}_{\mathcal{U}}(x,t-w). \qquad (9.40)$$

Note that prior to solving the system (9.39)–(9.40), we need the point availabilities from the system (9.5)–(9.6).

For the *joint interval reliability*, we have [35]

$$\begin{aligned}&\mathbf{JIR}_{\mathcal{U}}(x_1,t_1,x_2,t_2) = \\ &\int_{[0,t_1)} \mathbf{K}_{\mathcal{UD}}(dw)\mathbf{JIR}_{\mathcal{D}}(x_1,t_1-w,x_2,t_2-w) \\ &+\int_{(t_1+x_1,t_2)} \mathbf{K}_{\mathcal{UD}}(dw)\mathbf{IR}_{\mathcal{D}}(x_2,t_2-w) \\ &+\mathbf{1} - \mathbf{K}_{\mathcal{UD}}(t_2+x_2)\mathbf{1}, \qquad (9.41)\end{aligned}$$

$$\begin{aligned}&\mathbf{JIR}_{\mathcal{D}}(x_1,t_1,x_2,t_2) = \\ &\int_{[0,t_1]} \mathbf{K}_{\mathcal{DU}}(dw)\mathbf{JIR}_{\mathcal{U}}(x_1,t_1-w,x_2,t_2-w). \qquad (9.42)\end{aligned}$$

Note that now, in order to solve the system (9.41)–(9.42), we need the interval reliability from (9.37)–(9.38).

The set reliability [35], of which the last three are instances, satisfies the following system of integral equations

$$\mathbf{SR}_{\mathcal{U}}(T) = \int_{[0,+\infty)\setminus T} \mathbf{K}_{\mathcal{U}\mathcal{D}}(dw)\mathbf{SR}_{\mathcal{D}}((T-w)\cap[0,+\infty)),$$

$$\mathbf{SR}_{\mathcal{D}}(T) = \int_{[0,\inf(T)]} \mathbf{K}_{\mathcal{D}\mathcal{U}}(dw)\mathbf{SR}_{\mathcal{U}}(T),$$

where for $T \subset [0,+\infty)$, $w \in [0,+\infty)$, we put $T - w = \{t - w : t \in T\}$.

Before proceeding futher, let us add that systems of integral equations like the above, when written in the Laplace transform domain, can be used with a Tauberian argument to explore the steady state behaviour of the reliability measure under consideration (see, for example, [30] for corresponding results for the interval reliability).

Let us now turn our attention to the *cumulative up–time until final breakdown*, C. The assumption here is that Y is absorbing and its state space $\mathcal{S}$ is partitioned as $\mathcal{S} = \mathcal{U} \cup \mathcal{D} \cup \{\omega\}$. Results in the Laplace transform domain on the distribution of C can be found in [20] and [70]. (The discussions in [70] and [20] are set in the more general semi–Markov *reward* context. However, as far as the performance measure C is concerned, considering the accumulated reward until absorption does not entail more generality since, as shown in [20], the latter can be written by a 'change of pace' technique as a cumulative up–time variable C of a modified semi–Markov process.) In what follows, we want to concentrate on the evaluation of the cdf of C in the *discrete–parameter* framework. Then, C is defined by

$$C = \sum_{i=0}^{\infty} I_{\{Y_{\delta_i} \in \mathcal{U}\}},$$

and by [33] we have for

$$\boldsymbol{\phi}(t) = (\mathbb{P}(C \le t \mid Y_{\delta_0} = s, Y_{\delta_{-1}} \ne s) : s \in \mathcal{S}), \qquad t = 0, 1, \cdots,$$

the vector of cdfs of C, the following

$$\boldsymbol{\phi}_{\mathcal{U}}(0) = \mathbf{0}, \tag{9.43}$$

$$\begin{aligned}\boldsymbol{\phi}_{\mathcal{U}}(t) = &\sum_{i=1}^{t}(\mathbf{H}_{\mathcal{U}\mathcal{U}}(i) + \mathbf{H}_{\mathcal{U}\mathcal{D}}(i)(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}})\boldsymbol{\phi}_{\mathcal{U}}(t-i) \\ &+ \sum_{i=1}^{t} \mathbf{H}_{\mathcal{U}\mathcal{D}}(i)(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}\mathbf{1} \\ &+ \sum_{i=1}^{t} \mathbf{H}_{\mathcal{U}\{\omega\}}(i)\mathbf{1}, \qquad t = 1, 2, \cdots,\end{aligned} \tag{9.44}$$

$$\boldsymbol{\phi}_{\mathcal{D}}(t) = (\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}(\mathbf{P}_{\mathcal{D}\mathcal{U}}\boldsymbol{\phi}_{\mathcal{U}}(t) + \mathbf{P}_{\mathcal{D}\{\omega\}}\mathbf{1}), \qquad t = 0, 1, \cdots, \tag{9.45}$$

$$\boldsymbol{\phi}_{\{\omega\}} \equiv \mathbf{1}. \tag{9.46}$$

By (9.43)–(9.46) a recurrence relation is defined which can be used for the numerical evaluation of $\boldsymbol{\phi}$. Equations (9.43)–(9.46) also give rise to a formal closed form expression for $\boldsymbol{\phi}$. To see this, we define the (ordinary) convolution of two matrix–valued functions $\mathbf{M}$ and $\mathbf{N}$, defined on the non–negative integers, by

$$(\mathbf{M} * \mathbf{N})(t) = \sum_{i=1}^{t} \mathbf{M}(i)\mathbf{N}(t-i), \qquad t = 0, 1, \cdots.$$

The neutral element of the convolution operation is

$$\mathbf{id}(t) = \begin{cases} \mathbf{I}, \text{ for } t = 0, \\ \mathbf{0}, \text{ for } t \geq 1. \end{cases}$$

With this notation then, (9.43)–(9.44) imply the following *closed form representation* of $\phi_{\mathcal{U}}$

$$\phi_{\mathcal{U}} = \sum_{n=0}^{\infty} \mathbf{A}_{\mathcal{U}\mathcal{U}}{}^{*(n)} * \Big(\mathbf{H}_{\mathcal{U}\mathcal{D}} * \big((\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\{\omega\}}\mathbf{1}\big) + \mathbf{H}_{\mathcal{U}\{\omega\}} * \mathbf{1}\Big), \tag{9.47}$$

where $\mathbf{A}_{\mathcal{U}\mathcal{U}}(t) = \mathbf{H}_{\mathcal{U}\mathcal{U}}(t) + \mathbf{H}_{\mathcal{U}\mathcal{D}}(t)(\mathbf{I} - \mathbf{P}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{P}_{\mathcal{D}\mathcal{U}}$, $t = 0, 1, \cdots$.

The next dependability measure to be examined is the *cumulative operational time* $CO(t)$. For the continuous-parameter case, double-Laplace transform results are known under both the Markov and semi-Markov assumptions [103], [70], *i.e.* a closed form expression is available for the double-Laplace transform

$$\int_0^{+\infty}\int_0^{+\infty} \exp(-\tau_1 t_1 - \tau_2 t_2)\mathbf{f}(t_1, t_2)dt_1 dt_2,$$

$(\tau_1, \tau_2 \in \mathbb{C}, \Re(\tau_1), \Re(\tau_2) > 0)$, where for $t_1, t_2 \in [0, +\infty)$,

$$\mathbf{f}(t_1, t_2) = (\mathbb{P}(CO(t_1) \leq t_2 \mid Y_0 = s, Y_{0-} \neq s) : s \in \mathcal{S}). \tag{9.48}$$

For repairable systems, the corresponding integral equations in the time domain read as follows [37]

$$\mathbf{f}(t_1, t_2) = \mathbf{1}, \qquad 0 \leq t_1 \leq t_2 < +\infty, \tag{9.49}$$

$$\mathbf{f}_{\mathcal{U}}(t_1, t_2) = \int_{[0,t_2]} \mathbf{K}_{\mathcal{U}\mathcal{D}}(dw)\mathbf{f}_{\mathcal{D}}(t_1 - w, t_2 - w), \qquad 0 \leq t_2 < t_1 < +\infty, \tag{9.50}$$

$$\mathbf{f}_{\mathcal{D}}(t_1, t_2) = \int_{[0,t_1-t_2]} \mathbf{K}_{\mathcal{D}\mathcal{U}}(dw)\mathbf{f}_{\mathcal{U}}(t_1 - w, t_2) + \mathbf{1} - \mathbf{K}_{\mathcal{D}\mathcal{U}}(t_1 - t_2)\mathbf{1}, \ 0 \leq t_2 < t_1 < +\infty. \tag{9.51}$$

Note that (9.50) and (9.51) are not convolution equations and thus it will not be possible to find a formal closed form expression for the cdf of the cumulative operational time. In the discrete-parameter case, the cumulative operational time is written as

$$CO(t) = \sum_{i=0}^{t} I_{\{Y_{\delta_i} \in \mathcal{U}\}}, \qquad t \in \{0, 1, \cdots\},$$

and its cdf can be computed from the following recurrence relation [28]

$$\phi_{\mathcal{U}}(t_1, t_2) = \begin{cases} \mathbf{0}, & \text{if } t_2 = 0, \\ \sum_{i=1}^{t_2} \mathbf{H}_{\mathcal{U}\mathcal{U}}(t)\phi_{\mathcal{U}}(t_1 - t, t_2 - t) & \\ + \sum_{i=1}^{t_2} \mathbf{H}_{\mathcal{U}\mathcal{D}}(t)\phi_{\mathcal{D}}(t_1 - t, t_2 - t), & \text{if } t_1 \geq t_2 \geq 1, \\ \mathbf{1}, & \text{if } t_2 > t_1, \end{cases} \tag{9.52}$$

and

$$\boldsymbol{\phi}_{\mathcal{D}}(t_1,t_2)=\begin{cases}\mathbf{0}, & \text{if } t_1=t_2=0,\\ \sum_{i=1}^{t_1}\mathbf{H}_{\mathcal{D}\mathcal{U}}(t)\boldsymbol{\phi}_{\mathcal{U}}(t_1-t,t_2) & \\ +\sum_{i=1}^{t_1}\mathbf{H}_{\mathcal{D}\mathcal{D}}(t)\boldsymbol{\phi}_{\mathcal{D}}(t_1-t,t_2), & \text{if } t_1\geq t_2 \text{ and}\\ & (t_1,t_2)\neq(0,0),\\ \mathbf{1}, & \text{if } t_2>t_1,\end{cases} \tag{9.53}$$

with

$$\boldsymbol{\phi}(t_1,t_2)=(\mathbb{P}(CO(t_1)\leq t_2 \mid Y_{\delta_0}=s, Y_{\delta_{-1}}\neq s): s\in\mathcal{S})$$

for $t_1,t_2=0,1,\cdots$. It is seen that (9.52) and (9.53) do not show a close correspondence to their respective continuous–parameter counterpart, (9.50) and (9.51). This is because the discrete–parameter equations (9.52) and (9.53) have been set up so as to make them computationally tractable without first having to compute the discrete–parameter versions of $\mathbf{K}_{\mathcal{U}\mathcal{D}}$ and $\mathbf{K}_{\mathcal{D}\mathcal{U}}$. Both systems of equations are of course arrived at via the renewal argument: for (9.49)–(9.51), the regeneration points are taken to be the instants of change from $\mathcal{U}$ to $\mathcal{D}$ and $\mathcal{D}$ to $\mathcal{U}$, whereas for (9.52)–(9.53) they are defined by the instants of departures from and arrivals to individual states. (It is, of course, also possible to establish a system for the discrete–parameter case which mirrors (9.49)–(9.51).) Note that there is no closed form expression for the cdf of $CO(t)$ corresponding to (9.43)–(9.46) since, as in the continuous–parameter case, (9.52)–(9.53) cannot be written in terms of convolutions. However, using (9.1), we can approach the cumulative operational time via the work–mission availability which, as it will be shown next, will give rise to closed form expressions for both of these measures. The work–mission availability $\mathbf{WMA}(\mathbf{t})$, written as a function of $\mathbf{t}=(t_1,t_2)\in[0,+\infty)^2$, satisfies the following system of integral equations [39]

$$\begin{aligned}\mathbf{WMA}_{\mathcal{U}}(\mathbf{t}) &= \mathbf{1}-\mathbf{K}_{\mathcal{U}\mathcal{D}}(\pi_1(\mathbf{t}))\mathbf{1}\\ &\quad+\int_{[\mathbf{0},\mathbf{t}]}\mathbf{G}_{\mathcal{U}\mathcal{U}}(d\mathbf{w})\mathbf{WMA}_{\mathcal{U}}(\mathbf{t}-\mathbf{w}),\end{aligned} \tag{9.54}$$

$$\begin{aligned}\mathbf{WMA}_{\mathcal{D}}(\mathbf{t}) &= \mathbf{K}_{\mathcal{D}\mathcal{U}}(\pi_2(\mathbf{t}))\mathbf{1}-\mathbf{G}_{\mathcal{D}\mathcal{D}}(\mathbf{t})\mathbf{1}\\ &\quad+\int_{[\mathbf{0},\mathbf{t}]}\mathbf{G}_{\mathcal{D}\mathcal{D}}(d\mathbf{w})\mathbf{WMA}_{\mathcal{D}}(\mathbf{t}-\mathbf{w}),\end{aligned} \tag{9.55}$$

where $\pi_i(\mathbf{t})=t_i$, $i=1,2$, are the projection operators and the matrix–valued functions $\mathbf{G}_{\mathcal{U}\mathcal{U}}$ and $\mathbf{G}_{\mathcal{D}\mathcal{D}}$ on $\mathbf{t}=(t_1,t_2)\in[0,+\infty)^2$ are defined by

$$\begin{aligned}\mathbf{G}_{\mathcal{U}\mathcal{U}}(\mathbf{t})) &= \mathbf{K}_{\mathcal{U}\mathcal{D}}(\pi_1(\mathbf{t}))\mathbf{K}_{\mathcal{D}\mathcal{U}}(\pi_2(\mathbf{t})),\\ \mathbf{G}_{\mathcal{D}\mathcal{D}}(\mathbf{t})) &= \mathbf{K}_{\mathcal{D}\mathcal{U}}(\pi_2(\mathbf{t}))\mathbf{K}_{\mathcal{U}\mathcal{D}}(\pi_1(\mathbf{t})).\end{aligned}$$

As is also shown in [39], the two components of the work–mission–availability vector can be expressed in terms of each other as

$$\begin{aligned}\mathbf{WMA}_{\mathcal{U}}(t_1,t_2) &= \mathbf{1}-\mathbf{K}_{\mathcal{U}\mathcal{D}}(\pi_1(t_1))\mathbf{1}\\ &\quad+\int_{[0,t_1]}\mathbf{K}_{\mathcal{U}\mathcal{D}}(dw)\mathbf{WMA}_{\mathcal{D}}(t_1-w,t_2),\end{aligned} \tag{9.56}$$

$$\mathbf{WMA}_{\mathcal{D}}(t_1,t_2)=\int_{[0,t_2]}\mathbf{K}_{\mathcal{D}\mathcal{U}}(dw)\mathbf{WMA}_{\mathcal{U}}(t_1,t_2-w). \tag{9.57}$$

(In numerical work, it will therefore be sufficient to compute one of the components of **WMA** by solving (9.54) or (9.55) since the other component is easily obtained via either (9.56) or (9.57).) The integrals in (9.54)–(9.57) can, by analogy with all the previous formulae, be thought of as either (now two–dimensional) Stieltjes integrals or integrals with respect to appropriately defined finite measures on $[0,+\infty)^2$. To deduce a formal closed form solution of (9.54) and (9.55), the two–dimensional version of the Stieltjes convolution of matrix–valued functions is needed: by analogy with (9.24), for $\mathbf{M}$ and $\mathbf{N}$ on $[0,+\infty)^2$ we define $\mathbf{M} * \mathbf{N}$ for $\mathbf{t} \in [0,+\infty)^2$ by the Stieltjes integral

$$(\mathbf{M} * \mathbf{N})(\mathbf{t}) = \int_{[\mathbf{0},\mathbf{t}]} \mathbf{M}(d\mathbf{w})\mathbf{N}(\mathbf{t}-\mathbf{w}).$$

$\mathbf{id} = (I_{[0,+\infty)^2}(\mathbf{t})\mathbf{I} : \mathbf{t} \in [0,+\infty)^2)$ is now the neutral element of the Stieltjes convolution. Repeated i–fold convolution is defined analogously to that in the one–dimensional case. Now, the formal solution of (9.54) and (9.55) is respectively

$$\mathbf{WMA}_{\mathcal{U}} = \mathbf{1} - (\mathbf{K}_{\mathcal{UD}} \circ \pi_1)\mathbf{1} + \sum_{n=1}^{\infty} \mathbf{G}_{\mathcal{UU}}{}^{*(n)} * \big(\mathbf{1} - (\mathbf{K}_{\mathcal{UD}} \circ \pi_1)\mathbf{1}\big), \tag{9.58}$$

and

$$\mathbf{WMA}_{\mathcal{D}} = (\mathbf{K}_{\mathcal{UD}} \circ \pi_1)\mathbf{1} - \mathbf{G}_{\mathcal{DD}}\mathbf{1} + \sum_{n=1}^{\infty} \mathbf{G}_{\mathcal{DD}}{}^{*(n)} * \big((\mathbf{K}_{\mathcal{DU}} \circ \pi_2)\mathbf{1} - \mathbf{G}_{\mathcal{DD}}\mathbf{1}\big). \tag{9.59}$$

(The closed form expressions for $\mathbf{K}_{\mathcal{UD}}$ and $\mathbf{K}_{\mathcal{DU}}$ are given respectively by (9.27) and (9.28).) (9.58), (9.59) and (9.1) now give access to closed form expressions also for the cumulative operational time; for instance, in [39], it is shown that for a one–unit system modelled by an alternating renewal process (with $\mathcal{S} = \{u, d\}$), (9.58) translates to

$$\mathbb{P}(CO(t_1 \le t_2 \mid Y_0 = u, Y_{0-} \ne u) = F_{up}(t_2) - \sum_{n=1}^{\infty} \big(F_{up}{}^{*(n)}(t_2) - F_{up}{}^{*(n+1)}(t_2)\big) F_{down}{}^{*(n)}(t_1 - t_2), \tag{9.60}$$

where $F_{up} = F_{u,d}$ and $F_{down} = F_{d,u}$ stand for the time to failure and repair time cdf, respectively.

The methodology for treating the remaining dependability measures from Section 9.3 is similar to the approaches described above. For the *interval availability*, a system of integral equations in terms of $\mathbf{K}_{\mathcal{UD}}$ and $\mathbf{K}_{\mathcal{DU}}$ has been established in [37]; it is *not* a convolution equation, so that it will not be possible to arrive at a closed form expression for the interval availability, which would be analogous to (9.47) or (9.60). In [36], a system of integral equations akin to (9.54)–(9.55) is derived for the *mission availability*. Closed form expressions analogous to (9.60) were also established in [36] for the mission availability. $M_{\mathcal{D}}(t)$, the *number of repair periods* during the time interval $[0,t]$ has been addressed in a number of papers: based on (9.2), a closed form expression for the cdf of $M_{\mathcal{D}}(t)$ was established in [27] under the Markov assumption; under the semi–Markov assumption, various results for $M_{\mathcal{D}}(t)$ in the Laplace transform domain are available from [23], [25], [29] Chapters 9 and 10, [80] and [81]. The discrete–parameter version of $M_{\mathcal{D}}(t)$ under the semi–Markov assumption has been dealt with in [32].

9.6 Numerical Solution Techniques

In this section, the two main techniques for solving for reliability measures in continuous–time semi–Markov models will be addressed. First, a brief overview of the approach based on solving the corresponding integral equations is given in Section 9.6.1. In Section 9.6.2, the time–discretisation approach is then explored in more detail; this is based on replacing the original model by an approximating discrete–parameter semi–Markov model. For *numerical results* concerning actual systems, the reader will be referred to work available in the literature.

9.6.1 Solving the integral equations

For continuous–parameter models, most of the systems of integral equations are of the convolution type. (The systems (9.41)–(9.42) and (9.49)–(9.51), however, obviously do not fall into this category.) These convolution equations can be written in the form

$$\mathbf{H(t)} = \int_{[\mathbf{0,t}]} \mathbf{J}(d\mathbf{w})\mathbf{H}(\mathbf{t}-\mathbf{w})\mathbf{V(t)}, \tag{9.61}$$

where $\mathbf{J}$ and $\mathbf{V}$ are known matrix–valued functions on $[0,+\infty)^m$ and $\mathbf{H}$ is an unknown matrix–valued function, also defined on $[0,+\infty)^m$, $m = 1,2$. The following are examples of systems with the form (9.61): equation (9.7) with $m = 1$, $\mathbf{J}(t) = \mathbf{Q}_{\mathcal{UU}}(t)$, $\mathbf{V}(t) = \mathbf{Q}_{\mathcal{UD}}(t)$, $\mathbf{H}(t) = \mathbf{K}_{\mathcal{UD}}(t)$; equations (9.5)–(9.6) with $m = 1$,

$$\mathbf{J}(t) = \begin{bmatrix} \mathbf{0} & \mathbf{K}_{\mathcal{UD}}(t) \\ \mathbf{K}_{\mathcal{DU}}(t) & \mathbf{0} \end{bmatrix}, \tag{9.62}$$

$$\mathbf{V}(t) = \begin{bmatrix} \mathbf{1} - \mathbf{K}_{\mathcal{UD}}(t)\mathbf{1} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{H}(t) = \begin{bmatrix} \mathbf{PA}_{\mathcal{U}}(t) \\ \mathbf{PA}_{\mathcal{D}}(t) \end{bmatrix};$$

equations (9.39)–(9.40) (for fixed $x \geq 0$) with $m = 1$, $\mathbf{J}(t)$ as in (9.62), and

$$\mathbf{V}(t) = \begin{bmatrix} \int_{(t,t+x]} \mathbf{K}_{\mathcal{UD}}(dw)\mathbf{PA}_{\mathcal{D}}(t+x-w) + \mathbf{1} - \mathbf{K}_{\mathcal{UD}}(t+x)\mathbf{1} \\ \mathbf{0} \end{bmatrix},$$

$$\mathbf{H}(t) = \begin{bmatrix} \mathbf{JA}_{\mathcal{U}}(x,t) \\ \mathbf{JA}_{\mathcal{D}}(x,t) \end{bmatrix};$$

and, finally, (9.54) with $m = 2$, $\mathbf{J(t)} = \mathbf{G}_{\mathcal{UU}}(\mathbf{t})$, $\mathbf{V(t)} = \mathbf{1} - \mathbf{K}_{\mathcal{UD}}(\pi_1(\mathbf{t}))\mathbf{1}$, $\mathbf{H(t)} = \mathbf{WMA}_{\mathcal{U}}(\mathbf{t})$.

A simple yet effective method of solving (9.61) computationally is the *two–point trapezoidal rule*, described for $m = 1$ for example in [15]. A more comprehensive reference, though still addressing the case $m = 1$ only, is the book of Linz [75]. A two–dimensional version of that rule ($m = 2$) has been used in [39] for solving the system (9.54)–(9.55). If the system of integral equations is not of the convolution type (as is the case of (9.49)–(9.51)), it may still be possible to devise a numerical solution scheme based on the one–panel formula [37].

9.6.2 Discrete–parameter approximations

In this section, the process of replacing a continuous–time process Y by some discrete–parameter process $Y^{(\delta)}$ will be discussed [28], [33]. The sequence of states visited by $Y^{(\delta)}$ will be taken to be identical to that visited by Y; thus, the embedded Markov chain of $Y^{(\delta)}$ can be taken to be X. The index set of $Y^{(\delta)}$ is the lattice $\Delta^{(\delta)} = \{0, \delta, 2\delta, \cdots\}$, with $\delta > 0$ fixed. The unit of measurement for the holding times of $Y^{(\delta)}$ is δ; that means that for $k = 1, 2, \cdots$ and $s_1, s_2 \in \mathcal{S}$ $(s_1 \neq s_2)$, the holding time probability mass function $f^{(\delta)}_{s_1,s_2}(k)$ denotes the probability of the event that $Y^{(\delta)}$ will spend k instants (from $\Delta^{(\delta)}$) in s_1, given that the next state to be visited by $Y^{(\delta)}$ is s_2. These holding time distributions of $Y^{(\delta)}$ are obtained by approximating those of Y by appropriate discrete distributions whose support is $\{\delta, 2\delta, 3\delta \cdots\}$. In what follows, *discrete, parametric approximations* will be described for some of the more common classes of continuous distributions in reliability theory. T will stand for the continuous random variable on $(0, +\infty)$ whose distribution is to be approximated by that of some other random variable $T^{(\delta)}$ on $\{\delta, 2\delta, 3\delta \cdots\}$. $T^{(\delta)}$ will be found to converge in distribution on T as $\delta \to 0$.

If T is exponentially distributed with rate parameter $\theta \in (0, +\infty)$, then the distribution of the random variable $T^{(\delta)}$ with

$$\mathbb{P}(T^{(\delta)} = k\delta) = \theta\delta(1 - \theta\delta)^{k-1}, \qquad k = 1, 2, \cdots,$$

will converge for $\delta \to 0$ on that of T by the continuity theorem of Laplace transforms (*e.g.*, [47], [83]) since the following holds in the Laplace transform domain

$$\begin{aligned}
\mathbb{E}\left(\exp(-\tau T^{(\delta)})\right) &= \sum_{k=1}^{\infty} \exp(-\tau k\delta)\theta\delta(1 - \theta\delta)^{k-1} \\
&= \frac{\theta\delta \exp(-\tau\delta)}{1 - (1 - \theta\delta)\exp(-\tau\delta)} = \frac{\theta}{\theta + \tau} + o(1) \\
&= \mathbb{E}\left(\exp(-\tau T)\right) + o(1). \qquad (9.63)
\end{aligned}$$

Equation (9.63) is, of course, an instance of the well–known approximation of the exponential distribution by the *geometric distribution.*

If T has a Weibull distribution with scale parameter $\theta \in (0, +\infty)$ and shape parameter $\beta \in (0, +\infty)$, *i.e.*,

$$\mathbb{P}(T \leq t) = 1 - \exp\left(-(t\theta)^{\beta}\right), \qquad t > 0,$$

then $T^{(\delta)}$ is assumed to have the following discrete Weibull distribution [86], [91],

$$\mathbb{P}(T^{(\delta)} = k\delta) = \left(1 - (\theta\delta)^{\beta}\right)^{(k-1)^{\beta}} - \left(1 - (\theta\delta)^{\beta}\right)^{k^{\beta}}, \quad k = 1, 2, \cdots.$$

The cdf of $T^{(\delta)}$ is given by

$$\mathbb{P}(T^{(\delta)} \leq k\delta) = 1 - \left(1 - (\theta\delta)^{\beta}\right)^{k^{\beta}}, \qquad k = 0, 1, \cdots.$$

For $t > 0$, therefore, the following holds as $\delta \to 0$,

$$\mathbb{P}(T^{(\delta)} \le t) = \mathbb{P}(T^{(\delta)} \le [t/\delta]\delta) = 1 - \left(1 - \frac{[t/\delta]^\beta(\theta\delta)^\beta}{[t/\delta]^\beta}\right)^{[t/\delta]^\beta}$$
$$= 1 - \left(1 - \frac{(t\theta)^\beta + o(1)}{[t/\delta]^\beta}\right)^{[t/\delta]^\beta}$$
$$= 1 - \exp\left(-(t\theta)^\beta\right) + o(1).$$

$[x]$ denotes the integer part of x. The exponential case is obtained by putting $\beta = 1$.

If T has an *Erlang distribution, i.e.*, for some $\theta \in (0, +\infty)$ and $n = 1, 2, \cdots$,

$$\mathbb{P}(T \le t) = 1 - \exp(-t\theta) \sum_{\ell=0}^{n-1} \frac{(t\theta)^\ell}{\ell!}, \qquad t > 0, \tag{9.64}$$

then $T^{(\delta)}$ is assumed to have the following *negative binomial distribution*,

$$\mathbb{P}(T^{(\delta)} = k\delta) = \begin{cases} \binom{k-1}{n-1}(\theta\delta)^n(1-\theta\delta)^{k-n}, & \text{if } k \in \{n, n+1, \cdots\}, \\ 0, & \text{if } k \in \{1, 2, \cdots, n-1\}. \end{cases}$$

We use the continuity theorem again to establish that $T^{(\delta)}$ tends to T in distribution as $\delta \to 0$,

$$\mathbb{E}\left(\exp(-\tau T^{(\delta)})\right) = \sum_{k=n}^{\infty} \exp(-\tau k\delta) \binom{k-1}{n-1} (\theta\delta)^n (1-\theta\delta)^{k-n}$$
$$= \left(\frac{\theta\delta}{1-\theta\delta}\right)^n \sum_{k=n}^{\infty} \binom{k-1}{n-1} \left(\exp(-\tau\delta)(1-\theta\delta)\right)^k$$
$$= \left(\frac{\theta\delta \exp(-\tau\delta)}{1-(1-\theta\delta)\exp(-\tau\delta)}\right)^n$$
$$= \left(\frac{\theta\delta + O(\delta^2)}{(\tau+\theta)\delta + O(\delta^2)}\right)^n = \left(\frac{\theta}{\theta+\tau}\right)^n + o(1)$$
$$= \mathbb{E}\left(\exp(-\tau T)\right) + o(1). \tag{9.65}$$

Put $n = 1$ in (9.65) to obtain the exponential case again.

The random variable T in (9.64) has density

$$\mathbb{P}\left(T \in (t, t+dt)\right) = \frac{\theta}{\Gamma(n)}(t\theta)^{n-1}\exp(-t\theta)dt, \qquad t > 0. \tag{9.66}$$

Allowing $n \in (0, +\infty)$ in (9.66) to be nonintegral, we get the *gamma distribution* with scale parameter θ and shape parameter n. The distribution of $T^{(\delta)}$ will be a *shifted negative binomial distribution* on $\{\delta, 2\delta, 3\delta, \cdots\}$; this is best arrived at as follows. For $\theta\delta \in (0, 1)$, we get, by Taylor expansion of $(1-x)^{-n}$ about the origin (with $x = 1 - \theta\delta$),

$$(\theta\delta)^{-n} = \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \left(\prod_{j=0}^{\ell-1}(n+j)\right)(1-\theta\delta)^\ell,$$

from which it is seen that with $k = 1, 2, \cdots$,

$$\mathbb{P}(T^{(\delta)} = k\delta) = \frac{1}{(k-1)!}\left(\prod_{j=0}^{k-2}(n+j)\right)(\theta\delta)^n(1-\theta\delta)^{k-1},$$

a probability mass function is defined on $\{\delta, 2\delta, 3\delta, \cdots\}$. Now, the convergence in distribution of $T^{(\delta)}$ to T for $\delta \to 0$ is shown again by Laplace transforms,

$$\begin{aligned}
&\mathbb{E}\left(\exp(-\tau T^{(\delta)})\right) \\
&= \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \exp(-\tau k\delta) \left(\prod_{j=0}^{k-2}(n+j)\right) (\theta\delta)^n (1-\theta\delta)^{k-1} \\
&= (\theta\delta)^n \exp(-\tau\delta) \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \left(\prod_{j=0}^{k-2}(n+j)\right) (\exp(-\tau\delta)(1-\theta\delta))^{k-1} \\
&= \exp(-\tau\delta) \left(\frac{\theta\delta}{1-(1-\theta\delta)\exp(-\tau\delta)}\right)^n = \left(\frac{\theta}{\theta+\tau}\right)^n + o(1) \\
&= \mathbb{E}\left(\exp(-\tau T)\right) + o(1).
\end{aligned}$$

Note that, since the Erlang family is contained in the gamma class, the present approximation offers an alternative to the one discussed previously.

If T is *uniformly distributed* in some interval $(t_0, t_1) \subset (0, +\infty)$, then $T^{(\delta)}$ will be distributed according to the following *discrete uniform distribution*

$$\mathbb{P}(T^{(\delta)} = k\delta) = \begin{cases} \frac{1}{[t_1/\delta]-[t_0/\delta]}, & \text{if } k \in \{[t_0/\delta]+1, \cdots, [t_1/\delta]\}, \\ 0, & \text{if } k \notin \{[t_0/\delta]+1, \cdots, [t_1/\delta]\}. \end{cases}$$

$T^{(\delta)}$ converges in distribution to T as $\delta \to 0$ since

$$\begin{aligned}
\mathbb{E}\left(\exp(-\tau T^{(\delta)})\right) &= \sum_{k=[t_0/\delta]+1}^{[t_1/\delta]} \frac{\exp(\tau k\delta)}{[t_1/\delta]-[t_0/\delta]} \\
&= \frac{\exp\left(-\tau\delta([t_0/\delta]+1)\right) - \exp\left(-\tau\delta([t_1/\delta]+1)\right)}{([t_1/\delta]-[t_0/\delta])\left(1-\exp(-\tau\delta)\right)} \\
&= \frac{\exp(-\tau t_0) - \exp(-\tau t_1) + o(1)}{(t_1-t_0)\tau + o(1)} \\
&= \mathbb{E}\left(\exp(-\tau T)\right) + o(1).
\end{aligned}$$

It should be noted that the above approximations are all such that

$$\mathbb{P}(T^{(\delta)} = k\delta) \approx \mathbb{P}\left((k-1)\delta < T \leq k\delta\right), \qquad k = 1, 2, \cdots. \tag{9.67}$$

Equation (9.67) is one of the methods of approximating a distribution on $(0, +\infty)$ by what is considered to be a more manageable one. (An alternative method of approximation uses mixtures of Erlang distributions (*e.g.*, [112]). However, the approximation then does not yield a discrete–parameter semi–Markov process but a continuous–time Markov process, whose state space, incidentally, can become very large.)

9.7 Recent Developments, Conclusions and Further Work

Masuda and Yamakawa [82] have recently extended the analysis of a compound dependability measure from [38] to the semi–Markovian case. In [73], closed

form expression for some simple dependability measures for semi-Markov models are obtained by considering approximations to the solutions of the corresponding convolution equations brought about by truncating the infinite series solution. Noteworthy is also the general modelling framework reported by Limnios and Oprisan [74].

It would appear that the question of parameter estimation from observed realisations (possibly censored) will be of interest here [93]. However, it is more likely that estimates of the model parameters will be compiled from statistical information on the individual building blocks. In [95], Platis *et al.* consider dependability measures for non–homogeneous discrete parameter Markov models in power engineering. There seems to be some scope for examining the usual dependability measures discussed in this chapter for these processes. In [16], Bousfiha and Limnios deal with the numerical computation of reliability measures of semi–Markov models by approximating the holding times by phase–type distributions. Another noteworthy related reference is [57]. The numbers of papers on phase-type distributions are legion; in [1], Aalen gives a thorough overview of their use in survival analysis.

There are several possibilities for future work. The device of the Laguerre transform [69] for Laplace transform inversion has proved a rather useful tool for models where the quantity of interest can be represented in terms of convolutions of some other quantities with known Laplace transforms. A systematic exploration of this computational approach for the suggested dependability measures could be carried out. The dependability measures discussed here could be explored for non–semi–Markovian models. Many practical (and still rather simple) models are not semi–Markovian, and not even regenerative. For such models approximations to the dependability measures are of interest [114]. An interesting more general class of models are the self–exciting point processes [94]. For example, for a one–unit system (which would be modelled by the alternating renewal process if a semi–Markovian approach was taken) the self–exciting point process model allows the unit's (random) hazard rate to be a function of the unit's past; hence, quantities such as the unit's cumulative up–time or the number of past repair events can also be included in the model.

In Section 9.4.2 we have addressed the possibility of approximating a continuous–time semi–Markov model by a discrete–parameter one and then using its corresponding discrete–parameter analogue as an approximation to the original quantity of interest. Even though this is a perfectly valid practical approach, not a lot is known about the degree of accuracy achieved by this without some additional numerical experimentation. There is scope for work in deriving bounds on the distance of the approximation to the true value. (If the dependability measure concerned satisfies an integral equation, this question will probably lead to consideration of the error analysis in the numerical solution of integral equations.) The present modelling framework (semi–Markovian with a partitioned state space) also seems to have been explored for an entirely different application (ion channels in cell neurology) in quite great detail [4], [5]. There may be some profit in studying that literature and trying to see what is applicable from that field of research in the reliability area.

References

1. Aalen, O. O. (1995), "Phase-type distributions in survival analysis," *Scandinavian Journal of Statistics*, **22**, 447–463
2. Ahmad, A., Johannet, P. and Auriol, Ph. (1992), "Efficient inverse Laplace transform algorithm for transient overvoltage calculation," *IEE Proceedings-C: Generation, Transmission and Distribution*, **139**, 117–121
3. Ahmed, S. B., Alam, M. and Gupta, D. (1989), "Performance modelling and evaluation of flexible manufacturing systems using a semi-Markov approach," *International Journal of Computer Integrated Manufacturing*, **2**, 275–280
4. Ball, F. and Yeo, G. F. (1994), "Numerical evaluation of observed sojourn time distribution for a single ion channel incorporating time interval oImission," *Statistics and Computing*, **4**, 1–12
5. Ball, F. and Davies, S. (1997), "Clustering of bursts of openings in Markov and semi–Markov models of single channel gating," *Advances in Applied Probability*, **29**, 92–113
6. Baxter, L. A. (1981), "Availability measures for a two–state system," *Journal of Applied Probability*, **18**, 227–235
7. Baxter, L. A. (1982), "Compound availability measures," *Naval Research Logistics Quarterly*, **29**, 403–410
8. Beasley, M. (1991), Reliability for Engineers. Macmillan, Houndmills, Basingstoke, London
9. Beichelt, F. and Franken, P. (1983), Zuverlssigkeit und Instandhaltung. VEB Verlag Technik, Berlin
10. Bellman, R. E., Kalaba, R. E. and Lockett, J. A. (1966), Numerical Inversion of the Laplace Transform. Elsevier, New York
11. Bellman, R. E. and Roth, R. S. (1984), The Laplace Transform. World Scientific, Singapore
12. Billinton, R., Allan, R. N. and Salvaderi, L. (eds.) (1991), Applied Reliability Assessment in Electric Power Systems. IEEE Press, New York
13. Birolini, A. (1985), On the Use of Stochastic Processes in Modeling Reliability Problems. Lecture Notes in Economics and Mathematical Systems, **252**, Springer-Verlag, Berlin, Heidelberg, New York
14. Birolini, A. (1991), Qualität und Zuverlässigkeit Technischer Systeme: Theorie, Praxis, Management (3rd edn.). Springer-Verlag, Berlin, Heidelberg, New York
15. Boehme, T. K., Preuss, W. and van der Wall, V. (1991), "On a simple numerical method for computing Stieltjes integrals in reliability theory," *Probability in the Engineering and Informational Sciences*, **5**, 113–128
16. Bousfiha, A. and Limnios, N. (1997), "Ph-Distribution Method for Reliability Evaluation of semi–Markov Systems," in Advances in Safety &

Reliability-*Proceedings of the European Conference on Safety and Reliability - ESREL'97* (Soares, C. G. ed.). Elsevier, Oxford, **3**, 2149–2154
17. Bruneel, H. (1993), "Performance of discrete-time queueing systems," *Computers & Operations Research*, **20**, 303–320
18. Choi, B. D, Rhee, K. H. and Park, K. K. (1993), "The M/G/1 retrial queue with retrial rate control policy," *Probability in the Engineering and Informational Sciences*, **7**, 29–46
19. Chou, J.-H. and Horng, I.-R. (1986), "On a functional approximation for inversion of Laplace transforms via shifted Chebyshev series," *International Journal of Systems Science*, **17**, 735–739
20. Ciardo, G., Marie, R., Sericola, B. and Trivedi, K. (1990), "Performability analysis using semi-Markov reward processes," *IEEE Transactions on Computers*, **C–29**, 1251–1264
21. Crump, K. S. (1976), "Numerical inversion of Laplace transforms using a Fourier series approximation," *Journal of the Association of Computing Machinery*, **23**, 89–96
22. Csenki, A. (1991), "The joint distribution of sojourn times in finite semi-Markov processes," *Stochastic Processes and their Applications*, **39**, 287–299
23. Csenki, A. (1991), "Some renewal-theoretic investigations in the theory of sojourn times in finite semi-Markov processes," *Journal of Applied Probability*, **28**, 822–832
24. Csenki, A. (1992), "The joint distribution of sojourn times in finite Markov processes," *Advances in Applied Probability*, **24**, 141–160
25. Csenki, A. (1993), "Occupation frequencies for irreducible finite semi-Markov processes with reliability applications," *Computers & Operations Research*, **20**, 249-259
26. Csenki, A. (1993), "On a counting variable in the theory of discrete–parameter Markov chains," *Statistics & Probability Letters*, **18**, 105–112
27. Csenki, A. (1994), "The number of working periods of a repairable Markov system during a finite time interval," *IEEE Transactions on Reliability*, **R–3**, 163–169
28. Csenki, A. (1994), "Cumulative operational time analysis of finite semi-Markov reliability models," *Reliability Engineering and System Safety*, **44**, 17–25
29. Csenki, A. (1994), "Dependability for systems with a partitioned state space: Markov and semi-Markov theory and computational implementation," Lecture Notes in Statistics, **90**, Springer–Verlag, Berlin, Heidelberg, New York
30. Csenki, A. (1994), "On the interval reliability of systems modelled by finite semi-Markov processes," *Microelectronics and Reliability*, **34**, 1319–1335
31. Csenki, A. (1994), "Joint availability of systems modelled by finite semi-Markov processes," *Applied Stochastic Models and Data Analysis*, **10**, 279–293
32. Csenki, A. (1995), "The number of visits to a subset of the state space by a discrete–parameter semi-Markov process," *Statistics & Probability Letters*, **22**, 71–77
33. Csenki, A. (1995), "Total cumulative work until failure of a system modelled by a finite semi-Markov process," *International Journal of Systems Science*, **26**, 1511–1525
34. Csenki, A. (1995), "An integral equation approach to the interval reliability of systems modelled by finite semi-Markov processes," *Reliability Engineering and System Safety*, **47**, 37–45

35. Csenki, A. (1995), "Set reliability: a unified approach to dependability measures for semi-Markov reliability models," *Systems Analysis Modelling Simulation*, **20**, 173–186
36. Csenki, A. (1995), "Mission availability for repairable semi-Markov systems: analytical results and computational implementation," *Statistics*, **26**, 75–87
37. Csenki, A. (1995), "Transient analysis of interval availability for repairable systems modelled by finite semi-Markov processes," *IMA Journal of Mathematics Applied in Business and Industry*, **6**, 267–281
38. Csenki, A. (1996), "A compound measure of dependability for systems modeled by continuous-time absorbing Markov processes," *Naval Research Logistics*, **43**, 305–312
39. Csenki, A. (1996), "A new approach to the cumulative operational time for semi-Markov models of repairable systems," *Reliability Engineering and System Safety*, **54**, 11–21
40. Darroch, J. N. and Morris, K. W. (1967), "Some passage-time generating functions for discrete-time and continuous-time finite Markov chains," *Journal of Applied Probability*, **4**, 496–507
41. Delves, L. M. and Walsh, J. (eds.) (1974), Numerical Solution of Integral Equations. Oxford University Press, Oxford
42. Delves, L. M. and Mohamed, J. L. (1985), Computational Methods for Integral Equations. Cambridge University Press, Cambridge
43. De Souza e Silva, E. and Gail, H. R. (1986), "Calculating cumulative operational time distributions of repairable computer systems," *IEEE Transactions on Computers*, **C–35**, 322–332
44. De Souza e Silva, E. and Gail, H. R. (1992), "Performability analysis of computer systems: from model specification to solution," *Performance Evaluation*, **14**, 157–196
45. Dubner, H. and Abate, J. (1968), "Numerical inversion of Laplace transforms by relating them to the finite fourier cosine transform," *Journal of the Association for Computing Machinery*, **15**, 115–123
46. Durbin, F. (1974), "Numerical inversion of Laplace transforms: an efficient improvement to Dubner and Abate's method," *The Computer Journal*, **17**, 371–376
47. Feller, W. (1971), An Introduction to Probability Theory and Its Applications **II** (2nd edn.). John Wiley & Sons, New York
48. Garbow, B. S., Giunta, G. and Murli, A. (1988), "Software for an implementation of weeks' method for the inverse Laplace transform problem," *ACM Transactions on Mathematical Software*, **14**, 163-170
49. Garbow, B. S., Giunta, G., Lyness, J. N. and Murli, A. (1988), "Algorithm 662: A fortran software package for the numerical inversion of the Laplace transform based on Weeks' method," *ACM Transactions on Mathematical Software*, **14**, 171–176
50. Gaver, D. P. (1966), "Observing stochastic processes, and approximate transform inversion," *Operations Research*, **14**, 444–459
51. Goyal, A. and Tantawi, A. N. (1985), "Numerical evaluation of guaranteed availability," *Proceedings of the 15th International Conference on Fault-Tolerant Computing*, IEEE Computer Society Press, Silver Spring, Maryland, 324–329
52. Goyal, A. and Tantawi, A. N. (1988), "A measure of guaranteed availability and its numerical evaluation," *IEEE Transactions on Computers*, **C–37**, 25–32

53. Grosh, D. L. (1989), A Primer of Reliability Theory. John Wiley & Sons, New York
54. Gupta, S. M., Jaiswal, N. K. and Goel, L. R. (1982), "Analysis of a two-unit cold standby redundant system with allowed down-time," *International Journal of Systems Science*, **13**, 1385–1392
55. Herrmann, C. (1993), A Performance Model for Statistical Multiplexing of Correlated ATM Traffic Superpositions. Messung, Modellierung und Bewertung von Rechen- und Kommunikationssystemen (Walke, B. and Spaniol, O. eds.). *Proceedings of the 7th ITG/GI-Fachtagung Aachen*, Springer-Verlag, Berlin, 199–211
56. Höfle-Isphording, U. (1978), Zuverlässigkeitsrechnung. Springer-Verlag, Berlin
57. Hongler, M. O. and Salama, Y. (1996), "Semi-Markov processes with phase–type waiting times," *Zeitschrift für Angewandte Mathematik und Mechanik*, **76**, 461–462
58. Howard, R. A. (1971), Dynamic Probabilistic Systems, **II**: Semi-Markov and Decision Processes. John Wiley & Sons, New York
59. Hsueh, M. C, Iyer, R. K. and Trivedi, K. (1988), "Performability modeling based on real data: a case study," *IEEE Transactions on Computers*, **C–37**, 478–484
60. Islamov, R. T. (1994), "Using Markov reliability modelling for multiple repairable systems," *Reliability Engineering and System Safety*, **44**, 113–118
61. Jagerman, D. L. (1978), "An inversion technique for the Laplace transform with application to approximation," *The Bell System Technical Journal*, **57**, 669–710
62. Jagerman, D. L. (1982), "An inversion technique for the Laplace transform," *The Bell System Technical Journal*, **61**, 1995–2002
63. Janssen, J. (ed.) (1986), Semi-Markov Models: Theory and Applications. Plenum Press, New York
64. Kapur, P. K., Natarajan, T. V. and Chakrabarty, D. (1983), "Availability measures for an intermittently used repairable system," *Microelectronics and Reliability*, **23**, 841–844
65. Klaassen, K. B. and Peppen, J. C. L. (1989), System Reliability Concepts and Applications. Edward Arnold, London
66. Kohlas, J. (1982), Stochastic Methods of Operations Research. Cambridge University Press, Cambridge
67. Kohlas, J. (1987), Zuverlässigkeit und Verfügbarkeit. Teubner, Stuttgart
68. Kondo, J. (1991), Integral Equations. Clarendon Press, Oxford
69. Kubat, O., Sumita, U. and Masuda, Y. (1988), "Dynamic performance evaluation of communication/computer systems with highly reliable components," *Probability in the Engineering and Informational Sciences*, **2**, 185–213
70. Kulkarni, V. G., Nicola, V. F. and Trivedi, K. (1987), "The completion time of a job on multimode systems," *Advances in Applied Probability*, **19**, 932–954
71. Laprie, J.–C., Costes, A. and Landrault, C. (1981), "Parametric analysis of 2–unit redundant computer systems with corrective and preventive maintenance," *IEEE Transactions on Reliability*, **R–30**, 139–144
72. Lewis, E. E. (1987), Introduction to Reliability Engineering. John Wiley & Sons, New York
73. Limnios, N. (1997), "Dependability analysis of semi-Markov systems," *Reliability Engineering and System Safety*, **55**, 203–207

74. Limnios, N. and Oprisan, G. (1997), "A General framework for reliability and performability modelling of semi-Markov systems." *Proceedings of the 8th International Symposium on Applied Stochastic Models and Data Analysis*, **2**, Anacapry, Italy, 261–266
75. Linz, P. (1985), Analytical and Numerical Methods for Volterra Equations. SIAM, Philadelphia
76. Longman, I. M (1968), "On the numerical inversion of the Laplace transform of a discontinuous original," *Journal of the Institute of Mathematics and its Applications*, **4**, 320–328
77. Marszalek, W. (1983), "The block–pulse functions method of the two–dimensional Laplace transform," *International Journal of Systems Science*, **14**, 1311–1317
78. Marszalek, W. (1984), "On the inverse Laplace transform of irrational and transcendental transfer functions via block–pulse functions method," *International Journal of Systems Science*, **15**, 869–876
79. Marszalek, W. (1984), "On the nature of block–pulse operational matrices," *International Journal of Systems Science*, **15**, 983–989
80. Masuda, Y. and Sumita, U. (1987), "Analysis of a counting process associated with a semi-Markov process: number of entries into a subset of state space," *Advances in Applied Probability*, **19**, 767–783
81. Masuda, Y. and Sumita, U. (1991), "A multivariate reward process defined on a semi-Markov process and its first–passage–time distributions," *Journal of Applied Probability*, **28**, 360–373
82. Masuda, Y. and Yamakawa, S. (1997), A Compound Dependability Measure Arising from semi-Markov Reliability Model. Technical Report, Faculty of Science and Technology, Keio University, Yokohama, Japan
83. Medhi, J. (1982), Stochastic Processes. Wiley Eastern, New Delhi
84. Mode, C. J. (1985), Stochastic Processes in Demography and Their Computer Implementation. Springer–Verlag, Berlin
85. Mode, C. J. and Pickens, G. T. (1988), "Computational methods for renewal theory and semi-Markov processes with illustrative examples," *The American Statistician*, **42**, 143-151
86. Nagakawa, T. and Osaki, S. (1975), "The discrete Weibull distribution," *IEEE Transactions on Reliability*, **R–24**, 300–301
87. Nam, L. K. and Zin, C. N. (198), "A semi-Markov reliability analysis of alternating systems," *Transactions of the American Nuclear Society*, **60**, 413–414
88. Natarajan, R. and M. A. S. Hameed, M. A. S. (1986), "Some general measures for a complex n-unit standby redundant system," *Microelectronics and Reliability*, **26**, 619–623
89. Nollau, V. (1981), Semi-Markovsche Prozesse. Verlag Harri Deutsch, Thun, Frankfurt a. M.
90. Numerical Algorithms Group (1990) Nag Fortran Library User Manual, Mark 14, **1**. Oxford; (Routines: C06LAF, C06LBF, and C06LCF)
91. Osaki, S. (1985), Stochastic System Reliability Modeling. World Scientific, Singapore
92. Osaki, S. (1992), Applied Stochastic System Modeling. Springer-Verlag, Berlin
93. Ouhbi, B. and Limnios, N. (1996), "Non-parametric estimation for semi-Markov kernels with application to reliability analysis," *Applied Stochastic Models and Data Analysis*, **12**, 209–220
94. Para, A. F. and Garibba, S. (1980), Reliability Analysis of a Repairable Single Unit Under General Age-Dependence. Synthesis and Analysis Meth-

ods for Safety and Reliability Studies (Apostolakis, G., Garibba, S. and Volta, G. eds.). *Proceedings of the NATO Advanced Study Institute on Synthesis and Analysis Methods for Safety and Reliability Studies*, Urbino, Italy, Plenum Press, New York, 251–268
95. Platis, A., Limnios, N. and Le, Du M. (1996), "Performability of electric-power systems modeled by non-homogeneous Markov chains," *IEEE Transactions on Reliability*, **R–45**, 605–610
96. Rao, S. S. (1992), Reliability–Based Design. McGraw-Hill, New York
97. Rubino, G. and Sericola, B. (1993), "Sojourn times in semi–Markov reward processes: application to fault–tolerant systems modeling," *Reliability Engineering and System Safety*, **41**, 1–4
98. Rubino, G. and Sericola, B. (1992), "Interval availability analysis using operational periods," *Performance Evaluation*, **14**, 257–272
99. Sculli, D. and Choy, S. K. (1993), "Power plant boiler feed system reliability: a case study," *Computers in Industry*, **21**, 93–99
100. B. Sericola, B. (1990), "Closed–form solution for the distribution of the total time spent in a subset of states of a homogeneous Markov process during a finite observation period," *Journal of Applied Probability*, **27**, 713–719
101. Singh, C. and Billinton, R. (1977), System Reliability Modelling and Evaluation. Hutchinson, London
102. Singhal, K. and Vlash, J. (1975), "Computation of time domain response by numerical inversion of the Laplace transform," *Journal of the Franklin Institute*, **299**, 109–126
103. Smith, R. M., Trivedi, K. and Ramesh, A. V. (1988), "Performability analysis: measures, an algorithm, and a case study," *IEEE Transactions on Computers*, **C–37**, 406–417
104. Smith, D. J. (1993), Reliability Maintainability and Risk – Practical Methods for Engineers (4th edn.). Butterworth-Heinemann, Oxford
105. Stehfest, H. (1970), "Algorithm: 368 numerical inversion of Laplace transforms," *Communications of the ACM*, **13**, 47-49
106. Stepanek, E. (1971), "Eine numerische methode zur umkehr der Laplace-transformation," *Wissenschaftliche Zeitschrift für Elektrotechnik*, **17**, 261–272
107. Stepanek, E. (1974), "Erweiterung einer numerischen methode zur umkehr der Laplacetransformation," *Zeitschrift für Elektrische Informations und Energietechnik*, **4**, 41–44
108. Stepanek, E. (1974), Praktische Analyse Linearer Systeme durch Faltungsoperationen. Akademische Verlagsgesellschaft, Leipzig
109. Störmer, H. (1970), Semi–Markoff–Prozesse mit Endlich Vielen Zuständen. Lecture Notes in Operations Research and Mathematical Systems, **34**. Springer-Verlag, Berlin, Heidelberg, New York
110. Sundarajan, C. (1991), Guide to Reliability Engineering – Data, Analysis, Applications, Implementation, and Management. Van Nostrand Reinhold, New York
111. Talbot, A. (1979), "The accurate numerical inversion of Laplace transforms," *Journal of the Institute of Mathematics and its Applications*, **23**, 97–120
112. Tijms, H. C. (1990), Stochastic Modelling and Analysis: A Computational Approach. John Wiley & Sons, New York
113. Trivedi, K., Muppala, J. K., Woolet, S.P. and Haverkort, B. R. (1992), "Composite performance and dependability analysis," *Performance Evaluation*, **14**, 197–215

114. Van der Heiden, M. C. (1987), "Interval availability distribution for a 1–out–of–2 reliability system," *Probability in the Engineering and Informational Sciences*, **2**, 211–223
115. Weeks, W. T. (1966), "Numerical inversion of Laplace transforms," *Journal of the Association of Computing Machinery*, **13**, 419–429
116. Woodward, M. E. (1993), Communication and Computer Networks Modelling with Discrete–Time Queues. Pentech Press, London

10. Software Reliability Models

Shigeru Yamada
Department of Social Systems Engineering,
Tottori University
Tottori 680–8552, Japan

Summary.

Software reliability is one of the most important characteristics of software quality. Its measurement and management technologies during the software life-cycle are essential to produce and maintain quality/reliable software systems. In this chapter, we discuss software reliability modeling and its applications. As to software reliability modeling, hazard rate and NHPP models are investigated particularly for quantitative software reliability assessment. Further, imperfect debugging and software availability models are also discussed with reference to incorporating practical factors of dynamic software behavior. And three software management problems are discussed as an application technology of software reliability models: the optimal software release problem, statistical testing-progress control, and the optimal testing-effort allocation problem.

Keywords: software reliability measurement and assessment, reliability growth model, imperfect debugging model, availability model, optimal release problem, testing-progress control, optimal testing-effort allocation problem

10.1 Introduction

In recent years, many computer system failures have been caused by software faults introduced during the software development process. This is an inevitable problem, since a software system installed in the computer system is an intellectual product consisting of documents and source programs developed by human activities. Then, total quality management (TQM) is considered to be one of the key technologies needed to produce more highly quality software products [1], [2]. In the case of TQM used for software development, all phases of the development process, *i.e.* requirement specification, design, coding, and testing, have to be controlled systematically to prevent the introduction of software bugs or faults as far as possible and to detect any introduced faults in the software system as early as possible. Basically, the concept of TQM means assuring the quality of the products in each phase to the next phase. Particularly, quality control carried out at the testing phase,

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

which is the last stage of the software development process, is very important. During the testing phase, the product quality and the software performance during the operation phase are evaluated and assured. In concrete terms, a lot of software faults introduced in the software system through the first three phases of the development process by human activities are detected, corrected, and removed. Figure 10.1 shows a general software development process called a waterfall paradigm.

Therefore, TQM for software development, *i.e. software TQM*, has been emphasized. Software TQM aims to manage the software life-cycle comprehensively, considering productivity, quality, cost and delivery simultaneously. In particular, the management technologies for improving software reliability are very important. The quality characteristic of software reliability is that computer systems can continue to operate regularly without the occurrence of failures in software systems.

In this chapter, we discuss a quantitative technique for software quality/reliability measurement and assessment as one of the key software reliability technologies, which is a so-called *software reliability model*, and its applications.

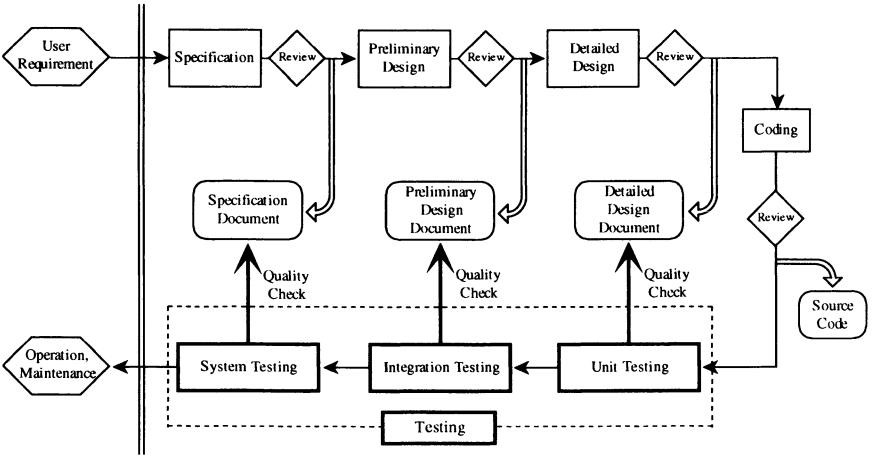


Fig. 10.1. A general software development process (water-fall paradigm)

10.2 Definitions and Software Reliability Model

Generally, a software failure caused by software faults latent in the system cannot occur except on a specific occasion when a set of specific data is put into the system under a specific condition, *i.e.* the program path including software faults is executed. Therefore, the software reliability is dependent on the input data and the internal condition of the program. We summarize the definitions of the technical terms related to the software reliability below.

A *software system* is a product which consists of the programs and documents produced through the software development process discussed in the

previous section (see Figure 10.1). The specification derived by analyzing user requirements for the software system is a document which describes the expected performance of the system. When the software performance deviates from the specification and an output variable has an improper value or the normal processing is interrupted, it is said that a software failure occurs. That is, *software failure* is defined as an unacceptable departure of program operation from the program requirements. The cause of software failure is called a software fault. Then, *software fault* is defined as a defect in the program which causes a software failure. The software fault is usually called a *software bug*. *Software error* is defined as human action that results in the software system containing a software fault [3], [4]. Thus, the software fault is considered to be a manifestation of software errors.

Based on the basic definitions above, we can describe a software behavior as Input(I)-Program(P)-Output(O) model [5], [6], as shown in Figure 10.2.

In this model a program is considered as a mapping from the input space constituting input data available on use to the output space constituting output data or interruptions of normal processing. Testing space T is an input subspace of I, the performance of which can be verified and validated by software testing. Software faults detected and removed during the testing phase map the elements of input subspace E into an output subspace O' constituting the events of a software failure. That is, the faults detected during the testing phase belong to the intersection of subspace E and T. Software faults remaining in the operation phase belong to the subspace E but not to the testing space T.

Under the definitions for technical terms above, *software reliability* is defined as the attribute that a software system will perform without causing software failures over a given time period under specified conditions, and is measured by its probability [3], [4]. A *software reliability model* is a mathematical analysis model for the purpose of measuring and assessing software quality/reliability quantitatively. Many software reliability models have been proposed and applied to practical use because software reliability is considered to be a "*must-be quality*" characteristic of a software product. The software reliability models can be divided into two classes [6], [7]. One treats the upper software development process, *i.e.* design and coding phases, and analyzes the reliability factors of the software products and processes. The other deals with testing and operation phases by describing a software failure-occurrence phenomenon or software fault-detection phenomenon, by applying the stochastic/statistics theories and can estimate and predict the software reliability.

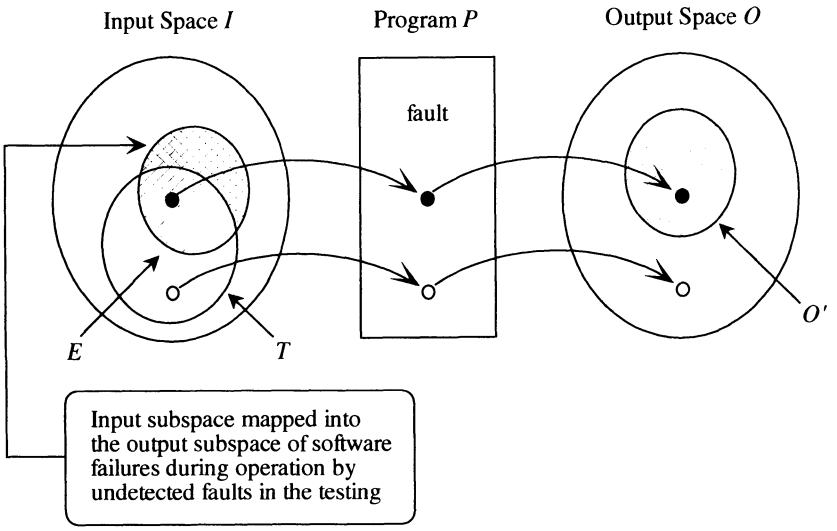
In the former class, a *software complexity model* is well known and can measure the reliability by assessing the complexity based the structural characteristics of products and the process features to produce the products. In the latter class, a *software reliability growth model* is especially well known. Further, this model is divided into three categories [6], [7]:

(1) *Software failure-occurrence time model*
The model which is based on the software failure-occurrence time or the software fault-detection time.

(2) *Software fault-detection count model*
The model which is based on the number of software failure-occurrences or the number of detected faults.

(3) *Software availability model*
The model which describes the time-dependent behavior of software system alternating up (operation) and down (fault correction) states.

The software reliability growth models are utilized for assessing the degree of achievement of software quality, deciding the time to software release for operational use, and evaluating the maintenance cost for faults undetected during the testing phase. We discuss the software reliability growth models and their applications below.



T : Input subspace of I, verified and validated by testing (Testing Domain)

E : Input subspace mapped into the output subspace of O' by software faults

O' : Output subspace of O, constituting events of software failures caused by software faults latent in program P

Fig. 10.2. An input-program-output model for software behavior

10.3 Software Reliability Growth Modeling

Generally, a mathematical model based on stochastic and statistical theories is useful to describe the software fault-detection phenomena or the software failure-occurrence phenomena and estimate the software reliability quantitatively. During the testing phase in the software development process, software faults are detected and removed with a lot of testing-effort expenditures. Then, the number of faults remaining in the software system decreases as the testing goes on. This means that the probability of software failure-occurrence is decreasing, so that the software reliability is increasing and the time interval

between software failures becoming longer with the testing time (see Figure 10.3).

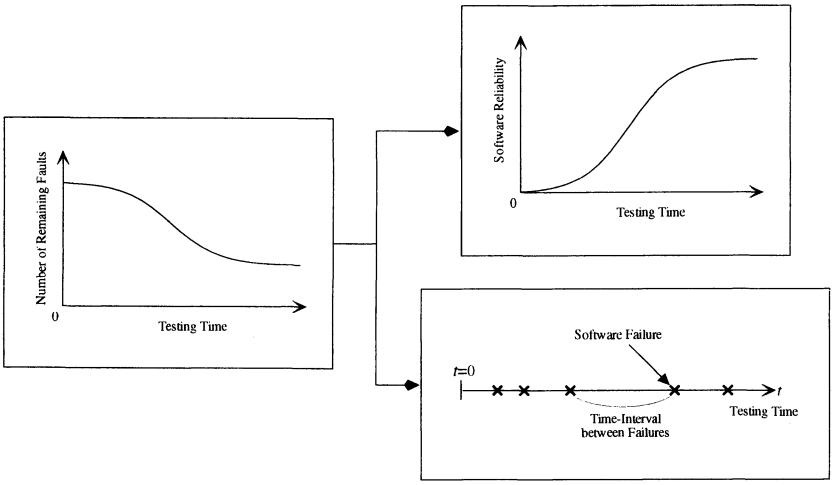
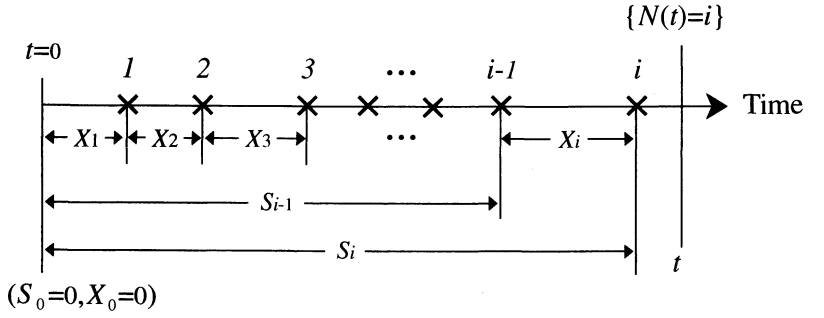


Fig. 10.3. Software reliability growth

A mathematical tool which describes software reliability aspect is a *software reliability growth model* [6], [8], [9].

Based on the definitions discussed in the previous section, we can develop a software reliability growth model based on the assumptions used for the actual environment during the testing phase or the operation phase. Then, we can define the following random variables on the number of detected faults and the software failure-occurrence time (see Figure 10.4):



(×: Software fault detection or software failure occurrence)

Fig. 10.4. The stochasic quantities related to a software fault-detection phenomenon or a software failure-occurrence phenomenon

$N(t)$ = the cumulative number of software faults (or the cumulative number of observed software failures) detected up to time t,
S_i = the i-th software-failure occurrence time ($i = 1, 2, \ldots; S_0 = 0$),
X_i = the time interval between $(i-1)$-st and i-th software failures ($i = 1, 2, \ldots; X_0 = 0$).

Figure 10.4 shows the occurrence of event $\{N(t) = i\}$ since i faults have been detected up to time t. From these definitions, we have

$$S_i = \sum_{k=1}^{i} X_k, \qquad X_i = S_i - S_{i-1}. \tag{10.1}$$

Assuming that the *hazard rate*, *i.e.* the *software failure rate*, for $X_i (i = 1, 2, \ldots)$, $z_i(x)$, is proportional to the current number of residual faults remaining in the system, we have

$$z_i(x) = (N - i + 1)\lambda(x), \qquad i = 1, 2, \ldots, N; x \geq 0, \lambda(x) > 0, \tag{10.2}$$

where N is the initial fault content and $\lambda(x)$ the software failure rate per fault remaining in the system at time x. If we consider two special cases in (10.2) as

$$\lambda(x) = \phi, \qquad \phi > 0, \tag{10.3}$$

$$\lambda(x) = \phi x^{m-1}, \qquad \phi > 0, m > 0, \tag{10.4}$$

then two typical software hazard rate models, respectively called the Jelinski-Moranda model [10] and the Wagoner model [11] can be derived, where ϕ and m are constant parameters. Usually, it is difficult to assume that a software system is completely fault free or failure free. Then, we have a software hazard rate model called the Moranda model [12] for the case of the infinite number of software failure occurrences as

$$z_i(x) = Dk^{i-1}, \qquad i = 1, 2, \ldots; D > 0, 0 < k < 1, \tag{10.5}$$

where D is the initial software hazard rate and k the decreasing ratio. Equation (10.5) describes a software failure-occurrence phenomenon where a software system has high frequency of software failure occurrence during the early stage of the testing or the operation phase and it gradually decreases thereafter. Based on the software hazard rate models above, we can derive the *software reliability* function for $X_i (i = 1, 2, \ldots)$ as

$$R_i(x) = \exp\left[-\int_0^x z_i(x)dx\right], \qquad i = 1, 2, \ldots. \tag{10.6}$$

In this section, we discuss NHPP models [8], [13]–[15], which are modeled for random variable $N(t)$ as typical software reliability growth models. In the NHPP models, a *nonhomogeneous Poisson process* (*NHPP*) is assumed for the random variable $N(t)$, the distribution function of which is given by

$$\Pr\{N(t) = n\} = \frac{\{H(t)\}^n}{n!} \exp[-H(t)], \qquad n = 1, 2, \ldots,$$
$$H(t) = \int_0^t h(x)dx, \tag{10.7}$$

where Pr{A} means the probability of event A. $H(t)$ in (10.7) is called a *mean value function* which indicates the expectation of $N(t)$, *i.e.* the expected cumulative number of faults detected (or the expected cumulative number of software failures occurred) in the time interval $(0, t]$, and $h(t)$ in (10.7) called an *intensity function* which indicates the instantaneous fault-detection rate at time t.

From (10.7), various software reliability assessment measures can be derived. For examples, the expected number of faults remaining in the system at time t is given by

$$n(t) = a - H(t), \tag{10.8}$$

where $a \equiv H(\infty)$, *i.e.* parameter a denotes the expected initial fault content in the software system. Given that the testing or the operation has been going on up to time t, the probability that a software failure does not occur in the time interval $(t, t + x](x \geq 0)$ is given by conditional probability $\Pr\{X_i > x | S_{i-1} = t\}$ as

$$R(x|t) = \exp[H(t) - H(x + t)], \qquad t \geq 0, x \geq 0. \tag{10.9}$$

$R(x|t)$ in (10.9) is a so-called *software reliability*. Measures of *MTBF* (mean time between software failures or fault detections) can be obtained follows:

$$MTBF_I(t) = \frac{1}{h(t)}, \tag{10.10}$$

$$MTBF_C(t) = \frac{t}{H(t)}. \tag{10.11}$$

*MTBF*s in (10.10) and (10.11) are called *instantaneous* and *cumulative MTBF*s, respectively.

It is obvious that the lower the value of $n(t)$ in (10.8), the higher the value $R(x|t)$ for specified x in (10.9), or the longer the value of *MTBF*s in (10.10) and (10.11), the higher the achieved software reliability is. Then, analyzing actual test data with accepted NHPP models, these measures can be utilized to assess software reliability during the testing or operation phase, where statistical inferences, *i.e.* parameter estimation and goodness-of-fit test, are usually performed by a method of maximum likelihood.

To assess the software reliability actually, it is necessary to specify the mean value function $H(t)$ in (10.7). Many NHPP models considering the various testing or operation environments for software reliability assessment have been proposed in the last decade. Typical NHPP models are summarized in Table 10.1. As discussed above, a software reliability growth is described as the relationship between the elapsed testing or operation time and the cumulative number of detected faults and can be shown as the reliability growth curve mathematically (see Figure 10.5). Among the NHPP models in Table 10.1, exponential and modified exponential software reliability growth models are appropriate when the observed reliability growth curve shows an exponential

curve ((A) in Figure 10.5). Similarly, delayed S-shaped and inflection S-shaped software reliability growth models are appropriate when the reliability grouth curbe is S shaped ((B) in Figure 10.5).

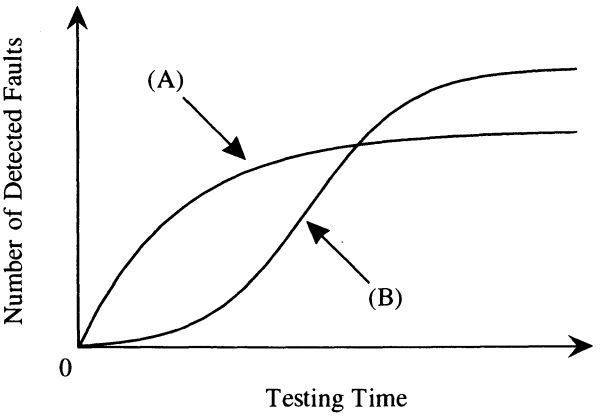


Fig. 10.5. Typical software reliability growth curves

In addition, as for computer makers or software houses in Japan, *logistic curve* and *Gompertz curve models* have often been used as software quality assessment models, on the assumption that software fault-detection phenomena can be shown by S-shaped reliability growth curves [30], [31]. In these deterministic models, the cumulative number of faults detected up to testing t is formulated by the following growth equations:

$$L(t) = \frac{k}{1 + me^{-\alpha t}}, \qquad m > 0, \alpha > 0, k > 0, \tag{10.12}$$

$$G(t) = ka^{(b^t)}, \qquad 0 < a < 1, 0 < b < 1, k > 0. \tag{10.13}$$

In (10.12) and (10.13), assuming that a convergence value of each curve ($L(\infty)$ or $G(\infty)$), *i.e.* parameter k, represents the initial fault content in the software system, it can be estimated by a regression analysis.

10.4 Imperfect Debugging Modeling

Most software reliability growth models proposed so far are based on the assumption of perfect debugging, *i.e.* that all faults detected during the testing and operation phases are corrected and removed perfectly. However, debugging actions in real testing and operation environments are not always performed perfectly. For example, typing errors invalidate the fault-correction activity or fault-removal is not carried out precisely due to incorrect analysis of test results [32]. We therefore have an interest in developing a software reliability growth model which assumes an *imperfect debugging* environment (cf. [33], [34]). Such an imperfect debugging model is expected to estimate reliability assessment measures more accurately.

Table 10.1. A summary of NHPP models

NHPP model	Mean Value Function $H(t)$	Intensity Function $h(t)$	Environment
Exponential software reliability growth model [16], [17]	$m(t) = a(1-e^{-bt})$ $(a>0, b>0)$	$h_m(t) = abe^{-bt}$	A software failure-occurrence phenomenon with a constant fault-detection rate at an arbitrary time is described.
Modified exponential software reliability growth model [18], [19]	$m_p(t) = a\sum_{i=1}^{2} p_i(1-e^{-b_i t})$ $(a>0, 0<b_2<b_1<1,$ $\sum_{i=1}^{2} p_i = 1, 0<p_i<1)$	$h_p(t) = a\sum_{i=1}^{2} p_i b_i e^{-b_i t}$	A difficulty of software fault-detection during the testing is considered, (b_1 is the fault-detection rate for easily detectable faults: b_2 is the fault-detection rate for hardly detectable faults).
Delayed S-shaped software reliability growth model [20], [21]	$M(t) = a[1-(1+bt)e^{-bt}]$ $(a>0, b>0)$	$h_M(t) = ab^2te^{-bt}$	A software fault-detection process is described by two successive phenomena, i.e. failure-detection process and fault-isolation process.
Inflection S-shaped software reliability growth model [22], [23]	$I(t) = \frac{a(1-e^{-bt})}{(1+ce^{-bt})}$ $(a>0, b>0, c>0)$	$h_I(t) = \frac{ab(1+c)e^{-bt}}{(1+ce^{-bt})^2}$	A software failure-occurrence phenomenon with mutual dependency of detected faults is described.
Testing-effort-dependent software reliability growth model [24], [25]	$T(t) = a[1-e^{-rW(t)}]$ $W(t) = \alpha(1-e^{-\beta t^m})$ $(a>0, 0<r<1,$ $\alpha>0, \beta>0, m>0)$	$h_T(t) = ar\alpha\beta$ $\cdot m t^{m-1} e^{-rW(t)}$	The time-dependent behavior of the amount of testing effort and the cumulative number of detected faults is considered.
Testing-domain-dependent software reliability growth model [26], [27]	$D(t) = a[1-\frac{1}{v-b}(ve^{-bt}$ $-be^{-vt})]$ $(v \neq b)$	$h_D(t) = \frac{avb}{v-b}(e^{-bt}-e^{-vt})$	The testing domain, which is the set of software functions influenced by executed test cases, is considered.
Logarithmic Poisson execution time model [28], [29]	$\mu(t) = \frac{1}{\theta}\ln(\lambda_0\theta t+1)$ $(\lambda_0>0, \theta>0)$	$\lambda(t) = \frac{\lambda_0}{(\lambda_0\theta t+1)}$	When the testing or operation time is measured on the basis of the number of CPU hours, an exponentially decreasing software failure rate is considered with respect to the cumulative number of software failures.

a = the expected number of initial fault content in the software system
b = the parameter representing the fault-detection rate
c = the parameter representing the inflection factor of test personnel
α, β, m = the parameters which determine the testing-effort function $W(t)$
v = the testing-domain growth rate
λ_0 = the initial software failure rate
θ = the reduction rate of software failure rate

10.4.1 Imperfect debugging model with perfect correction rate

To model an imperfect debugging environment, the following assumptions are made:

(1) Each fault which causes a software failure is corrected perfectly with probability $p(0 \leq p \leq 1)$. It is not corrected with probability $q(= 1 - p)$. We call p the perfect debugging rate or the perfect correction rate.
(2) The hazard rate is given by (10.5) and decreases geometrically each time a detected fault is corrected.
(3) The probability that two or more software failures occur simultaneously is negligible.
(4) No new faults are introduced during the debugging. At most one fault is removed when it is corrected, and the correction time is not considered.

Let $X(t)$ be a random variable representing the cumulative number of faults corrected up to the testing time t. Then, $X(t)$ forms a *Markov process* [35]. That is, from assumption (1), when i faults have been corrected by arbitrary testing time t,

$$X(t) = \begin{cases} i, & \text{with probability } q, \\ i+1, & \text{with probability } p, \end{cases} \tag{10.14}$$

(see Figure 10.6). Then, the one-step transition probability for the Markov process that after making a transition into state i, the process $\{X(t), t \geq 0\}$makes a transition into state j by time t is given by

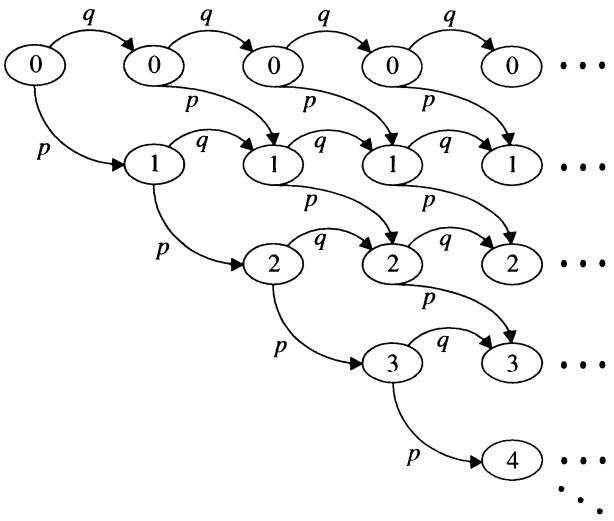


Fig. 10.6. A diagrammatic representation of transitions between states of $X(t)$

$$Q_{ij}(t) = p_{ij}(1 - \exp[-Dk^i t]), \tag{10.15}$$

where p_{ij} are the transition probabilities from state i to state j and are given by

$$p_{ij} = \begin{cases} q, & (i = j), \\ p, & (j = i + 1), \\ 0, & (\text{elsewhere}), \end{cases} \qquad i, j = 0, 1, 2, \ldots. \tag{10.16}$$

Equation (10.15) represents the probability that if i faults have been corrected at time zero, j faults are corrected by time t after the next software failure occurs. Therefore, based on Markov analysis by using the assumptions and stochastic quantities above, we have the software reliability function and the mean time between software failures for X_i $(i = 1, 2, \ldots)$ as

$$R_i(x) = \sum_{s=0}^{i-1} \binom{i-1}{s} p^s q^{i-1-s} \exp[-Dk^s x], \tag{10.17}$$

$$\mathrm{E}[X_i] = \int_0^\infty R_i(x)dx = \frac{(\frac{p}{k} + q)^{i-1}}{D}. \tag{10.18}$$

And if the initial fault content in the system, N, is specified, the expected cumulative number of faults debugged imperfectly up to time t is given by

$$M(t) = \frac{q}{p} \sum_{n=1}^{N} \sum_{i=0}^{n-1} A_{i,n}(1 - \exp[-pDk^i t]), \tag{10.19}$$

where $A_{i,n}$ is

$$\left.\begin{aligned} A_{0,1} &\equiv 1 \\ A_{i,n} &= \frac{k^{(1/2)n(n-1)-i}}{\prod_{\substack{j=0 \\ j \neq i}}^{n-1} (k^j - k^i)}, \quad n = 2, 3, \ldots; i = 0, 1, 2, \ldots, n-1 \end{aligned}\right\}. \tag{10.20}$$

10.4.2 Imperfect debugging model for introduced faults

Besides the imperfect debugging factor above in fault-correction activities, we consider the possibility of introducing new faults in the debugging process. It is assumed that the following two kinds of software failures exist in the dynamic environment [36], [37], *i.e.* the testing or user operation phase:

(F1) software failures caused by faults originally latent in the software system prior to the testing (which are called inherent faults),
(F2) software failures caused by faults introduced during the software operation owing to imperfect debugging.

In addition, it is assumed that one software failure is caused by one fault and that it is impossible to discriminate whether the fault that caused the software failure that has occurred is F1 or F2. As to the software failure-occurrence rate due to F1, the inherent faults are detected with the progress of the operation time. In order to consider two kinds of time dependencies on the decreases of F1, let $a_i(t)(i = 1, 2)$ denote the software failure-occurrence rate for F1. On the other hand, the software failure-occurrence rate due to F2 is denoted as constant $\lambda(\lambda > 0)$, since we assume that F2 occurs randomly throughout the operation. When we consider the software failure-occurrence phenomena due to F1 and F2 simultaneously, the software failure-occurrence rate at operation time t is given by

$$h_i(t) = \lambda + a_i(t), \qquad i = 1, 2. \tag{10.21}$$

From (10.21), the expected cumulative number of software failures in the time interval $(0, t]$ (or the expected cumulative number of detected faults) is given by

$$\left.\begin{aligned} H_i(t) &= \lambda t + A_i(t), \\ A_i(t) &= \int_0^t a_i(x)dx, \qquad i = 1, 2 \end{aligned}\right\}. \tag{10.22}$$

Then, we have two imperfect debugging models based on an NHPP discussed in Section 10.3, where $h_i(t)$ in (10.21) and $H_i(t)$ in (10.22) are used as the intensity functions and the mean value functions $(i = 1, 2)$ for an NHPP, respectively. Especially, exponential and delayed S-shaped software reliability growth models are assumed for describing software failure-occurrence phenomena attributable to the inherent faults as (see Table 10.1)

$$a_1(t) = abe^{-bt}, \qquad a > 0, b > 0, \tag{10.23}$$

$$a_2(t) = ab^2te^{-bt}, \qquad a > 0, b > 0, \tag{10.24}$$

where a is the expected number of initially latent inherent faults and b the software failure-occurrence rate per inherent fault. Therefore, the mean value functions of NHPP models for the imperfect debugging factor are given by

$$H_1(t) = \lambda t + a(1 - e^{-bt}), \tag{10.25}$$

$$H_2(t) = \lambda t + a[1 - (1 + bt)e^{-bt}]. \tag{10.26}$$

From these imperfect debugging models we can derive several software reliability measures for the next software failure-occurrence time interval X since current time t, such as the software reliability function $R_i(x|t)$, the software hazard rate $z_i(x|t)$, and the mean time between software failures $\mathrm{E}_i[X|t](i = 1, 2)$:

$$R_i(x|t) = \exp[H_i(t) - H_i(t + x)], \qquad t \geq 0, x \geq 0, \tag{10.27}$$

$$z_i(x|t) = -\frac{d}{dx}R_i(x|t)/R_i(x|t) = h_i(t + x), \tag{10.28}$$

$$\mathrm{E}_i[X|t] = \int_0^\infty R_i(x|t)dx. \tag{10.29}$$

10.5 Software Availability Modeling

Recently, software performance measures such as the possible utilization factors have begun to be interesting for metrics as well as the hardware products. That is, it is very important to measure and assess *software availability*, which is defined as the probability that the software system is performing successfully, according to the specification, at a specified time point [38]–[40]. Several stochastic models have been proposed so far for software availability measurement and assessment. One group [41] has proposed a software availability model considering a reliability growth process, taking account of the cumulative number of corrected faults. Others [42]–[44] have constructed software availability models describing the uncertainty of fault removal. Still others [45] and [46] have incorporated the increasing difficulty of fault removal.

The actual operational environment needs to be more clearly reflected in software availability modeling, since software availability is a customer-oriented metrics. In [46] and [47] the development of a plausible model is described, which assumes that there exist two types of software failure occurring during the operation phase. Furthermore, in [48] an operational software availability model is built up from the viewpoint of restoration scenarios.

The above models have employed Markov processes for describing the stochastic time-dependent behaviors of the systems which alternate between the up state, operating regularly, and the restoration state (down state) when a system is inoperable [49]. Several stochastic metrics for software availability measurement in dynamic environment are derived from the respective models.

We discuss a fundamental software availability model [44] below.

10.5.1 Model description

The following assumptions are made for software availability modeling:

(1) The software system is unavailable and starts to be restored as soon as a software failure occurs, and the system cannot operate until the restoration action is complete.
(2) The restoration action implies debugging activity, which is performed perfectly with probability $a(0 < a \leq 1)$ and imperfectly with probability $b(= 1 - a)$. We call a the perfect debugging rate. One fault is corrected and removed from the software system when the debugging activity is perfect.
(3) When n faults have been corrected, the time to the next software failure occurrence and the restoration time follow exponential distributions with means of $1/\lambda_n$ and $1/\mu_n$, respectively.
(4) The probability that two or more software failures will occur simultaneously is negligble.

Consider a stochastic process $\{X(t), t \geq 0\}$ with the state space $(\boldsymbol{W}, \boldsymbol{R})$ where up state vector $\boldsymbol{W} = \{W_n; n = 0, 1, 2, \ldots\}$ and down state vector $\boldsymbol{R} = \{R_n; n = 0, 1, 2, \ldots\}$. Then, the events $\{X(t) = W_n\}$ and $\{X(t) = R_n\}$ mean that the system is operating and inoperable, respectively, due to the restoration action at time t, when n faults have already been corrected.

From assumption (2), when the restoration action has been completed in $\{X(t) = R_n\}$,

$$X(t) = \begin{cases} W_n, & \text{with probablity } b, \\ W_{n+1}, & \text{with probablity } a. \end{cases} \tag{10.30}$$

We use the Moranda model discussed in Section 10.3 to describe the software failure-occurrence phenomenon, *i.e.* when n faults have been corrected, the software hazard rate λ_n is given by

$$\lambda_n = Dk^n, \qquad n = 0, 1, 2, \ldots; D > 0, 0 < k < 1. \tag{10.31}$$

The expression of (10.31) comes from the point of view that software reliability depends on the debugging efforts, not the residual fault content. We do not note how many faults remain in the software system.

Next, we describe the time-dependent behavior of the restoration action. The restoration action for software systems includes not only the data recovery and the program reload, but also the debugging activities for manifested faults. From the viewpoint of the complexity, there are cases where the faults detected during the early stage of the testing or operation phase have low complexity and are easy to correct/remove, and as the testing is in progress, detected faults have higher complexity and are more difficult to correct/remove [8]. In the above case, it is appropriate that the mean restoration time becomes longer with the increasing number of corrected faults. Accordingly, we express μ_n as follows:

$$\mu_n = Er^n, \qquad n = 0, 1, 2, \ldots; E > 0, 0 < r \leq 1, \tag{10.32}$$

where E and r are the initial restoration rate and the decreasing ratio of the restoration rate, respectively. In (10.32) the case of $r = 1$, *i.e.* $\mu_n = E$, means that the complexity of each fault is random.

Let T_n and $U_n (n = 0, 1, 2, \ldots)$ be the random variables representing the next software failure occurrence and the next restoration time intervals when n faults have been corrected, in other words the sojourn times in states W_n and R_n, respectively. Furthermore, let $Y(t)$ be the random variable representing the cumulative number of faults corrected up to time t. The sample behavior of $Y(t)$ is illustrated in Figure 10.7. It is noted that the cumulative number of corrected faults is not always coincident with that of software failures or restoration actions. The sample state transition diagram of $X(t)$ is illustrated in Figure 10.8.

10.5.2 Software availability measures

We can obtain the state occupancy probabilities that the system is in states W_n and R_n at time point t as

$$\begin{aligned} P_{W_n}(t) &\equiv \Pr\{X(t) = W_n\} \\ &= \frac{g_{n+1}(t)}{a\lambda_n} + \frac{g'_{n+1}(t)}{a\lambda_n\mu_n}, \qquad n = 0, 1, 2, \ldots, \end{aligned} \tag{10.33}$$

$$\begin{aligned} P_{R_n}(t) &\equiv \Pr\{X(t) = R_n\} \\ &= \frac{g_{n+1}(t)}{a\mu_n}, \qquad n = 0, 1, 2, \ldots, \end{aligned} \tag{10.34}$$

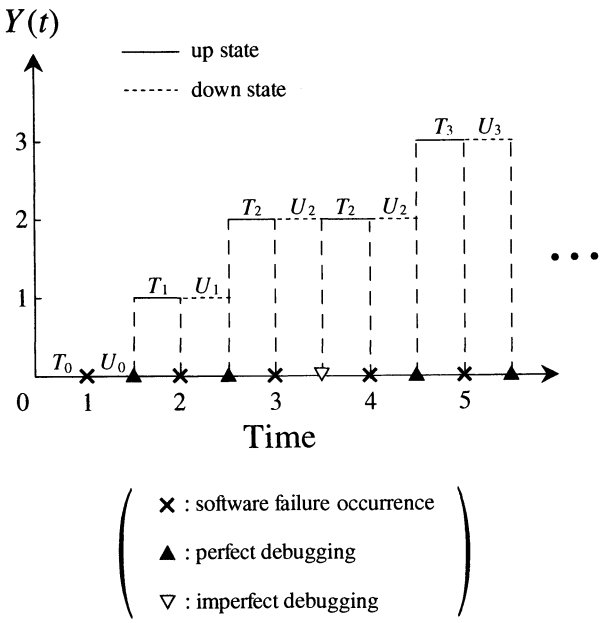


Fig. 10.7. A sample realization of $Y(t)$

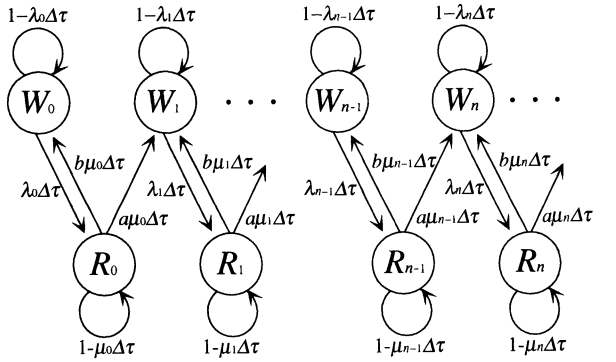


Fig. 10.8. A state transition diagram for software availability modeling

respectively, where $g_n(t)$ is the probability density function of random variable S_n, which denotes the first passage time to state W_n, and $g'_n(t) \equiv dg_n(t)/dt$. $g_n(t)$ and $g'_n(t)$ can be given analytically.

The following equation holds for arbitrary time t:

$$\sum_{n=0}^{\infty}[P_{W_n}(t) + P_{R_n}(t)] = 1. \tag{10.35}$$

The *instantaneous availability* is defined as

$$A(t) \equiv \sum_{n=0}^{\infty} P_{W_n}(t), \tag{10.36}$$

which represents the probability that the software system is operating at specified time point t. Furthermore, the *average software availability* over $(0, t]$ is defined as

$$A_{av}(t) \equiv \frac{1}{t}\int_0^t A(x)dx, \tag{10.37}$$

which represents the ratio of system's operating time to the time interval $(0, t]$. Using (10.33) and (10.34), we can express (10.36) and (10.37) as

$$\begin{aligned} A(t) &= \sum_{n=0}^{\infty}\left[\frac{g_{n+1}(t)}{a\lambda_n} + \frac{g'_{n+1}(t)}{a\lambda_n\mu_n}\right] \\ &= 1 - \sum_{n=0}^{\infty}\frac{g_{n+1}(t)}{a\mu_n}, \end{aligned} \tag{10.38}$$

$$\begin{aligned} A_{av}(t) &= \frac{1}{t}\sum_{n=0}^{\infty}\left[\frac{G_{n+1}(t)}{a\lambda_n} + \frac{g_{n+1}(t)}{a\lambda_n\mu_n}\right] \\ &= 1 - \frac{1}{t}\sum_{n=0}^{\infty}\frac{G_{n+1}(t)}{a\mu_n}, \end{aligned} \tag{10.39}$$

respectively, where $G_n(t)$ is the distribution function of S_n.

10.6 Application of Software Reliability Assessment

It is very important to apply the results of software reliability assessment to management problems with software projects for attaining higher productivity and quality. We discuss three software management problems as application technologies of software reliability models.

10.6.1 Optimal software release problem

Recently, it is becoming increasingly difficult for the developers to produce highly reliable software systems efficiently. Thus, it has been necessary to control a software development process in terms of quality, cost, and release time. In the last phase of the software development process, testing is carried out to detect and fix software faults introduced by human work, prior to its release for the operational use. The software faults that cannot be detected and fixed remain in the released software system after the testing phase. Thus, if a software failure occurs during the operational phase, then a computer system stops working and it may cause serious damage in our daily life.

If the duration of software testing is long, we can remove many software faults in the system and its reliability increases. However, this increases the testing cost and delays software delivery. In contrast, if the length of software testing is short, a software system with low reliability is delivered and it includes many software faults which have not been removed in the testing phase. Thus, the maintenance cost during the operation phase increases.

It is therefore very important in terms of software management that we find the optimal length of software testing, which is called an *optimal release time.* Such a decision problem is called an *optimal software release problem* [50]–[57]. These decision problems have been studied in the last decade by many researchers. We discuss optimal software release problems which consider both a present value and a warranty period (in the operational phase) during which the developer has to pay the cost for fixing any faults detected. It is very important with respect to software development management, then that we solve the problem of an optimal software testing time by integrating the total expected maintenance cost and the reliability requirement.

10.6.1.1 Maintenance cost model. The following notations are defined:

c_0 = the cost for the minimum quantity of testing which must be done,
c_t = the testing cost per unit time,
c_w = the maintenance cost per one fault during the warranty period,
T = the software release time, *i.e.* additional total testing time,
T^* = the optimum software release time.

We discuss a *maintenance cost model* for formulation of the optimal release problem. The maintenance cost during the warranty period is considered. The concept of a present value is also introduced into the cost factors. Then, the total expected software maintenance cost $WC(T)$ can be formulated as:

$$WC(T) \equiv c_0 + c_t \int_0^T e^{-\alpha t} dt + C_w(T), \tag{10.40}$$

where $C_w(T)$ is the maintenance cost during the warranty period. The parameter α in (10.40) is a discount rate of the cost. When we apply an exponential software reliability growth model based on an NHPP with mean value function $m(t)$ and intensity function $h_m(t)$ discussed in Section 10.3 (see Table 10.1), we discuss the following three cases in terms of the behavior of $C_w(T)$ (see Figure 10.9):

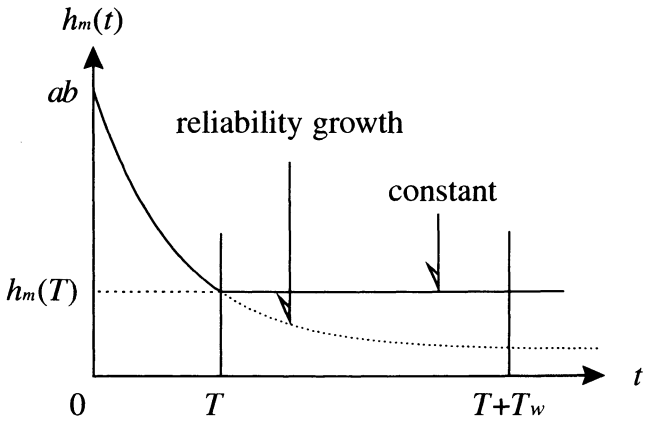


Fig. 10.9. Software reliability growth aspects during the warranty period

(Case 1)
When the length of the warranty period is constant and the software reliability growth is not assumed to occur after the testing phase, $C_w(T)$ is represented as:

$$C_w(T) = c_w \int_T^{T+T_w} h_m(T) e^{-\alpha t} dt. \tag{10.41}$$

(Case 2)
When the length of the warranty period is constant and the software reliability growth is assumed to occur even after testing, $C_w(T)$ is given by:

$$C_w(T) = c_w \int_T^{T+T_w} h_m(t) e^{-\alpha t} dt. \tag{10.42}$$

(Case 3)
When the length of the warranty period obeys a distribution function $W(t)$ and the software reliability growth is assumed to occur even after the testing phase, $C_w(T)$ is represented as:

$$C_w(T) = c_w \int_0^{\infty} \int_T^{T+T_w} h_m(t) e^{-\alpha t} dt dW(T_w), \tag{10.43}$$

where we assume that the distribution of the warranty period is a truncated normal distribution:

$$\frac{dW(t)}{dt} = \frac{1}{A\sqrt{2\pi}\sigma} \exp[-(t-\mu)^2/(2\sigma^2)], \quad t \geq 0, \mu > 0, \sigma > 0, \tag{10.44}$$

$$A = \frac{1}{\sqrt{2\pi}\sigma} \int_0^{\infty} \exp[-(t-\mu)^2/(2\sigma^2)] dt. \tag{10.45}$$

Let us consider the optimal release policies for minimizing $WC(T)$ in (10.40) with respect to T of Case 1, which is a typical case for optimal software release problems. Substituting (10.41) into (10.40), we rewrite it as:

$$WC(T) = c_0 + c_1 \int_0^T e^{-\alpha t} dt + c_w h_m(T) \int_T^{T+T_w} e^{-\alpha t} dt. \tag{10.46}$$

Differentiating (10.46) in terms of T and equating it to zero yields:

$$h_m(T) = \frac{c_t}{c_w T_w (b+\alpha)}. \tag{10.47}$$

Note that $WC(T)$ is a convex function with respect to T because $d^2WC(T)/dT^2 > 0$. Thus, the equation $dWC(T)/dT = 0$ has only one finite solution when the condition $h_m(0) > c_t/[c_w T_w(b+\alpha)]$ holds. The solution T_1 of (10.47) and the optimum release time can be shown as follows:

$$T^* = T_1 = \frac{1}{b} \ln \left[\frac{abc_w T_w (b+\alpha)}{c_t} \right], \qquad 0 < T_1 < \infty. \tag{10.48}$$

When the condition $h_m(0) \leq c_t/[c_w T_w(b+\alpha)]$ holds, $WC(T)$ in (10.46) is a monotonically increasing function in terms of the testing time T. Then, the optimum release time $T^* = 0$. Therefore, we can obtain the optimal release policies as follows:

[Optimal Release Policy 1]

(1.1) If $h_m(0) > c_t/[c_w T_w(b+\alpha)]$, then the optimum release time is $T^* = T_1$.
(1.2) If $h_m(0) \leq c_t/[c_w T_w(b+\alpha)]$, then the optimum release time is $T^* = 0$.

Similarly, we can obtain the optimal release policies for Case 2 and Case 3 [55], [58].

10.6.1.2 Maintenance cost model with reliability requirement.

Next, we discuss the optimal release problem with the requirement for software reliability. In the actual software development, the manager must spend and control the testing resources with a view to minimizing the total software cost and satisfying reliability requirements rather than only minimizing the cost. From the exponential software reliability growth model, the software reliability function can be defined as the probability that a software failure does not occur during the time interval $(T, T+x]$ after the total testing time T, *i.e.* the release time. The software reliability function is given as follows:

$$R(x|T) = \exp[-\{m(T+x) - m(T)\}]. \tag{10.49}$$

From (10.49), we derive the software reliability function as follows:

$$R(x|T) = \exp[-e^{-bT} \cdot m(x)]. \tag{10.50}$$

Let the software reliability objective be $R_0 (0 < R_0 \leq 1)$. We can evaluate optimum release time $T = T^*$ which minimizes (10.40) while satisfying the software reliability objective R_0. Thus, the optimal software release problem is formulated as follows:

$$\text{minimize } WC(T) \qquad \text{subject to} \qquad R(x|T) \geq R_0. \tag{10.51}$$

For the optimal release problem formulated by (10.51), let T_R be the optimum release time with respect to T satisfying the relation $R(x|T) = R_0$ for specified x. By applying the relation $R(x|T) = R_0$ into (10.50), we can obtain the solution T_R as follows:

$$T_R = \frac{1}{b}\left\{\ln m(x) - \ln \ln\left(\frac{1}{R_0}\right)\right\}. \tag{10.52}$$

Then, we can derive the optimal release policies to minimize the total expected software maintenance cost and to satisfy the software reliability objective R_0.

For Case 1,the optimal release policies are given as follows:

[Optimal Release Policy 2]

(2.1) If $h_m(0) > c_t/[c_w T_w(b+\alpha)]$ and $R(x|0) < R_0$,then the optimum release time is $T^* = \max\{T_1, T_R\}$.
(2.2) If $h_m(0) > c_t/[c_w T_w(b+\alpha)]$ and $R(x|0) \geq R_0$,then the optimum release time is $T^* = T_1$.
(2.3) If $h_m(0) \leq c_t/[c_w T_w(b+\alpha)]$ and $R(x|0) < R_0$,then the optimum release time is $T^* = T_R$.
(2.4) If $h_m(0) \leq c_t/[c_w T_w(b+\alpha)]$ and $R(x|0) \geq R_0$,then the optimum release time is $T^* = 0$.

Similarly, we can obtain the optimal release policies for Case 2 and Case 3 [58].

10.6.2 Statistical software testing-progress control

As well as quality/reliability assessment, software-testing managers should assess the degree of testing progress. We can construct a statistical method for software testing-progress control based on a *control chart* method as follows [6], [59]. This method is based on several instantaneous fault-detection rates derived from software reliability growth models based on an NHPP. For example, the intensity function based on the delayed S-shaped software reliability growth model in Section 10.3 (see Table 10.1) is given by

$$h_M(t) = \frac{dM(t)}{dt} = ab^2 te^{-bt}, \qquad a > 0, b > 0. \tag{10.53}$$

From (10.53), we can derive

$$\ln Z_M(t) = \ln a + 2 \cdot \ln b - bt, \tag{10.54}$$

$$Z_M(t) = \frac{h_M(t)}{t}. \tag{10.55}$$

The mean value of the instantaneous fault-detection rate represented by (10.55) is defined as the *average-instantaneous fault-detection rate.* Equation (10.54) means that the relation between the logarithm value of $Z_M(t)$ and the testing time has a linear property. If the testing phase progresses smoothly and

the reliability growth is stable in the testing, the logarithm of the average-instantaneous fault-detection rate decreases linearly with the testing time. From (10.54), we can also estimate the unknown parameters a and b by the method of least squares, and assess the testing progress by applying a regression analysis to the observed data. It is assumed that the form of the data is $(t_k, Z_k)(k = 1, 2, \ldots, n)$ where t_k is the kth testing time and Z_k is the realization of average-instantaneous fault-detection rate $Z_M(t)$. Letting the estimated unknown parameters be $\hat{a}$ and $\hat{b}$, we obtain the estimator of $Y(= \ln Z_M(t))$ as follows:

$$\hat{Y} = \ln \hat{Z}_M(t) = \ln \hat{a} + 2 \cdot \ln \hat{b} - \hat{b}t = \bar{Y} - \hat{b}(t - \bar{t}), \tag{10.56}$$

where

$$\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k, \qquad Y_k = \ln Z_k, \qquad \bar{t} = \frac{1}{n}\sum_{k=1}^{n} t_k, \quad k = 1, 2, \ldots, n.$$

The variation, which is explained as the regression to the dependent variable, Y, is

$$S_b = \sum_{k=1}^{n} (\hat{Y}_k - \bar{Y})^2 = \hat{b}^2 \sum_{k=1}^{n} (t_k - \bar{t})^2. \tag{10.57}$$

On the other hand, the error-variation not explained as the regression is represented as the summation of residual squares. That is,

$$S_e = \sum_{k=1}^{n} (Y_k - \hat{Y}_k)^2. \tag{10.58}$$

The unbiased variances from (10.57) and (10.58) are:

$$V_b = S_b, \qquad V_e = \frac{S_e}{n-2}. \tag{10.59}$$

With reference to (10.56), we discuss the logarithm of average-instantaneous fault-detection rate $Y_0 = \ln Z_M(t_0)$ at $t = t_0 (t_0 \geq t_n)$ by using the results of the analysis of variance. The $100(1 - \alpha)$ percent confidence interval to $\hat{Y}_0$ is given by

$$\begin{aligned} &\hat{Y}_0 \pm t\left(n - 2, 1 - \frac{\alpha}{2}\right)\sqrt{\mathrm{Var}[\hat{Y}_0]}, \\ &\mathrm{Var}[\hat{Y}_0] = \left\{1 + \frac{1}{n} + \frac{(t_0 - \bar{t})^2}{\sum_{k=1}^{n}(t_k - \bar{t})^2}\right\} V_e. \end{aligned} \tag{10.60}$$

$\mathrm{Var}[\hat{Y}_0]$ in (10.60) is the variance of $\hat{Y}_0$. $t(h, p)$ in (10.60) is $100p$ percent point of t-distribution at degree of freedom h. We now make the control chart which consists of the center line by the logarithm of the average-instantaneous fault-detection rate, and the upper and lower control limits which are given by

(10.60). We can assess the testing progress by applying a regression analysis to the observed data.

The testing-progress assessment indices for the other NHPP models are given by the following intensity function:

- $h_m(t) = abe^{-bt}$ with relation $\ln h_m(t) = (\ln a + \ln b) - bt$ (for the exponential software reliability growth model),
- $h_\mu(t) = \lambda_0/(\lambda_0\theta t + 1)$ with relation $\ln h_\mu(t) = \ln\lambda_0 - \theta\mu(t)$ (for the logarithmic Poisson execution time model),
- $h_\lambda(t) = \lambda\beta t^{\beta-1}$ with relation $\ln h_\lambda(t) = (\ln\lambda + \ln\beta) + (\beta - 1)\ln t$ (for the Weibull process model [6], [8]).

The procedure of testing-progress control is shown as follows:

Step 1: An appropriate model is selected to apply and the model parameters are estimated by the method of least squares.
Step 2: To certify goodness-of-fit of the estimated regression equation for the observed data, we use the F-test.
Step 3: Based on the result of the F-test, the center line and upper and lower control limits of the control chart are calculated. The control chart is drawn.
Step 4: The observed data are plotted on the control chart and the stability of the testing progress is judged.

10.6.3 Optimal testing-effort allocation problem

We discuss a management problem to achieve a reliable software system efficiently during module testing in the software development process by applying a testing-effort-dependent software reliability growth model based on an NHPP (see Table 10.1). We take account of the relationship between the testing effort spent during the module testing and the detected software faults where the testing effort is defined as resource expenditures spent on software testing, *e.g.* manpower, CPU hours, and executed test cases. The software development manager has to decide how to use the specified testing effort effectively in order to maximize the software quality and reliability [60]. That is, to develop a quality and reliable software system, it is very important for the manager to allocate the specified amount of testing-effort expenditure for each software module under some constraints. We can observe the software reliability growth in the module testing in terms of a time-dependent behavior of the cumulative number of faults detected during the testing stage.

Based on the testing-effort-dependent software reliability growth model, we consider the following testing-effort allocation problem [61], [62]:

(1) The software system is composed of M independent modules. The number of software faults remaining in each module can be estimated by the model.
(2) The total amount of testing-effort expenditure for module testing is specified.
(3) The manager has to allocate the specified total testing-effort expenditure to each software module so that the number of software faults remaining in the system may be minimized.

The following are defined:

a = the expected initial fault content,
r = the fault-detection rate per unit of testing-effort expenditure $(0 < r < 1)$,
i = the subscript for each software module number $i = 1, 2, \ldots, M$,
w_i = the weight for each module$(w_i > 0)$,
n_i = the expected number of faults remaining in each module,
q_i, Q = the amount of testing-effort expenditure for each module to be allocated and the total testing-effort expenditure before module testing $(q_i \geq 0, Q > 0)$.

From (10.8) and Table 10.1, *i.e.* $n(t) = a \cdot \exp[-rW(t)]$, the estimated number of remaining faults for module i is formulated by

$$n_i = a_i \cdot \exp[-r_i q_i], \qquad i = 1, 2, \ldots, M. \tag{10.61}$$

Thus, the *optimal testing-effort allocation problem* is formulated as:

$$\text{minimize} \sum_{i=1}^{M} w_i n_i = \sum_{i=1}^{M} w_i a_i \cdot \exp[-r_i q_i], \tag{10.62}$$

$$\text{so that} \sum_{i=1}^{M} q_i \leq Q, \quad q_i \geq 0, \qquad i = 1, 2, \ldots, M, \tag{10.63}$$

where it is supposed that the parameter a_i and r_i have already been estimated by the model.

To solve the problem above, we consider the following Lagrangian:

$$L = \sum_{i=1}^{M} w_i a_i \cdot \exp[-r_i q_i] + \lambda \left(\sum_{i=1}^{M} q_i - Q \right), \tag{10.64}$$

and the necessary and sufficient conditions [63] for the minimum are

$$\left.\begin{array}{lll} \dfrac{\partial L}{\partial q_i} & = -w_i a_i r_i \cdot \exp[-r_i q_i] + \lambda \geq 0, & \\ q_i \cdot \dfrac{\partial L}{\partial q_i} & = 0, & i = 1, 2, \ldots, M, \\ \displaystyle\sum_{i=1}^{M} q_i & = Q, & \\ q_i & \geq 0, & i = 1, 2, \ldots, M, \end{array}\right\} \tag{10.65}$$

where λ is a Lagrange multiplier.

Without loss of generality, setting $A_i = w_i a_i r_i (i = 1, 2, \ldots, M)$, we can assume that the following condition is satisfied for the tested modules:

$$A_1 \geq A_2 \geq \cdots \geq A_{k-1} \geq A_k \geq A_{k+1} \geq \cdots \geq A_M. \tag{10.66}$$

This means that it is arranged in order of fault detectability for the tested modules. Now, if $A_k > \lambda \geq A_{k+1}$, from (10.65) we have

$$q_i = \max\left\{0, \frac{1}{r_i}(\ln A_i - \ln \lambda)\right\},$$

i.e.

$$\left.\begin{array}{ll} q_i = \dfrac{1}{r_i}(\ln A_i - \ln \lambda), & i = 1, 2, \ldots, k, \\ q_i = 0, & i = k+1, \ldots, M, \end{array}\right\}. \tag{10.67}$$

From (10.65) and (10.67), $\ln \lambda$ is given by

$$\ln \lambda = \frac{\sum_{i=1}^{k} \frac{1}{r_i} \ln A_i - Q}{\sum_{i=1}^{k} \frac{1}{r_i}}, \qquad k = 1, 2, \ldots, M. \tag{10.68}$$

Let λ_k denote the value of the right-hand side of (10.68). Then, the optimal Lagrange multiplier λ^* exists in the set $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$. Hence, we can obtain λ^* by the following procedures:

(1) Set $k = 1$.
(2) Compute λ_k by (10.68).
(3) If $A_k > \lambda_k \geq A_{k+1}$, then $\lambda^* = \lambda_k$(stop). Otherwise, set $k = k + 1$ and go back to (2).

The optimal solutions $q_i^*(i = 1, 2, \ldots, M)$ are given by

$$\left.\begin{array}{ll} q_i^* = \dfrac{1}{r_i}(\ln A_i - \ln \lambda^*), & i = 1, 2, \ldots, k, \\ q_i^* = 0, & i = k+1, \ldots, M, \end{array}\right\}, \tag{10.69}$$

which means that the amount of testing-effort expenditure is needed more for the tested modules containing more faults.

References

1. Kanno, A. (1992), Introduction to Software Production Engineering (in Japanese). JUSE Press, Tokyo
2. Matsumoto, Y. and Ohno, Y. (eds.) (1989), Japanese Perspectives in Software Engineering. Addison-Wesley, Singapore
3. Lyu, M. R. (ed.) (1996), Handbook of Software Reliability Engineering. IEEE Computer Society Press, Los Alamitos, California
4. Pham, H. (2000), Software Reliability. Springer-Verlag, Singapore
5. Yamada, S. and Ohtera, H. (1990), Software Reliability: Theory and Practical Application (in Japanese). Soft Research Center, Tokyo
6. Yamada, S. (1994), Software Reliability Models: Fundamentals and Applications (in Japanese). JUSE Press, Tokyo
7. Yamada, S. and Takahashi, M. (1993), Introduction to Software Management Model (in Japanese). Kyoritsu-shuppan, Tokyo
8. Musa, J. D., Iannino, A. and Okumoto, A. (1987), Software Reliability: Measurement, Prediction, Application. McGraw-Hill, New York
9. Ramamoorthy, C. V. and Bastani, F. B. (1982), "Software reliability- status and perspectives," *IEEE Transactions on Software Engineering*, **SE-8**, 354–371
10. Jelinski, Z. and Moranda, P. B. (1972), Software Reliability Research, in Statistical Computer Performance Evaluation (Freiberger W. ed.). 465–484, Academic Press, New York
11. Wagoner, W. L. (1973), "The final report on a software reliability measurement study,"*Report TOR-0074(4112)-1*, Aerospace Corporation
12. Moranda, P. B. (1979), "Event-altered rate models for general reliability analysis," *IEEE Transactions on Reliability*, **R-28**, 376–381
13. Ascher, H. and Feingold, H. (1984), Repairable Systems Reliability: Modeling, Inference, Misconceptions, and Their Causes. Marcel Dekker, New York
14. Yamada, S. (1991), "Software quality/reliability measurement and assessment: software reliability growth models and data analysis," *Journal of Information Processing*, **14**, 254–266
15. Yamada, S. and Osaki, S. (1985), "Software reliability growth modeling: models and applications," *IEEE Transactions on Software Engineering*, **SE-11**, 1431–1437
16. Goel, A. L. and Okumoto, K. (1979), "Time-dependent error-detection rate model for software reliability and other performance measures," *IEEE Transactions on Reliability*, **R-28**, 206–211
17. Goel, A. L. (1980), "Software error detection model with applications," *Journal of Systems and Software*, **1**, 243–249

18. Yamada, S. and Osaki, S. (1984), " Nonhomogeneous error detection rate models for software reliability growth," in Stochastic Models in Reliability Theory (Osaki, S. and Hatoyama, Y. eds.). 120–143, Springer-Verlag, Berlin
19. Yamada, S., Osaki, S. and Narihisa, H. (1985), "A software reliability growth model with two types of errors," *R. A. I. R. O. Operations Research*, **19**, 87–104
20. Yamada, S., Ohba, M. and Osaki, S. (1983), "S-shaped reliability growth modeling for software error detection," *IEEE Transactions on Reliability*, **R-32**, 475–478, 484
21. Yamada, S., Ohba, M. and Osaki, S. (1984), " S-shaped software reliability growth models and their applications," *IEEE Transactions on Reliability*, **R-33**, 289–292
22. Ohba, M. (1984), "Inflection S-shaped software reliability growth model," in Stochastic Models in Reliability Theory (Osaki, S. and Hatoyama, Y. eds.). 144–162, Springer-Verlag, Berlin
23. Ohba, M. and Yamada, S. (1984), "S-shaped software reliability growth models," *Proceedings of the 4th International Conference on Reliability and Maintainability*, 430–436
24. Yamada, S., Ohtera, H., Narihisa, H. (1986), "Software reliability growth models with testing-effort," *IEEE Transactions on Reliability*, **R-35**, 19–23
25. Yamada, S., Hishitani, J. and Osaki, S. (1993), "Software-reliability growth with a Weibull test-effort function," *IEEE Transactions on Reliability*, **R-42**, 100–106
26. Ohtera, H., Yamada, S. and Ohba, M. (1990), "Software reliablity growth model with testing-domain and comparisons of goodness-of-fit," *Proceedings of the International Symposium on Reliability and Maintainability*, 289–294
27. Yamada, S., Ohtera, H. and Ohba, M. (1992), "Testing-domain dependent software reliability growth models," *Computers & Mathematics with Applications*, **24**, 79–86
28. Musa, J. D. and Okumoto, K. (1984), " A logarithmic Poisson execution time model for software reliability measurement," *Proceedings of the 7th International Conference on Software Engineering*, 230–238
29. Okumoto, K. (1985), "A statistical method for software quality control," *IEEE Transactions on Software Engineering*, **SE-11**, 1424–1430
30. Kanno, A. (1979), Software Engineering (in Japanese). JUSE Press, Tokyo
31. Mitsuhashi, T. (1981), A Method of Software Quality Evaluation (in Japanese). JUSE Press, Tokyo
32. Shoomam, M. L. (1983), Software Engineering: Design, Reliability, and Management. McGraw-Hill, New York
33. Ohba, M. and Chou, X. (1989), "Does imperfect debugging affect software reliability growth?," *Proceedings of the 11th International Conference on Software Engineering*, 237–244
34. Shanthikumar, J. G. (1981), "A state- and time-dependent error occurrence-rate software reliability model with imperfect debugging," *Proceedings of the National Computer Conference*, 311–315
35. Ross, S. M. (1996), Stochastic Processes. John Wiley & Sons, New York
36. Yamada, S. and Miki, T. (1998), "Imperfect debugging models with introduced software faults and their comparisons (in Japanese)," *Transactions of IPS Japan*, **39**, 102–110
37. Yamada, S. (1998). " Software reliability growth models incorporating imperfect debugging with introduced faults," *Electronics and Communications in Japan*, **81**, 33–41

38. Xie, M. (1991), Software Reliability Modelling. World Scientific, Singapore
39. Laprie, J. -C., Kanoun, K., Béounes, C. and Kaâniche, M. (1991), "The KAT (Knowledge-Action-Transformation) approach to the modeling and evaluation of reliability and availability growth," *IEEE Transactions on Software Engineering*, **17**, 370–382
40. Laprie, J. -C. and Kanoun, K. (1992), "X-ware reliability and availability modeling," *IEEE Transactions on Software Engineering*, **18**, 130–147
41. Tokuno, K. and Yamada, S. (1995), "A Markovian software availability measurement with a geometrically decreasing failure-occurrence rate," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, **E78-A**, 737–741
42. Okumoto, K. and Goel, A. L. (1978), "Availability and other performance measures for system under imperfect maintenance," *Proceedings of the COMPSAC '78*, 66–71
43. Kim, J. H., Kim, Y. H. and Park, C. J. (1982), "A modified Markov model for the estimation of computer software performance," *Operations Research Letters*, **1**, 253–257
44. Tokuno, K. and Yamada, S. (1997), "Markovian software availability modeling for performance evaluation," in Stochastic Modelling in Innovative Manufacturing (Christer, A. H., Osaki, S. and Thomas, L. C. eds.). 246–256, Springer-Verlag, Berlin
45. Tokuno, K. and Yamada, S. (1997), "Software availability model with a decreasing fault-correction rate (in Japanese)," *Journal of Reliability Engineering Association of Japan*, **19**, 3–12
46. Tokuno, K. and Yamada, S. (1997), "A Markovian modeling for software availability measurement (in Japanese)," *Journal of Japan Society for Software Science and Technology*, **14**, 38–44
47. Tokuno, K. and Yamada, S. (1997), "Markovian software availability modeling with two types of software failures for operational use," *Proceedings of the 3rd ISSAT International Conference on Reliability and Quality in Design*, 97–101
48. Tokuno, K. and Yamada, S. (1998), "A Markovian software availability model for operation use (in Japanese)," *Journal of Japan Society for Software Science and Technology*, **15**, 17–24
49. Tokuno, K. and Yamada, S. (1998), "Operational software availability measurement with two kinds of restoration actions," *Journal of Quality in Maintenance Engineering*, **4**, 273–283
50. Cho, B. C. and Park, K. S. (1994), "An optimal time for software testing under the user's requirement of failure-free demonstration before release," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, **E77-A**, 563–570
51. Foreman, E. H. and Singpurwalla, N. D. (1979), "Optimal time intervals for testing-hypotheses on computer software errors," *IEEE Transactions on Reliability*, **R-28**, 250–253
52. Kimura, M. and Yamada, S. (1995), "Optimal software release policies with random life-cycle and delivery delay," *Proceedings of the 2nd ISSAT International Conference on Reliability and Quality in Design*, 215–219
53. Koch, H. S. and Kubat, P. (1983), "Optimal release time for computer software," *IEEE Transactions on Software Engineering*, **SE-9**, 323–327
54. Okumoto, K. and Goel, A. L. (1980), "Optimum release time for software system based on reliability and cost criteria," *Journal of Systems and Software*, **1**, 315–318

55. Yamada, S. (1994), "Optimal release problems with warranty period based on a software maintenance cost model (in Japanese)," *Transactions of IPS Japan*, **35**, 2197–2202
56. Yamada, S., Kimura, M., Teraue, E. and Osaki, S. (1993), "Optimal software release problem with life-cycle distribution and discount rate (in Japanese)," *Transactions of IPS Japan*, **34**, 1188–1197
57. Yamada, S. and Osaki, S. (1987), "Optimal software release policies with simultaneous cost and reliability requirements," *European Journal of Operational Research*, **31**, 46–51
58. Kimura, M., Toyota, T. and Yamada, S. (1999), "Economic analysis of software release problems with warranty cost and reliability requirement," *Reliability Engineering and System Safety*, **66**, 49–55
59. Yamada, S. and Kimura, M. (1999), "Software reliability assessment tool based on object-oriented analysis and its application," *Annals of Software Engineering*, **8**, 223–238
60. Kubat, P. and Koch, H. S. (1983), "Managing test procedures to achieve reliable software," *IEEE Transactions on Reliability*, **R-32**, 299–303
61. Ohtera, H. and Yamada, S. (1990), "Optimal allocation and control problem for software testing-resources," *IEEE Transactions on Reliability*, **R-39**, 171–176
62. Yamada, S., Ichimori, T. and Nishiwaki N. (1995), "Optimal allocation policies for testing-resource based on a software reliability growth model," *Mathematical and Computer Modelling*, **22**, 295–301
63. Bazaraa, M. S. and Shetty, C. M. (1979), Nonlinear Programming: Theory and Algorithms. John Wiley & Sons, New York

11. Reliability Models in Data Communication Systems

Kazumi Yasui, Toshio Nakagawa
Department of Industrial Engineering,
Aichi Institute of Technology
Toyota 470-0392, Japan
and
Hiroaki Sandoh
Faculty of Information Science,
University of Marketing and Distribution Sciences
Kobe 651-2188, Japan

Summary.

We survey the data transmission models in a communication system from the viewpoint of reliability: data transmission often fails owing to errors generated by disconnections, noises or distortions in a communication line. To protect against such errors and to ensure accurate transmission of data, the following three schemes of error-control procedures are mainly used: (i) Forward-error-correction (FEC) scheme, (ii) automatic-repeat-request (ARQ) scheme, and (iii) hybrid ARQ schemes, which combines FEC with ARQ. ARQ schemes are mainly employed in data transmission systems to achieve high reliability of communication. Further, three protocols of stop-and-wait (SW), go-back-N (GBN) and selective-repeat (SR) have been well known in ARQ schemes. In this chapter, we formulate three typical stochastic models of SW ARQ, SR ARQ and hybrid ARQ schemes, and discuss analytically and numerically optimal policies to improve the data throughput.

Keywords : communication system, data transmission, error control, ARQ, retransmission number, mean time to success, throughput

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

11.1 Introduction

One of the important problems in a communication system is how to transmit the data accurately and rapidly to a recipient. However, errors in data transmission are unavoidable because of disturbing factors such as disconnections, noises or distortions in a communication line [1], [2]. Error-control procedures are indispensable to transmission of high-quality data, and their various techniques have been considered from the viewpoints of reliability and accuracy. The main techniques in error-correcting strategies are classified into the following three schemes: (i) forward-error-correction (FEC) scheme, (ii) automatic-repeat-request (ARQ) scheme and (iii) hybrid ARQ scheme.

An FEC scheme is the policy of transmitting the data together with error-detecting and error-correcting codes without making any retransmissions: the correcting code is used for the correction of expected error patterns which occur frequently. Since no retransmission is made, the data throughput of the system is held at a constant level, but the data transmission with unexpected error patterns may be passed without correcting errors. For this reason, this scheme is not used except in special cases.

An ARQ scheme is the policy of retransmitting the data when some errors have been detected: a receiver requires retransmission of the same data when errors have been detected in its data transmission. Since the coding and decoding for error corrections are omitted, the error-control procedure is simple, and can also cope easily with unexpected error patterns. However, the data throughput decreases significantly in the case where a large number of retransmissions are required. The three main protocols of stop-and-wait (SW), go-back-N (GBN) and selective-repeat (SR) have become familiar in ARQ schemes.

A hybrid ARQ scheme is the policy of combining FEC with ARQ to cover up defects of two schemes, and can be further divided into two types of schemes, type-I hybrid ARQ and type-II hybrid ARQ: this aims to reduce the number of retransmissions and to gain a constant data throughput even in a worse environment, by coding error patterns which occur frequently.

In the past, a variety of ideas for the above error-correcting strategies and their modified communication systems were proposed and studied, and a great many protocols of ARQ schemes have appeared in research reports. In recent years, the broadcast-type protocols of ARQ schemes have seen conspicuous development. As a result, research on ARQ protocols has progressed in the following directions[3], [4] :

1) Modifications of SW and GBN protocols to improve the data throughput are [4], [5], [6], [7], [8], [9], [10], [11], [12].
2) Modifications of SR protocols with finite buffer area [13], [14], [15], [16], [17], [18], [19].

3) Integrations of coding techniques into ARQ protocols are found in [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38].
4) Techniques of broadcast-type ARQ protocols are found in [39], [40], [3], [41], [42], [43], [44], [45],[46].
5) Queueing behaviors of ARQ protocols are found in [47], [48].
6) Methodologies in analyzing the performance of ARQ protocols are found in [49], [50], [51], [52], [53].

ARQ schemes are mainly used in data communication systems to achieve high reliability of communication. Here, we summarize three typical stochastic models of SW ARQ, SR ARQ and hybrid ARQ schemes and discuss analytically their optimal policies for improving the data throughput.

11.2 SW ARQ Model with Intermittent Faults

In this section, we formulate a stochastic model of ARQ scheme with SW protocols. The ARQ strategy is widely employed in point-to-point data transmission because its error control is easy and simple.

We model data transmissions with intermittent faults as follows [54], [55]: faults in a communication system occur intermittently and are hidden, and sometimes cause errors in data transmissions. Some faults occur repeatedly, and consequently become permanent failures from hidden faults [56].

It is assumed that when the system is in a normal condition, *i.e.*, no fault occurs, errors of data transmission occur with probability q_0. If some faults are hidden, errors of data transmission occur with probability q_1 $(q_1 \geq q_0)$. Further, if the duration of hidden faults exceeds a threshold time, faults become permanent failures and errors occur with probability q_2 $(q_2 > q_1)$.

When errors have occurred in a communication system, it is supposed that the data transmission will fail with certainty and an ARQ strategy is made. However, the data throughput decreases significantly if retransmissions are repeated without limitation. To keep the level of data throughput, if all numbers N of ARQ strategies have failed, the system is inspected and maintained. We repeat the above procedure until data transmission is successful. Forming the generating functions of probabilities that the transmission fails at k times successively, we derive the mean time to successful data transmission and discuss an optimal number N^* which minimizes it. Numerical examples are finally given for several values of q_0, q_1 and q_2.

11.2.1 Intermittent faults

Faults in a communication system occur according to an exponential distribution $(1 - e^{-\lambda t})$ and are hidden. If the duration time X of hidden faults exceeds an upper limit time Y, then faults become permanent failures, and

otherwise, faults are recovered from a hidden state. This indicates that if the event $\{X \leq Y\}$ occurs hidden faults disappear, and if the event $\{X > Y\}$ occurs faults become permanent failures. It is assumed that both random variables X and Y are independent and have exponential distributions, *i.e.*, $\Pr\{X \leq t\} = 1 - e^{-\mu t}$ and $\Pr\{Y \leq t\} = 1 - e^{-\theta t}$.

We define the following states of the above model :

State 0 : No fault occurs and the system is in a normal condition.
State 1 : Hidden fault occurs.
State 2 : Permanent failure occurs.

The mass functions (one-step transition probabilities) $Q_{ij}(t)$ from state i to state j of a Markov renewal process are

$$Q_{01}(t) = 1 - e^{-\lambda t}, \tag{11.1}$$

$$Q_{10}(t) = \int_0^t e^{-\theta x} \mu e^{-\mu x} dx = \frac{\mu}{\mu + \theta} \left[1 - e^{-(\mu+\theta)t}\right], \tag{11.2}$$

$$Q_{12}(t) = \int_0^t \theta e^{-\theta x} e^{-\mu x} dx = \frac{\theta}{\mu + \theta} \left[1 - e^{-(\mu+\theta)t}\right]. \tag{11.3}$$

Using the above mass functions, the transition probabilities that the system is in state j at time t, given that it was in state i at time 0 are, from [54],

$$P_{00}(t) = \frac{1}{\gamma_1 - \gamma_2} \left[(\mu + \theta - \gamma_2)e^{-\gamma_2 t} - (\mu + \theta - \gamma_1)e^{-\gamma_1 t}\right], \tag{11.4}$$

$$P_{01}(t) = \frac{\lambda}{\gamma_1 - \gamma_2} \left[e^{-\gamma_2 t} - e^{-\gamma_1 t}\right], \tag{11.5}$$

$$P_{10}(t) = \frac{\mu}{\gamma_1 - \gamma_2} \left[e^{-\gamma_2} - e^{-\gamma_1 t}\right], \tag{11.6}$$

$$P_{11}(t) = \frac{1}{\gamma_1 - \gamma_2} \left[(\lambda - \gamma_2)e^{-\gamma_2 t} - (\lambda - \gamma_1)e^{-\gamma_1 t}\right], \tag{11.7}$$

$$P_{i0}(t) + P_{i1}(t) + P_{i2}(t) = 1, \qquad i = 0, 1, \tag{11.8}$$

where

$$\gamma_1 \equiv \frac{1}{2}\left[\lambda + \mu + \theta + \sqrt{(\lambda + \mu + \theta)^2 - 4\lambda\theta}\right],$$

$$\gamma_2 \equiv \frac{1}{2}\left[\lambda + \mu + \theta - \sqrt{(\lambda + \mu + \theta)^2 - 4\lambda\theta}\right].$$

11.2.2 ARQ policy

We consider the following ARQ strategy, which detects errors in a communication system :

(1) We transmit a certain amount of data, which is named *unit data*, to a receiver. A time a is required for transmission of unit data.

(a) If there is no fault in a communication line, *i.e.*, the system is in state 0, errors of data transmission occur with probability q_0 $(0 \leq q_0 < 1)$. When errors do not occur with probability $1-q_0$, the data is transmitted correctly and we call it *success of data transmission.*

(b) If there are hidden faults, *i.e.*, the system is in state 1, errors of data transmission occur with probability q_1 $(q_0 \leq q_1 < 1)$ and do not with probability $1 - q_1$.

(c) If there are permanent failures, *i.e.*, the system is in state 2, errors occur with probability q_2 $(q_1 < q_2 \leq 1)$ and do not with probability $1 - q_2$.

(2) When errors have occurred, the data transmission fails. We call it *failure of data transmission.*

(3) We transmit the data until the total number of retransmissions and the first transmission reaches a specified number N or the data transmission succeeds, whichever occurs first. If the transmission fails at N times successively, the system is inspected and maintained. After that, the system can return to a normal condition and the same transmission is made again from the beginning. A time v is required for all times resulting from inspection, maintenance and the restart of transmission.

Let $Q_i(k)$ $(i = 0, 1)$ be the probabilities that the transmission fails at k $(k = 1, 2, \cdots)$ times successively, starting from state i, and $Q_i(0) \equiv 1$. Then, we have the following renewal equations :

$$Q_i(k) = P_{i0}(a)q_0 Q_0(k-1) + P_{i1}(a)q_1 Q_1(k-1) + P_{i2}(a)q_2^k, \qquad i = 0, 1. \quad (11.9)$$

Let us introduce the generating functions $V_i(z) \equiv \sum_{k=0}^{\infty} Q_i(k)z^k$ $(i = 0, 1)$. Then, forming the generating functions of (11.9),

$$\begin{aligned} V_i(z) &= 1 + \sum_{k=1}^{\infty} Q_i(k)z^k \\ &= 1 + P_{i0}(a)q_0 z V_0(z) + P_{i1}(a)q_1 z V_1(z) + P_{i2}(a)\frac{q_2 z}{1 - q_2 z}, \quad i = 0, 1. \end{aligned} \quad (11.10)$$

Solving (11.10) for $V_0(z)$, we have

$$V_0(z) = \frac{\left[\begin{array}{c} 1 - q_1 z[P_{11}(a) - P_{01}(a)] + q_2 z \\ \times \{P_{02}(a) - q_1 z[P_{11}(a)P_{02}(a) - P_{01}(a)P_{12}(a)]\}/(1 - q_2 z) \end{array}\right]}{[1 - q_0 z P_{00}(a)]\,[1 - q_1 z P_{11}(a)] - q_0 q_1 z^2 P_{01}(a) P_{10}(a)}. \tag{11.11}$$

Thus, from Appendix 2 in [55] and the definition of $V_0(z)$,

$$Q_0(k) = \frac{A_1}{s_1 - s_2} + \frac{A_2}{s_1 - s_2}, \qquad k = 0, 1, 2, \cdots, \tag{11.12}$$

where

$$\begin{aligned} s_1 &\equiv \frac{1}{2}\{q_0 P_{00}(a) + q_1 P_{11}(a) \\ &\quad + \sqrt{[q_0 P_{00}(a) - q_1 P_{11}(a)]^2 + 4 q_0 q_1 P_{01}(a) P_{10}(a)}\}, \\ s_2 &\equiv \frac{1}{2}\{q_0 P_{00}(a) + q_1 P_{11}(a) \\ &\quad - \sqrt{[q_0 P_{00}(a) - q_1 P_{11}(a)]^2 + 4 q_0 q_1 P_{01}(a) P_{10}(a)}\}, \end{aligned}$$

$$\begin{aligned} A_1 &\equiv \{s_1 - q_1[P_{11}(a) - P_{01}(a)]\}\, s_1^k - \{s_2 - q_1[P_{11}(a) - P_{01}(a)]\}\, s_2^k, \\ A_2 &\equiv q_2 \left\{ \left[s_1 P_{02}(a) - q_1[P_{11}(a)P_{02}(a) - P_{01}(a)P_{12}(a)] \right] \cdot \frac{s_1^k - q_2^k}{s_1 - q_2} \right. \\ &\quad \left. - \left[s_2 P_{02}(a) - q_1[P_{11}(a)P_{02}(a) - P_{01}(a)P_{12}(a)] \right] \cdot \frac{s_2^k - q_2^k}{s_2 - q_2} \right\}. \end{aligned}$$

It is evident that $Q_0(0) = 1$ and $Q_0(1) = q_0 P_{00}(a) + q_1 P_{01}(a) + q_2 P_{02}(a)$.

Using $Q_0(k)$, the probability that the transmission succeeds at the k-th time for the first time is

$$1 - Q_0(k) - [1 - Q_0(k-1)] = Q_0(k-1) - Q_0(k). \tag{11.13}$$

Hence, the expected number of transmissions until all N transmissions fail or some transmission succeeds is

$$\begin{aligned} E_0(N) &= \sum_{k=1}^{N} k[Q_0(k-1) - Q_0(k)] + N Q_0(N) \\ &= \sum_{k=0}^{N-1} Q_0(k), \qquad N = 1, 2, \cdots. \end{aligned} \tag{11.14}$$

Further, the mean time to the success of data transmission is given by the renewal function

$$M_0(N) = aE_0(N) + Q_0(N)[v + M_0(N)]. \tag{11.15}$$

Solving (11.15) for $M_0(N)$, we have

$$M_0(N) = \frac{a\sum_{k=0}^{N-1} Q_0(k) + vQ_0(N)}{1 - Q_0(N)}, \qquad N = 1, 2, \cdots. \tag{11.16}$$

11.2.3 Optimal retransmission number

We seek an optimal retransmission number N^*, including the first transmission, which minimizes $M_0(N)$ in (11.16). Let $q(N) \equiv Q_0(N+1)/Q_0(N)$ be the probability that the retransmission fails at the $(N+1)$-th time, given that it has failed at N times successively. Then, forming the inequality $M_0(N+1) - M_0(N) \geq 0$, we have

$$\frac{1 - Q_0(N)}{1 - q(N)} - \sum_{k=0}^{N-1} Q_0(k) \geq \frac{v}{a}, \qquad N = 1, 2, \cdots. \tag{11.17}$$

Denoting the left-hand side of (11.17) by $L(N)$,

$$L(N) - L(N-1) = [1 - Q_0(N)]\left[\frac{1}{1 - q(N)} - \frac{1}{1 - q(N-1)}\right].$$

Then, if $q(N)$ is strictly increasing in N, then $L(N)$ is also strictly increasing in N. Further, we easily have

$$L(N) > \frac{1 - Q_0(1)}{1 - q(N)} - 1, \qquad N = 2, 3, \cdots. \tag{11.18}$$

Hence, if $q(\infty) \equiv \lim_{N\to\infty} q(N) \geq [v + aQ_0(1)]/(a + v)$, then there exists a finite N^* which satisfies (11.17).

Therefore, we have the following optimal policy:

(i) If $L(1) < v/a < L(\infty)$ then there exists a unique minimum N^* $(2 \leq N^* < \infty)$ which satisfies (11.17).
(ii) If $L(1) \geq v/a$ then $N^* = 1$, *i.e.*, when the first data transmission has failed, a communication system is inspected immediately.
(iii) If $L(\infty) \leq v/a$ then $N^* = \infty$, *i.e.*, the system is not inspected at all.

11.2.4 Numerical examples and remarks

Suppose that the mean time of fault occurrences is $1/\lambda = 21,600a$, the mean duration of hidden faults is $1/\mu = 300a$, the mean upper limit time is $1/\theta = 300a$, and the inspection time after N failures of retransmissions is $v = 60a$. For example, when $a = 1$ second, $1/\lambda = 6$ hours, $1/\mu = 1/\theta = 5$ minutes and $v = 1$ minute.

Table 11.1. Optimal retransmission numbers N^* when $1/\lambda = 21,600a$, $1/\mu = 300a$, $1/\theta = 300a$ and $v = 60a$

q_1	$q_0 = 0.0$			$q_0 = 0.1$		
	q_2			q_2		
	0.99	0.999	1.0	0.99	0.999	1.0
0.1	5	5	5	10	9	9
0.2	7	7	7	10	9	9
0.3	9	8	8	10	9	9
0.4	12	10	10	12	11	11
0.5	14	13	13	15	13	13

Table 11.1 gives the optimal numbers N^* for q_i $(i = 0, 1, 2)$ which satisfy (11.17). This shows that N^* increase with q_1. It is indicated that we should adopt an ARQ strategy when hidden faults have occurred. However, from [55], N^* depend little on $1/\lambda$ and $1/\mu$. That is, N^* are roughly determined by q_i $(i = 0, 1, 2)$. For example, when $q_1 = 0.1 \sim 0.3$, N^* are $5 \sim 10$ and take almost the same values as actual numbers in practical fields.

We have defined the three states of no fault, hidden fault, and permanent failure according to the occurrences of intermittent faults, and have discussed the optimal ARQ strategy which minimizes the mean time to the success of data transmission. The methods used in this section and the results derived here could be applied to the analysis of more advanced hybrid ARQ schemes.

11.3 SR ARQ Model with Retransmission Number

In this section, we formulate a stochastic model of an ARQ scheme with SR protocol: we transmit blocks of data continuously and retransmit only the blocks in which errors are detected. Hence, the SR protocol is the most efficient in the ARQ scheme. However, the implementation of the SR protocol requires an extensive buffer (theoretically, an infinite buffer) at the receiver side.

We consider the following model of SR ARQ scheme [57]: we transmit a certain amount of data, which is named a *block*, successively from a sender to a receiver. If a receiver detects errors in a certain block, it returns a negative acknowledgement (NAK) to a sender. Conversely, if the block is transmitted

successfully, a receiver sends a positive acknowledgement (ACK) to a sender. When a sender receives a NAK for a certain block transmission, the same block is retransmitted to a receiver. If a sender receives the successive $N+1$ NAK's, we stop the data transmission for some time to inspect and maintain a communication system. After that, the system returns to a normal condition.

For the above model, we obtain the expected numbers of ACK's and retransmitted blocks and the mean time until a receiver sends the successive $N+1$ NAK's for the first time. Using these results, optimal numbers N^* which minimize the mean time per successful blocks and the mean loss time per the mean time are derived.

11.3.1 Model and analysis

We consider the following SR ARQ strategy which detects and corrects errors in a communication system:

(1) We transmit n blocks successively to a receiver.
(2) A receiver checks every block whether errors have occurred or not. If no errors have been detected, a receiver returns ACK's for these successful blocks successively and accepts them. We call this *success of data transmission.* Conversely, if some errors have been detected, a receiver returns NAK's for these blocks to a sender. We call this *failure of data transmission.*
(3) When a sender has received NAK, we retransmit the erroneous block until the success of data transmission. If a sender has received the successive $N+1$ NAK's, we stop the data transmission for some time v to transmit its block to a receiver again.
(4) A mean time a is required to prepare one block for sending and b is required for all times resulting from transmission, which includes transmitting, checking and returning ACK or NAK, where $(n-1)a \leq b$. That is, the mean time from preparation of n blocks for sending to receipt of ACK or NAK for the first block is $a+b$.

In the above ARQ strategy, we introduce the following probabilities of errors :

(5) Errors of each transmitted block occur with probability q_0 $(\equiv 1-p_0)$, and do not occur with $1-q_0$ in a normal condition. When the $(j-1)$-th consecutive errors have occurred, errors in the next transmitted block occur with probability q_j $(\equiv 1-p_j)$ $(j=1,2,\cdots,N)$, where it is assumed that $q_0 < q_1 < \cdots < q_N$. This assumption indicates that if data transmissions fail repeatedly at many times there might be faults in a communication line and a transmission environment then becomes worse with the number of transmissions.

(6) When no errors have occurred in a block, errors occur the next transmitted block with probability q_0. That is, if the data transmission succeeds after some NAK's, it returns to a normal condition and the probability of errors becomes q_0.

We investigate reliability properties of the above ARQ strategy until the successive $N+1$ NAK's. The mean time to the successive $N+1$ NAK's, after we have received ACK or NAK, is given by a renewal function

$$h(N) = \sum_{j=1}^{\infty} p_0^{j-1} \sum_{k=1}^{N} [F(k-1) - F(k)] [(j-1)a + (k+1)a + h(N)] + \sum_{j=1}^{\infty} p_0^{j-1} F(N) [(j-1)a + Na], \tag{11.19}$$

where $F(k) \equiv q_0 q_1 \cdots q_k$ $(k = 0, 1, 2, \cdots)$, which represents the probability on the successive $k+1$ NAK's. The first term on the right-hand side in (11.19) is the mean time after $j-1$ $(j = 1, 2, \cdots)$ ACK's until we receive the successive $N+1$ NAK's. The second term is the mean time after $j-1$ $(j = 1, 2, \cdots)$ ACK's and the successive k $(k = 1, 2, \cdots, N)$ NAK's until we receive one ACK and hence, the data transmission returns to a normal condition. Solving (11.19) for $h(N)$, we have

$$\begin{aligned} h(N) &= \frac{\left[\sum_{j=1}^{\infty} p_0^{j-1} F(N) [(j-1)a + Na] + \sum_{j=1}^{\infty} p_0^{j-1} \sum_{k=1}^{N} [F(k-1) - F(k)] [(j-1)a + (k+1)a] \right]}{1 - \sum_{j=1}^{\infty} p_0^{j-1} \sum_{k=1}^{N} [F(k-1) - F(k)]} \\ &= \frac{a}{F(N)} \left[1 + \sum_{k=0}^{N-1} F(k) \right] - a. \end{aligned} \tag{11.20}$$

Thus, the mean duration of the data transmission from preparation of sending n blocks to the receipt of the successive $N+1$ NAK's is

$$\begin{aligned} H(N) &= a + b + h(N) \\ &= b + \frac{a}{F(N)} \left[1 + \sum_{k=0}^{N-1} F(k) \right], \qquad N = 1, 2, \cdots. \end{aligned} \tag{11.21}$$

Similarly, the expected number $M_1(N)$ of ACK blocks and the expected number $M_2(N)$ of retransmitted blocks until the successive $N+1$ NAK's are, respectively,

$$M_1(N) = \frac{\left[\begin{array}{c} \displaystyle\sum_{j=1}^{\infty} p_0^{j-1} F(N) \\ \displaystyle + \sum_{j=1}^{\infty} p_0^{j-1} \sum_{k=1}^{N} [F(k-1) - F(k)]\,[(j-1)+1] \end{array}\right]}{1 - \displaystyle\sum_{j=1}^{\infty} p_0^{j-1} \sum_{k=1}^{N} [F(k-1) - F(k)]}$$

$$= \frac{1 - F(N)}{F(N)}, \qquad N = 1, 2, \cdots, \tag{11.22}$$

$$M_2(N) = \frac{\left[\begin{array}{c} \displaystyle\sum_{j=1}^{\infty} p_0^{j-1} N \cdot F(N) \\ \displaystyle + \sum_{j=1}^{\infty} p_0^{j-1} \sum_{k=1}^{N} k\,[F(k-1) - F(k)] \end{array}\right]}{1 - \displaystyle\sum_{j=1}^{\infty} p_0^{j-1} \sum_{k=1}^{N} [F(k-1) - F(k)]}$$

$$= \frac{1}{F(N)} \sum_{k=0}^{N-1} F(k), \qquad N = 1, 2, \cdots. \tag{11.23}$$

Evidently, we have the relation

$$h(N) = a\,[M_1(N) + M_2(N)]. \tag{11.24}$$

11.3.2 Optimal policy

From (11.21) and (11.22), the mean time to $N+1$ NAK's per the successful transmitted blocks is given by

$$L_1(N_1) \equiv \frac{H(N_1)}{M_1(N_1)}$$

$$= \frac{(a+b) + a \displaystyle\sum_{k=0}^{N_1-1} F(k)}{1 - F(N_1)} - b, \qquad N_1 = 1, 2, \cdots. \tag{11.25}$$

We seek an optimal number N_1^* which minimizes $L_1(N_1)$. From the inequality $L_1(N_1+1) - L_1(N_1) \geq 0$,

$$\frac{1-F(N_1)}{1-q_{N_1+1}} - \sum_{k=0}^{N_1-1} F(k) \geq \frac{a+b}{a}, \qquad N_1 = 1, 2, \cdots . \tag{11.26}$$

Let denote the left side of (11.26) by $Q_1(N_1)$. Then, we easily have

$$Q_1(1) = \frac{p_0 + q_0(q_2 - q_1)}{1 - q_2}, \tag{11.27}$$

$$Q_1(N_1) - Q_1(N_1 - 1) = \frac{1 - F(N_1 - 1)}{(1 - q_{N_1})(1 - q_{N_1+1})} (q_{N_1+1} - q_{N_1}) > 0. \tag{11.28}$$

Thus, $Q_1(N_1)$ strictly increases from $Q_1(1)$. Further, we have

$$Q_1(N_1) \geq \frac{1}{1 - q_{N_1+1}} \sum_{k=0}^{N_1-1} F(k)(q_{N_1+1} - q_{k+1}). \tag{11.29}$$

Hence, if $\lim_{j\to\infty} q_j = 1$, then a finite solution to (11.26) exists.

Therefore, we have the following optimal policy :

(i) If $\lim_{j\to\infty} q_j = 1$ and $Q_1(1) < (a+b)/a$, then there exists a unique minimum number N_1^* $(2 \leq N_1^* < \infty)$ which satisfies (11.26), and the resulting mean time is

$$\frac{a}{1 - q_{N_1^*}} < L_1(N_1^*) + b \leq \frac{a}{1 - q_{N_1^*+1}}.$$

(ii) If $Q_1(1) \geq (a+b)/a$ then $N_1^* = 1$, that is, when the first data transmission has failed, a communication system is inspected immediately.

Next, from (11.21) and (11.23), the mean loss time due to retransmissions and inspections per the mean recurrence time to an initial state is

$$\begin{aligned} L_2(N_2) &\equiv \frac{aM_2(N_2) + v}{H(N_2) + v} \\ &= \frac{vF(N_2) + a \sum_{k=0}^{N_2-1} F(k)}{(n+v)F(N_2) + a\left[1 + \sum_{k=0}^{N_2-1} F(k)\right]}, \qquad N_2 = 1, 2, \cdots . \end{aligned} \tag{11.30}$$

We seek an optimal number N_2^* which minimizes $L_2(N_2)$. Forming the inequality $L_2(N_2+1) - L_2(N_2) \geq 0$,

$$\frac{a/b + F(N_2)}{1 - q_{N_2+1}} + \sum_{k=0}^{N_2-1} F(k) \geq \frac{v}{b}, \qquad N_2 = 1, 2, \cdots . \tag{11.31}$$

Let us denote the left side of (11.31) as $Q_2(N_2)$. Then, we have

$$Q_2(1) = \frac{a/b + q_0(1 - q_2 + q_1)}{1 - q_2}, \tag{11.32}$$

$$Q_2(N_2) - Q_2(N_2 - 1) = \frac{a/b + F(N_2)}{(1 - q_{N_2})(1 - q_{N_2+1})} (q_{N_2+1} - q_{N_2}) > 0. \tag{11.33}$$

Thus, $Q_2(N_2)$ strictly increases from $Q_2(1)$.

Therefore, we have the following optimal policy :

(iii) If $\lim_{j\to\infty} q_j = 1$ and $Q_2(1) < v/b$ then there exists a unique minimum number N_2^* $(2 \leq N_2 < \infty)$ which satisfies (11.31).
(iv) If $Q_2(1) \geq v/b$ then $N_2^* = 1$.

11.3.3 Numerical examples and remarks

We compute numerically optimal numbers N_1^* and N_2^* which minimize $L_1(N_1)$ in (11.25) and $L_2(N_2)$ in (11.30), respectively. Suppose that the transmission time is $b = 5a \sim 40a$ and the mean inspection time after $N+1$ NAK's is $v = 2a$, $5a$,$10a$. Further, the probability that errors of the j-th transmitted block occur, after the $(j-1)$-th consecutive error is $q_j = 1 - p^{j+1}$ $(j = 0, 1, 2, \cdots)$.

Table 11.2. Optimal numbers N_1^* to minimize $L_1(N_1)$ when $q_j = 1 - p^{j+1}$

	b/a							
p	5	10	15	20	25	30	35	40
0.92	20	27	32	35	38	40	41	43
0.90	16	21	25	27	29	31	33	34
0.84	9	12	14	16	17	18	19	20
0.80	7	9	11	12	13	14	15	15
0.70	4	5	6	7	8	8	9	9
0.60	3	4	4	5	5	6	6	6
0.50	2	3	3	3	4	4	4	4

Table 11.2 gives optimal numbers N_1^* which satisfy (11.26). It is shown from Table 11.2 that N_1^* increase with p and b/a. For example, when the length of block is $8,192$ bits and a bit error rate is 3.7×10^{-5}, *i.e.*, p is 0.70,

N_1^* are $4 \sim 9$ and take almost the same values as ARQ numbers in practical fields. However, when p is larger, N_1^* are also larger. This indicates that when the error rates in the data transmission system become smaller, we should continue to transmit blocks whenever possible.

Table 11.3. Optimal numbers N_2^* to minimize $L_2(N_2)$ when $q_j = 1 - p^{j+1}$

p	$v/b = 2$				$v/b = 5$				$v/b = 10$			
	b/a				b/a				b/a			
	5	10	20	30	5	10	20	30	5	10	20	30
0.92	26	34	42	47	37	45	53	58	45	54	62	67
0.90	20	26	33	37	29	35	42	46	36	42	49	53
0.84	11	15	19	21	17	21	25	27	21	25	29	31
0.80	8	11	14	15	12	15	18	20	16	19	22	23
0.70	3	4	6	6	6	8	9	10	9	10	11	12
0.60	1	1	1	1	1	2	2	2	3	3	3	3

Table 11.3 gives optimal numbers N_2^* which satisfy (11.31). From these values, we could also estimate the allowed inspection time v after $N + 1$ NAK's.

11.4 Hybrid ARQ Models with Response Time

In this section, we propose two stochastic models of hybrid ARQ schemes which combine the advantages of both FEC and SW ARQ schemes [58], [59]: when the first transmission of hybrid ARQ scheme has failed, we repeatedly transmit the same data together with FEC until success of data transmission. This is called type-I hybrid ARQ. However, the data throughput is decreased even in a good environment by bit increments of correcting codes. To improve the throughput, we would rather transmit the data with only error-correcting codes when the transmission has failed. This is called type-II hybrid ARQ.

We derive the mean times to success of data transmission for both hybrid ARQ schemes using Markov renewal processes, and compare them. Further, we discuss analytically and numerically which scheme is better.

Suppose that a certain amount of data, which is named *unit data*, is transmitted successively from a sender to a receiver. Then, we call a unit data with error-detecting code a *data block* and one with both error-detecting and error-correcting codes, a *coded data block*. Furthermore, we call only the error-correcting code with error-detecting code the *correcting code block*.

It is assumed that errors of each block transmission occur with certain probabilities as follows: the probabilities of errors in data block and correcting code block are the same q_0, and the probability of errors in a coded data block is q_1, where $0 \leq q_0 \leq q_1 < 1$. This is because while both data block and correcting code block are almost the same size, the coded data block is bigger

than the other blocks since it consists of redundant bits of error-detecting and error-correcting codes. It is also assumed that when errors of each block have occurred, they are detected.

Similarly, both mean times required for each transmission of data block and correcting code block are the same value a and the mean time for coded data block is $a+b$, where $0 < b < a$. That is, the mean time b represents an additional transmission time for the error-correcting code.

When the data transmission of blocks has failed, we repeatedly retransmit the same blocks. If the data transmission has not succeeded by time T, we stop it, and inspect and maintain the transmission system for some time v, where T is called the *deadline time* [60], [61].

11.4.1 Type-I hybrid ARQ

The coded data block is transmitted from a sender to a receiver. When the errors have been detected they can be corrected by themselves with probability α $(0 < \alpha < 1)$. If the error correction fails, we retransmit the same block at the request of a receiver. Note that $q_1(1-\alpha)$ is the probability that the errors of coded data block are detected, but are not corrected.

We define the following states of the data transmission:

State 0 : The coded data block begins to be transmitted.
State S : The data transmission succeeds.
State F : The data transmission has failed before time T.

The above system states form a Markov renewal process [62], where state S is an absorbing state and both states 0 and F are regenerating points.

Let $A_1(t)$ be a degenerate distribution placing unit mass at $a+b$, *i.e.*, $A_1(t) \equiv 1$ for $t \geq a+b$, 0 for $t < a+b$, and $Q_{ij}(t)$ $(i = 0, F; j = 0, S, F)$ be mass functions of a Markov renewal process. Then, the Laplace-Stieltjes (LS) transforms of $Q_{ij}(t)$ are

$$q_{0S}(s) = \sum_{j=1}^{\infty} [q_1(1-\alpha)]^{j-1}[1 - q_1(1-\alpha)] \int_0^T e^{-st} dA_1^{(j)}(t), \tag{11.34}$$

$$q_{0F}(s) = \sum_{j=1}^{\infty} [q_1(1-\alpha)]^{j-1} e^{-sT}[A_1^{(j-1)}(T) - A_1^{(j)}(T)], \tag{11.35}$$

$$q_{F0}(s) = e^{-sv}, \tag{11.36}$$

where $A^{(j)}(t)$ denotes the j-fold Stieltjes convolution of a distribution $A(t)$ with itself, *i.e.*, $A^{(j)}(t) \equiv \int_0^t A^{(j-1)}(t-u)dA(u)$ $(j = 1, 2, \cdots)$, and $A^{(0)}(t) \equiv 1$ for $t \geq 0$.

We obtain the mean time to success of transmission of the coded data block. Let $H_I(t)$ be the first-passage time distribution from state 0 to state S. Then, we have the renewal equation

$$H_I(t) = Q_{0S}(t) + Q_{0F}(t) * Q_{F0}(t) * H_I(t), \tag{11.37}$$

where the asterisk denotes the Stieltjes convolution. Taking the LS transforms on both sides of (11.37) and arranging them, we have

$$h_I(s) = \frac{q_{0S}(s)}{1 - q_{0F}(s)q_{F0}(s)}. \tag{11.38}$$

Thus, the mean time to the success of data transmission is given by

$$\begin{aligned} \ell_I(T) &\equiv \lim_{s\to 0}\left[-\frac{dh_I(s)}{ds}\right] \\ &= \frac{a+b}{1-q_1(1-\alpha)} + \frac{[T+v-N_1(a+b)][q_1(1-\alpha)]^{N_1}}{1-[q_1(1-\alpha)]^{N_1}}, \end{aligned} \tag{11.39}$$

where $N_1 \equiv [T/(a+b)]$ for $T \geq a+b$ which denotes the greatest integer contained in $[\,\cdot\,]$.

11.4.2 Type-II hybrid ARQ

The data block is transmitted from a sender to a receiver. When the errors have been detected, only the correcting code block is transmitted from a sender at the request of a receiver. If the error correction has failed, the same data block is retransmitted.

(1) The data block D_1 is transmitted :

(a) If no errors have been detected in D_1, D_1 is accepted. We call the transmission *success of data transmission.*

(b) If errors have been detected, a receiver keeps D_1 and requests transmission of the correcting code block C_1 only.

(2) The correcting code block C_1 is transmitted :

(a) If no errors have been detected in C_1, a receiver can correct errors of D_1 with probability α_1 $(0 < \alpha_1 < 1)$, and D_1 is accepted. We call the transmission *success of data transmission.* Otherwise, a receiver keeps C_1 and requests transmission of the data block D_2.

(b) If errors have been detected, the transmission of D_2 is required.

(3) The data block D_j $(j = 2, 3, \cdots)$ is retransmitted :

(a) If no errors have been detected in D_j, D_j is accepted (success of data transmission).

(b) If errors have been detected, they can be corrected by C_{j-1}, which has been kept, with probability α_2 $(0 < \alpha_2 < 1)$ (success of data transmission). Otherwise, a receiver keeps D_j and requests transmission of the correcting code block C_j only.

(4) The correcting code block C_j $(j = 2, 3, \cdots)$ is retransmitted :

(a) If no errors have been detected in C_j, C_j can correct D_j with probability α_1 (success of data transmission). Otherwise, a receiver keeps C_j and requests transmission of D_{j+1}.

(b) If the errors have been detected, the transmission of D_{j+1} is required.

(5) Steps (3) and (4) are repeated until the success of data transmission.

(6) If the data transmission has not succeeded by time T, it is stopped and is restarted from the beginning of (1) after time v.

Let $x \equiv \alpha_1(1-q_0)$, $X \equiv q_0(1-\alpha_2)(1-x)$. Then, it is easily seen that x is the probability that the correcting code block C_j $(j = 1, 2, \cdots)$ can correct errors in data block D_j, which has been kept, and X is the probability that C_j fails to correct errors of D_j and also cannot correct errors of D_{j+1} when they are detected.

We define the following system states:

State 0: The data block D_1 begins to be transmitted.

States S and F: The same as the states of Section 11.4.1.

To derive mass functions $Q_{ij}(t)$, we use the following notations:

$$
\begin{aligned}
&D_S(1) \equiv 1 - q_0, \\
&D_S(j) \equiv q_0 X^{j-2}(1-x-X), \qquad j = 2, 3, \cdots, \\
&D_F(j) \equiv q_0 X^{j-1}, \qquad j = 1, 2, \cdots, \\
&C_S(j) \equiv q_0 x X^{j-1}, \qquad j = 1, 2, \cdots, \\
&C_F(0) \equiv 1, \\
&C_F(j) \equiv q_0(1-x)X^{j-1}, \qquad j = 1, 2, \cdots.
\end{aligned}
$$

Note that $D_S(j)$ and $D_F(j)$ $(j = 2, 3, \cdots)$ represent the probabilities that the transmissions of data block have failed successively at $(j-1)$ times and the transmission of D_j succeeds or fails, respectively, and $C_S(j)$ and $C_F(j)$ $(j = 2, 3, \cdots)$ are the probabilities that the transmissions of correcting code block have failed successively at $(j-1)$ times and the transmission of C_j succeeds or fails, respectively. It is easily seen that

$$
\begin{aligned}
&D_S(j) + D_F(j) = C_F(j-1), \qquad j = 1, 2, \cdots, \\
&C_S(j) + C_F(j) = D_F(j), \qquad j = 1, 2, \cdots.
\end{aligned}
$$

Let $A_0(t) \equiv 1$ for $t \geq a$, 0 for $t < a$. Then, the LS transforms of $Q_{ij}(t)$ are given by

$$q_{0S}(s) = \sum_{j=1}^{\infty} D_S(j) \int_0^T e^{-st} dA_0^{(2j-1)}(t) + \sum_{j=1}^{\infty} C_S(j) \int_0^T e^{-st} dA_0^{(2j)}(t), \tag{11.40}$$

$$q_{0F}(s) = \sum_{j=0}^{\infty} C_F(j) e^{-sT} [A_0^{(2j)}(T) - A_0^{(2j+1)}(T)] + \sum_{j=1}^{\infty} D_F(j) e^{-sT} [A_0^{(2j-1)}(T) - A_0^{(2j)}(T)]. \tag{11.41}$$

It is evident that $q_{0S}(0) + q_{0F}(0) = 1$. The first term on the right-hand side in (11.40) is the LS transform of the probability that errors of D_{j-1} cannot be corrected by C_{j-1} and the transmission of D_j succeeds. The second term is the LS transform of the probability that the transmission of D_j has failed and C_j can correct errors of D_j.

We obtain the mean time $\ell_{II}(T)$ to success of data block transmission. Substituting (11.36), (11.40) and (11.41) into (11.38), and arranging them, we have, for $2ja \le T < (2j+1)a$ $(j = 1, 2, \cdots)$,

$$h_{II}(s) = \frac{\sum_{j=1}^{N_2} [D_S(j) e^{-(2j-1)sa} + C_S(j) e^{-2jsa}]}{1 - C_F(N_2) e^{-s(T+v)}}, \tag{11.42}$$

and for $(2j+1)a \le T < 2(j+1)a$ $(j = 1, 2, \cdots)$,

$$h_{II}(s) = \frac{\sum_{j=1}^{N_2+1} D_S(j) e^{-(2j-1)sa} + \sum_{j=1}^{N_2} C_S(j) e^{-2jsa}}{1 - D_F(N_2+1) e^{-s(T+v)}}, \tag{11.43}$$

where $N_2 \equiv [T/(2a)]$ for $T \ge 2a$. Thus, the mean time to success of data block transmission is, for $2ja \le T < (2j+1)a$ $(j = 1, 2, \cdots)$,

$$\ell_{II}(T) = \frac{a + v + T - 2aN_2 + aq_0 \left[\frac{(2-x)(1-X^{N_2-1})}{1-X} + X^{N_2-1} \right]}{1 - q_0(1-x)X^{N_2-1}} + 2aN_2 - T - v, \tag{11.44}$$

and for $(2j+1)a \le T < 2(j+1)a$ $(j = 1, 2, \cdots)$,

$$\ell_{II}(T) = \frac{a+v+T-(2N_2+1)a+\frac{aq_0(2-x)(1-X^{N_2})}{1-X}}{1-q_0X^{N_2}} + (2N_2+1)a - T - v. \tag{11.45}$$

11.4.3 Comparison of type-I and type-II hybrid ARQs

We compare the mean times of type-I and type-II hybrid ARQ's, taking note of the probability q_0 of errors of data block, which is the most fundamental probability between two models.

First, we compare three probabilities α, α_1 and α_2 of correcting errors. It would be reasonable to assume that $\alpha_2 \leq \alpha_1$ because the correcting code block by which errors were detected has not been kept and hence, the errors in D_j might not be amenable to correction by C_{j-1}. The transmission of a coded data block for type-I policy often fails since the decoding at the receiver side may be made imperfectly owing to complicated correcting codes, and hence, we would have $\alpha \leq \alpha_1$. We cannot explain clear relations between α and α_2; however, there exists an inequality $\alpha \geq \alpha_2$ in many practical cases.

From (11.39) and (11.44), we have

$$\begin{aligned} L_2(q_0) &\equiv \ell_I(T) - \ell_{II}(T) \\ &= \frac{a+b}{1-q_1(1-\alpha)} - \frac{1}{1-q_0(1-x)X^{N_2-1}} \\ &\quad \times \left(\frac{\left\{ \begin{array}{c} [T+v-2aN_2]q_0(1-x)X^{N_2-1} \\ -[T+v-N_1(a+b)][q_1(1-\alpha)]^{N_1} \\ +[2aN_2-N_1(a+b)][q_1(1-\alpha)]^{N_1}q_0(1-x)X^{N_2-1} \end{array} \right\}}{1-[q_1(1-\alpha)]^{N_1}} \right. \\ &\quad \left. + a + aq_0X^{N_2-1} + \frac{aq_0(2-x)(1-X^{N_2-1})}{1-X} \right). \end{aligned} \tag{11.46}$$

We easily have

$$L_2(0) = \frac{aq_1(1-\alpha)+b}{1-q_1(1-\alpha)} + \frac{[T+v-N_1(a+b)][q_1(1-\alpha)]^{N_1}}{1-[q_1(1-\alpha)]^{N_1}} > 0, \tag{11.47}$$

$$\begin{aligned} L_2(1) &= \frac{a+b}{\alpha} - \frac{2a}{\alpha_2} - \frac{1}{1-(1-\alpha_2)^{N_2-1}} \left(a + a(1-\alpha_2)^{N_2-1} \right. \\ &\quad \left. + \frac{\left\{ \begin{array}{c} [T+v-2aN_2](1-\alpha_2)^{N_2-1} \\ -[T+v-N_1(a+b)](1-\alpha)^{N_1} \\ +[2aN_2-N_1(a+b)](1-\alpha)^{N_1}(1-\alpha_2)^{N_2-1} \end{array} \right\}}{1-[(1-\alpha)]^{N_1}} \right), \end{aligned} \tag{11.48}$$

$$L_2'(q_0) = \frac{-N_2(1-x+q_0\alpha_1)X^{N_2-1}}{[1-q_0(1-x)X^{N_2-1}]^2}\left[\frac{aq_0\alpha_2(2-x)}{1-X}+T+v-2aN_2\right]$$
$$+\frac{-a[2-x+q_0\alpha_1+q_0^2\alpha_1(1-\alpha_2)]}{1-q_0(1-x)X^{N_2-1}}\cdot\frac{1-X^{N_2}}{(1-X)^2}<0. \qquad (11.49)$$

Hence, $L_2(q_0)$ is strictly decreasing in q_0 from $L_2(0)$ to $L_2(1)$. Thus, if $L_2(1) < 0$ then there exists a unique solution $\hat{q}_0$ $(0 < \hat{q}_0 < 1)$ which satisfies $L_2(q_0) = 0$. It is easily seen that if $\alpha \geq \alpha_2$ then $L_2(1) < 0$. We have the same results in the case where $\ell_{II}(T)$ is given in (11.45).

Therefore, we have the following results, where $\hat{q}_0 \equiv 1$ for $L_2(1) \geq 0$:

(i) If $0 \leq q_0 < \hat{q}_0$ then $\ell_I(T) > \ell_{II}(T)$, *i.e.*, type-II ARQ is better than type-I ARQ.
(ii) If $\hat{q}_0 < q_0 < 1$ then $\ell_I(T) < \ell_{II}(T)$, *i.e.*, type-I ARQ is better than type-II ARQ.

11.4.4 Numerical examples and remarks

Table 11.4. Comparison between the mean times until the success of data transmission for type-I and type-II hybrid ARQ's

T	q_0	$\hat{q}_0$	$\ell_I(T)$	$\ell_{II}(T)$
	0.1	0.177	2.947	* 2.283
	0.2		* 4.462	4.831
2a	0.3		* 6.052	8.954
	0.4		* 7.722	15.245
	0.5		* 9.477	24.828
	0.1		1.536	* 1.124
	0.2		1.577	* 1.295
4a	0.3	0.323	1.629	* 1.561
	0.4		* 1.696	1.997
	0.5		* 1.781	2.702
	0.1		1.535	* 1.119
	0.2		1.572	* 1.259
8a	0.3		1.610	* 1.420
	0.4	0.428	1.651	* 1.606
	0.5		* 1.694	1.820
	0.1		1.535	* 1.119
	0.2		1.572	* 1.259
∞	0.3		1.610	* 1.420
	0.4	0.428	1.651	* 1.605
	0.5		* 1.693	1.818

* : minimum values

We compute and compare numerically the mean times until the transmission of a data block succeeds. Suppose that the probability of errors of coded data block is $q_1/q_0 = 1.2$, which is relative to probability q_0 of errors in a data

block. That is, the rate of increment bits of error-correcting code is about 20%. The total additional times increased by increment bits and required for coding and decoding blocks are $b/a = 0.5$, which is relative to the time a for block transmission. Further, the time for restart after inspections and maintenances of the system is $v/a = 60$. Moreover, the probabilities α and α_2 of error corrections are $\alpha/\alpha_1 = 0.9$ and $\alpha_2/\alpha_1 = 0.9$, which are relative to the probability α_1 of error correction of the conecting code block.

Table 11.4 gives the mean times $\ell_I(T)$, $\ell_{II}(T)$ and $\hat{q}_0$ for $q_0 = 0.1 \sim 0.5$ when $\alpha_1 = 0.9$ and $T = 2a, 4a, 8a, \infty$. It can be seen that both mean times increase with q_0 and decrease with the response time T. Type-I ARQ becomes better than type-II ARQ as q_0 is increasing, and conversely, type-II ARQ is better than type-I ARQ when q_0 is relatively small. This indicates that type-II ARQ would be more effective than type-I ARQ in a normal environment of a conventional communication system.

It is easily noted that the mean times $\ell_I(T)$ and $\ell_{II}(T)$ and the probabilities $\hat{q}_0$ are almost unchanged when $T = 8a$ and $T = \infty$. Since the upper bound of retransmission numbers is nearly 8 in practice, $\hat{q}_0 = 0.428$ gives the probability of a turning point to show which scheme is better.

We have formulated two stochastic models of hybrid ARQ's with response time T, and have derived the mean times for successful data transmission. Comparing them, we have analyzed and discussed which scheme is better and have evaluated these results in numerical examples. Type-I hybrid ARQ is better than type-II one in a worse environment, while on the other hand, type-II ARQ is better in a normal environment. That is, to improve data transmission, we should retransmit only the error-correcting code in normal conditions.

Further studies to develop more useful techniques of error-correcting coding and more practical error-control technologies of hybrid ARQ can be expected in these research areas.

References

1. Schwartz, M. (1987), Telecommunication Networks : Protocols, Modeling and Analysis. Addison Wesley, Massachusetts
2. I.E.I.C.E. (ed.) (1988), Handbook for Electronics, Information and Communication Engineers. Ohm Publication, Tokyo
3. Chang, J. F. and Yang, T. H. (1993), "Multichannel ARQ protocols," *IEEE Transactions on Communications*, **41**, 592-598
4. Lu, D. L. and Chang, J. F. (1993), "Performance of ARQ protocols in nonindependent channel errors," *IEEE Transactions on Communications*, **41**, 721-730
5. Bruneel, H. and Moeneclaey, M. (1986), "On the throughput performance of some continuous ARQ strategies with repeated transmission," *IEEE Transactions on Communications*, **COM-34**, 244-249
6. Fantacci, R. (1990), "Performance evaluation of efficient continuous ARQ protocols," *IEEE Transactions on Communications*, **38**, 773-781
7. Fantacci, R. (1991), "Performance evaluation of some ARQ schemes using efficient modulation techniques and noncoherent detection," *IEEE Transactions on Communications*, **39**, 445-451
8. Kim, S. R. and Un, C. K. (1992), "Throughput analysis for two ARQ schemes using combined transition matrix," *IEEE Transactions on Communications*, **40**, 1679-1683
9. Hayashida, Y. (1993), "Throughput analysis of tandem-type go-back-N ARQ scheme for satellite communications," *IEEE Transactions on Communications*, **41**, 1517-1524
10. Cho, Y. J. and Un, C. K. (1994), "Performance analysis of ARQ error controls under Markovian block error pattern," *IEEE Transactions on Communications*, **42**, 2051-2061
11. Yao, Y. D. (1995), "An effective go-back-N ARQ scheme for variable-error-rate channels," *IEEE Transactions on Communications*, **43**, 20-23
12. Cam, R. and Leung, C. (1997), "Throughput analysis of some ARQ protocols in the presence of feedback errors," *IEEE Transactions on Communications*, **45**, 35-44
13. Miller, M. J. and Lin, S. (1981), "The analysis of some selective-repeat ARQ schemes with finite receiver buffer," *IEEE Transactions on Communications*, **COM-29**, 1307-1315
14. Wang, Y. M. and Lin, S. (1983), "A modified selective-repeat type-II hybrid ARQ system and its performance analysis," *IEEE Transactions on Communications*, **COM-31**, 593-607
15. Shacham, N. and Towsley, D. (1991), "Resequencing delay and buffer occupancy in selective repeat ARQ with multiple receivers," *IEEE Transactions on Communications*, **39**, 928-937

16. Shacham, N. and Shin, B. C. (1992), "A selective-repeat-ARQ protocol for parallel channels and its resequencing analysis," *IEEE Transactions on Communications*, **40**, 773-782
17. Benelli, G. (1993), "Some ARQ protocols with finite receiver buffer," *IEEE Transactions on Communications*, **41**, 513-523
18. Benelli, G. (1993), "A selective ARQ protocol with a finite-length buffer," *IEEE Transactions on Communications*, **41**, 1102-1111
19. Chang, J. F. and Yang, T. H. (1994), "End-to-end delay of an adaptive selective repeat ARQ protocol," *IEEE Transactions on Communications*, **42**, 2926-2928
20. Lin, S. and Yu, P. S. (1982), "A hybrid ARQ scheme with parity retransmission for error control of satellite channels," *IEEE Transactions on Communications*, **COM-30**, 1701-1719
21. Benelli, G. (1985), "An ARQ scheme with memory and soft error detectors," *IEEE Transactions on Communications*, **COM-33**, 285-288
22. Kasami, T., Fujiwara, T. and Lin, S. (1986), "A concatenated coding scheme for error control," *IEEE Transactions on Communications*, **COM-34**, 481-488
23. Krishna, H. and Morgera, S. D. (1987), "A new error control scheme for hybrid ARQ systems," *IEEE Transactions on Communications*, **COM-35**, 981-989
24. Kallel, S. (1990), "Analysis of a type II hybrid ARQ scheme with code combining," *IEEE Transactions on Communications*, **38**, 1133-1137
25. Kallel, S. and Haccoun, D. (1990), "Generalized type II hybrid ARQ scheme using punctured convolutional coding," *IEEE Transactions on Communications*, **38**, 1938-1946
26. Kallel, S. and Haccoun, D. (1991), "Sequential decoding with an efficient partial retransmission ARQ strategy," *IEEE Transactions on Communications*, **39**, 208-213
27. Wicker, S. B. (1991), "Adaptive rate error control through the use of diversity combining and majority-logic decoding in a hybrid-ARQ protocol," *IEEE Transactions on Communications*, **39**, 380-385
28. Kousa, M. A. and Rahman, M. (1991), "An adaptive error control system using hybrid ARQ schemes," *IEEE Transactions on Communications*, **39**, 1049-1057
29. Lin, M. C. and Guu, M. Y. (1991), "The performance analysis of a concatenated ARQ scheme using parity retransmissions," *IEEE Transactions on Communications*, **39**, 1869-1874
30. Rice, M. D. and Wicker, S. B. (1992), "Modified majority logic decoding of cyclic codes in hybrid-ARQ systems," *IEEE Transactions on Communications*, **40**, 1413-1417
31. Kallel, S. (1992), "Sequential decoding with an efficient incrémental redundancy ARQ scheme," *IEEE Transactions on Communications*, **40**, 1588-1594
32. Benelli, G. (1992), "A new method for the integration of modulation and channel coding in an ARQ protocol," *IEEE Transactions on Communications*, **40**, 1594-1606
33. Benelli, G. (1993), "New ARQ protocols using concatenated codes," *IEEE Transactions on Communications*, **41**, 1013-1019
34. Kallel, S. (1994), "Efficient hybrid ARQ protocols with adaptive forward error correction," *IEEE Transactions on Communications*, **42**, 281-289
35. Wicker, S. B. and Bartz, M. J. (1994), "Type-II hybrid-ARQ protocols using punctured MDS codes," *IEEE Transactions on Communications*, **42**, 1431-1440
36. Deng, R. H. (1994), "Hybrid ARQ schemes employing coded modulation and sequence combining," *IEEE Transactions on Communications*, **42**, 2239-2245
37. Deng, R. H. and Lin, M. L. (1995), "A type I hybrid ARQ system with adaptive code rates," *IEEE Transactions on Communications*, **43**, 733-737

38. Rasmussen, L. K. and Wicker, S. B. (1995), "Trellis-coded, type-I hybrid-ARQ protocols based on CRC error-detecting codes," *IEEE Transactions on Communications*, **43**, 2569-2575
39. Towsley, D. (1985), "An analysis of a point-to-multipoint channel using a go-back-N error control protocol," *IEEE Transactions on Communications*, **COM-33**, 282-285
40. Chandran, S. R. and Lin, S. (1992), "Selective-repeat-ARQ scheme for broadcast links ," *IEEE Transactions on Communications*, **40**, 12-19
41. Deng, R. H. (1993), "Hybrid ARQ schemes for point-to-multipoint communication over nonstationary broadcast channels," *IEEE Transactions on Communications*, **41**, 1379-1387
42. Wang, J. L. and Silvester, J. A. (1993), "Optimal adaptive multireceiver ARQ protocols," *IEEE Transactions on Communications*, **41**, 1816-1829
43. Sakakibara, K. and Kasahara, M. (1995), "A multicast hybrid ARQ scheme using MDS codes and GMD decoding," *IEEE Transactions on Communications*, **43**, 2933-2940
44. Shiozaki, A. (1996), "Adaptive type-II hybrid broadcast ARQ system," *IEEE Transactions on Communications*, **44**, 420-422
45. Cam, R. and Leung, C. (1998), "Multiplexed ARQ for time-varying channels– Part I : system model and throughput analysis," *IEEE Transactions on Communications*, **46**, 41-51
46. Cam, R. and Leung, C. (1998), "Multiplexed ARQ for time-varying channels– Part II : postponed retransmission modification and numerical results," *IEEE Transactions on Communications*, **46**, 314-326
47. Anagnostou, M. E. and Protonotarios, N. E. (1986), "Performance analysis of the selective repeat ARQ protocol," *IEEE Transactions on Communications*, **COM-34**, 127-135
48. Yoshimoto, M., Takine, T., Takahashi, Y. and Hasegawa, T. (1993), "Waiting time and queue length distributions for go-back-N and selective-repeat ARQ protocols," *IEEE Transactions on Communications*, **41**, 1687-1693
49. Lu, D. L. and Chang, J. F. (1989), "Analysis of ARQ protocol via signal flow graphs," *IEEE Transactions on Communications*, **COM-37**, 245-251
50. Oduol, V. K. and Morgera, S. D. (1993), "Performance evaluation of the Generalized type II hybrid ARQ scheme with noisy feedback on Markov channels," *IEEE Transactions on Communications*, **41**, 32-40
51. Djuknic, G. M. and Schilling, D. L. (1994), "Performance analysis of an ARQ transmission scheme for meteor burst communications," *IEEE Transactions on Communications*, **42**, 268-271
52. Rice, M. (1994), "Application of generalized minimum distance decoding to hybrid-ARQ error control," *IEEE Transactions on Communications*, **42**, 640-647
53. Zorzi, M. and Rao, R. R. (1996), "On the use of renewal theory in the analysis of ARQ protocols," *IEEE Transactions on Communications*, **44**, 1077-1081
54. Nakagawa, T., Yasui, K. and Sandoh, H. (1993), "An optimal policy for a data transmission system with intermittent faults," *The Transactions of the Institute of Electronics, Information and Communication Engineers*, **J76-A**, 1201-1206 (in Japanese)
55. Yasui, K., Nakagawa, T. and Sandoh, H. (1995), "An ARQ policy for a data transmission system with three type of error probabilities," *The Transactions of the Institute of Electronics, Information and Communication Engineers*, **J78-A**, 824-830 (in Japanese)
56. Nakagawa, T. and Yasui, K. (1989), "Optimal testing-policies for intermittent faults," *IEEE Transactions on Reliability*, **38**, 577-580

57. Yasui, K. and Nakagawa, T. (1995), "Reliability considerations of a selective-repeat ARQ policy for a data communication system," *Microelectronics and Reliability*, **35**, 41-44
58. Yasui, K. and Nakagawa, T. (1997), "Reliability analysis of a hybrid ARQ system with finite response time," *The Transactions of the Institute of Electronics, Information and Communication Engineers*, **J80-A**, 221-227 (in Japanese)
59. Yasui, K., Nakagawa, T. and Imaizumi, M. (1998), "Reliability evaluations of hybrid ARQ policies for a data communication system," *International Journal of Reliability, Quality and Safety Engineering*, **5**, 15-28
60. Zheng, Q. and Shin, K. G. (1994), "On the ability of establishing real-time channels in point-to-point packet-switched networks," *IEEE Transactions on Communications*, **42**, 1096-1105
61. Zheng, Q., Shin, K. G. and Shen, C. (1995), "Real-time communication in ATM networks," *IEICE Transactions of the Institute of Electronics, Information and Communication Engineering*, **J78-D-I**, 639-648
62. Osaki, S. (1992), Applied Stochastic System Modeling. Springer-Verlag, Berlin

12. Quick Monte Carlo Methods in Stochastic Systems and Reliability

C. Papadopoulos and N. Limnios

Division Mathématiques Appliquées
Centre de Recherches de Royallieu,
Université de Technologie de Compiègne
BP 20529, 60205 Compiègne, France

Summary.

This paper aims with the evaluation of probability of rare events defined on Markovian systems by Monte Carlo methods. Optimal change of measure and failure biaising techniques are presented. This model is used to describe highly reliable Markovian systems. Finally, we discuss reliability evaluation by analytical statistical methods for non-Markovian systems of type k-out-of-n.

Keywords: Monte Carlo methods, rare event, Markovian system, non-Markovian system, k-out-of-n system, importance sampling, optimal importance sampling, failure biasing scheme, analytical statistical method

12.1 Introduction

In many systems design or analysis, an event whose probability of occurrence is quite small, is a key parameter of the system's efficacy. Consequently, in cases where the requirements for the reliability of the system are extremely high, such events cannot be neglected. Power systems, nuclear stations, communications systems, and computer networks are examples of systems that must attain a high reliability. The last one may be obtained either by choosing individual components having a highly reliable nature, or by increasing components' redundancy. Systems of this type may contain a large number of components, that may interact to each other in a complicated way. In order to circumvent the state space size explosion problem stemming from the fact that the state space of a system comprising n components may contain 2^n distinct states, state lumping or state aggregation techniques have been suggested [1], [17]. However, when such techniques are employed, a considerable amount of computer time and memory is needed. More importantly, the flexibility in modeling all interdependencies between different components of the system is limited, and it is usually very difficult to assess the error incurred through the state aggregation process. Moreover, the analysis of a system that is non-Markovian in nature is much more complicated, and in the general

S. Osaki (ed.), *Stochastic Models in Reliability and Maintenance*

case, analytical solutions are not available. For this type of system, the only tractable solution seems to be simulation.

Besides its simplicity, one of the major advantages of the Monte Carlo simulation method is that among all numerical methods using n points to produce an approximate solution in s-dimensional spaces, the Monte Carlo method has an absolute error that decreases as $n^{-1/2}$, while all other methods have errors decreasing as $n^{-1/s}$ at best. However, direct simulation of a rare event (whose probability of occurrence is usually less than 10^{-6}) is a formidable task. To get an idea of the probability we have to hit the event several times. Thus, direct simulation suffers from the drawback that the effort needed to estimate the probability of a rare event at a given level of precision increases as the event becomes rarer and rarer. Therefore, modifications of the direct simulation scheme have to be carried out in order to accelerate the simulation procedure. This may be done by reducing the variance of the corresponding estimator, and the associated methods are called variance reduction methods [2], [8], [18], [33], [34]. In this chapter, we deal with some of the variance reduction methods actually used in reliability systems, especially those ones that are based on the idea of importance sampling.

The organization of this chapter is as follows : In Section 12.2 we present the problem associated with direct simulation of a rare event, while in Section 12.3 we give some background theory concerning importance sampling. Section 12.4 discusses the optimal change of measure which in some cases may be obtained analytically. Some particular cases in which this change of measure can be recursively calculated are given in Section 12.5. The model used to describe highly reliable Markovian systems is briefly presented in Section 12.6, while Section 12.7 discusses the regenerative method of simulation. Some of the commonly used failure biasing schemes are presented in Section 12.8, while some principles of the optimal change of measure are used in Section 12.9 to develop heuristics for the unreliability estimation of the systems. Finally, in Section 12.10, we discuss the analytical statistical method for the unreliability estimation of k-out-of-n systems. We close this chapter with some global remarks.

12.2 The Problem with Direct Simulation

Consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and let X be a real random variable having density f. Suppose that we are interested in estimating the following quantity

$$\gamma = \mathbf{E}_f[h(X)] = \int_{-\infty}^{+\infty} h(x) f(x) dx, \tag{12.1}$$

where $h : \mathbf{R} \to \mathbf{R}$. In such cases, an unbiased estimator for γ will be $\hat{\gamma}_n = \frac{1}{n}\sum_{i=1}^{n} h(\omega^i)$, where $(\omega^1, \ldots, \omega^n)$ is an n-sample of f. Clearly, if $h(x) = \mathbf{1}_A(x)$ with $\mathbf{1}_A(x) = 1$, if $x \in A$ and 0 otherwise, then $\gamma = \mathbf{P}(X \in A)$. Furthermore, the variance of $\hat{\gamma}_n$ is equal to $\frac{\gamma(1-\gamma)}{n}$. Thus, the associated $100 \times (1-\delta)\%$, confidence interval will be

$$\left[\hat{\gamma}_n - z_{\delta/2}\sqrt{\frac{\hat{\gamma}_n(1-\hat{\gamma}_n)}{n}}, \quad \hat{\gamma}_n + z_{\delta/2}\sqrt{\frac{\hat{\gamma}_n(1-\hat{\gamma}_n)}{n}}\right],$$

where $z_{\delta/2}$ is defined by the equation $\delta/2 = \mathbf{P}(Z > z_{\delta/2})$ and Z denotes a random variable having the standard normal distribution $N(0,1)$. The natural way to construct the confidence interval will be to continue the simulation until the interval's half width becomes less than κ times the value of the parameter that we are trying to estimate. In other words, the stopping criterion for our simulation will be (for any $\kappa \in]0,1[$)

$$z_{\delta/2}\sqrt{\frac{\hat{\gamma}_n(1-\hat{\gamma}_n)}{n}} < \kappa\gamma, \text{ implying that } z_{\delta/2}\frac{\sqrt{\frac{\hat{\gamma}_n(1-\hat{\gamma}_n)}{n}}}{\gamma} < \kappa. \tag{12.2}$$

Thus, the relative error (RE) of an estimator $\hat{\gamma}_n$, which is defined as the ratio of its standard deviation to its expected value, will be given by (as $n \to +\infty$)

$$RE(\hat{\gamma}_n) = z_{\delta/2}\frac{\sqrt{\frac{\hat{\gamma}_n(1-\hat{\gamma}_n)}{n}}}{\gamma} \approx z_{\delta/2}\frac{1}{\sqrt{n\gamma}}, \quad \text{since } \hat{\gamma}_n \overset{n\to+\infty}{\longrightarrow} \gamma.$$

The last equation clearly illustrates the inconvenience of using direct simulation: the relative error of the estimator remains unbounded, while the event becomes rarer and rarer (*i.e.* $RE(\hat{\gamma}_n) \to +\infty$ when $\gamma \to 0$). It also means that in order for the equation (12.2) to be satisfied and thus obtain the desired relative precision of estimation, we have to considerably increase the size n of the sample. Put another way, in order to estimate γ up to a certain level of precision, one has to increase the number of iterations of the algorithm as the probability of the event becomes smaller and smaller.

12.3 Importance Sampling

In order to overcome this difficulty, one needs to apply importance sampling. Importance sampling changes the original probabilistic dynamics of the system, and at the same time modifies the function to be integrated.

Consider that we make the following change of measure in (12.1)

$$\gamma = \int_{-\infty}^{+\infty} h(x)\frac{f(x)}{f'(x)}f'(x)dx = \mathbf{E}_{f'}[h(X)L(X)], \tag{12.3}$$

where $L(X) = \frac{f(X)}{f'(X)}$ represents the likelihood ratio and the subscript f' means that the expected value is taken with respect to the new density f'.

The term importance sampling stems from the fact that the process is sampled in the areas that are important for the estimation of γ, for example the areas where the event $\{X \in A\}$ is realized if $h(x) = \mathbf{1}_A(x)$. In other words, f' has to be chosen in such a way as to make the rare event under consideration more likely to occur. Moreover, in order to keep the unbiasedness of the estimates, the results obtained have to be multiplied by the appropriate likelihood ratio. This is the compensatory factor, since the system has been simulated using a density which is not directly associated with the system's model.

Equation (12.3) is valid only in the case that $f'(x) > 0$, for every $x \in \mathbb{R}$ with $f(x) > 0$ and $h(x) > 0$, which implies that a possible value of X under f is also possible under f'. Note, however, that we can have $f'(x) = 0$ and $f(x) > 0$ for any $x \in \mathbb{R}$ with $h(x) = 0$. Consequently, the new unbiased estimator of

γ will be $\hat{\gamma}_n(f') = \frac{1}{n}\sum_{i=1}^{n} h(\omega^i)L(\omega^i)$, where the new n-sample $(\omega^1, \ldots, \omega^n)$ has now been generated using f'. Furthermore, its variance is given by

$$\begin{aligned} Var_{f'}[\hat{\gamma}_n(f')] &= \int_{-\infty}^{+\infty} h(x)\left(\frac{f(x)}{f'(x)}\right)^2 f'(x)dx - \gamma^2 \\ &= \mathbf{E}_f[h(X)L(X)] - \gamma^2. \end{aligned} \tag{12.4}$$

The aim of importance sampling is to find a suitable - and easily implementable - new density f', in order to minimize the variance of $\hat{\gamma}_n(f')$ and, by doing this, reduce the cost of the estimation procedure. In other words, when importance sampling is used a rare event must be realized more often, signifying that its new probability must be greater than the original one. Consequently, the L term in equation (12.4) should be kept as small as possible. Moreover, if L is uniformly less than one, then $Var_{f'}[\hat{\gamma}_n(f')] < \gamma - \gamma^2 = Var_f[\hat{\gamma}_n(f)]$ and we will certainly obtain a variance reduction. Another alternative would be to choose f' such that $\mathbf{E}_f[h(X)L(X)]$ is of the same order of magnitude as γ^2. In such cases the associated change of measure is sometimes called asymptotically efficient or asymptotically optimal (see the survey of Heidelberger [19] for a discussion on this matter).

Of course, the unconstrained optimal solution for (12.4), given by

$$f^*(x) = \frac{h(x)f(x)}{\gamma}, \tag{12.5}$$

always exists and is called "the optimal change of measure."

In Kuruganti and Strickland [24], an optimal change of measure is defined as the measure that results in a zero variance estimator for the unknown quantity. Using the optimal change of measure for the simulation, the exact value of the parameter will be obtained in the first trial. However, this has the disadvantage of containing γ, the parameter that we are trying to estimate, and for this reason it is not directly exploitable. Nevertheless, as will be shown in what follows, we can explicitly construct this optimal change of measure in certain special cases. This will enable us not only to estimate γ at a minimum cost, but also - and more importantly - to find its exact value as a by-product of the intermediate calculations [24], [25].

The interested reader can find the conditions for the applicability as well as the theoretical framework behind importance sampling in [13] by Glynn and Iglehart. In their work, importance sampling is extended to problems arising in the simulation of both discrete time and continuous time Markov processes, as well as in semi-Markov processes. In the same paper, the authors discuss the problem of steady state quantities estimation, which can be carried out by exploiting the regenerative structure of the Markov chain, and also the estimation of transient quantities, where a different approach has to be used.

12.4 The Optimal Change of Measure

12.4.1 Remarks

Consider again expression (12.5) and let $h(x) = \mathbf{1}_A(x)$. Then

$$f^*(x) = \frac{f(x)}{\gamma}\mathbf{1}_A(x). \tag{12.6}$$

will be the optimal change of measure associated with the estimation of $\gamma = \mathbb{P}(X \in A)$. Equation (12.6) seems at first sight to be of no use, since it contains the unknown parameter. However, it provides us with a very useful insight concerning the choice of the new density $f'(\cdot)$, since the following hold [41] :

- All the mass of the probability is concentrated on the event $\{X \in A\}$, and thus only samples corresponding to the realization of this event will be produced when the optimal change of measure is used to carry out the simulation.
- On A, the new density is nothing else than the conditional density of X, given that $\{X \in A\}$ has occurred

$$f^*(x) = \frac{f(x) \cdot \mathbf{1}_A(x)}{\mathbb{P}(X \in A)} = \begin{cases} f(x|X \in A), & x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Hence the relative likelihood of the values of X on A is the same for the original and the new distribution :

$$\frac{dF^*(x_1)}{dF^*(x_2)} = \frac{dF(x_1)}{dF(x_2)}, \qquad \text{where } x_1, x_2 \in A$$

and $F(\cdot)$ represents the cdf corresponding to the density $f(\cdot)$.

12.4.2 Preliminary definitions

Consider a continuous time Markov chain $X = \{X_t : t \geq 0\}$, with state space $E = \{0, 1, \ldots, s\}$, $s < +\infty$, infinitesimal generator $Q = \{q(x,y) : x, y \in E\}$, and initial law $\mu(\cdot)$. The quantity $q(x) = -q(x,x)$ represents the total rate out of state x. Suppose also that the state space of the system is divided into two disjoint subsets, U and F, with $U \cup F = E$ and $U \cap F = \emptyset$. The set $U = \{0, 1, \ldots, m\}$ represents the set of operational states, while $F = \{m+1, \ldots, s\}$ stands for the set of failed states of the system. In state 0, all components are considered new. Define also $0 = T_0 < T_1 < \ldots < T_n < \ldots$, the sequence of the successive jump times of the chain X. Then $Y = \{Y_n,\ n \geq 0\}$, defined by

$$Y_n = X_{T_n}, \qquad n = 0, 1, \ldots,$$

will be the embedded discrete time Markov chain associated with X_t. The elements of its transition matrix $P = \{P(x,y)\colon x, y \in E\}$, are given by $P(x,y) = q(x,y)/q(x)$, when $x \neq y$ and 0 otherwise. Let us also define :

- $\Gamma = \{(x,y)\colon x, y \in E \text{ and } P(x,y) > 0\}$: the set of all feasible transitions;
- $F(x) = \{y \in E\colon (x,y) \text{ is a failure transition}\}$;
- $R(x) = \{y \in E\colon (x,y) \text{ is a repair transition}\}$;
- $T_B = \inf\{t \geq 0\colon X_t \in B\}$: the hitting time for $B \subset E$ (with the convention that $\inf \emptyset = +\infty$);
- $\tau_B = \inf\{n \geq 0\colon Y_n \in B\}$: the number of jumps for Y to enter $B \subset E$;
- $\gamma = \mathbf{P}(\tau_F < \tau_0)$: the probability that the system starting from state 0, reaches F without returning to the initial state.

The probability $\mathbf{P}(\tau_F < \tau_0)$ is a performance measure of great importance, since it is associated with:

- the mean time of failure of the system (MTTF) (see Section 12.7);
- the unreliability of the system at time t [38].

12.4.3 The recursive approach

Kuruganti and Strickland [24], [25] have obtained the optimal change of measure for the estimation of $\mathbf{P}(\tau_F < \tau_0)$, which also corresponds to a discrete time Markov chain. This optimal change of measure is described by the transition matrix $P^* = \{P^*(x,y)\colon x,y, \in E\}$, whose elements must be chosen such that the following conditions hold :

- **Unfeasibility conditions**
  Using the optimal change of measure, the event has to be realized at every trial. For the case of $\mathbf{P}(\tau_F < \tau_0)$ estimation, it is therefore necessary to eliminate all transitions going to state 0, from any state $x \in U - \{0\}$. For every pair of states (x,y), if $P(x,y) > 0$, or if all paths from y to F include state 0, then $P^*(x,y) = 0$.
- **Direct path conditions**
  Let us define a direct path, as a sequence of states visited exactly once and going from state 0 to any state in F. Then the likelihood ratio of the initial to the new measure for all direct paths of the chain for which the event $\{\tau_F < \tau_0\}$ is realized is constant and equal to γ.
- **Loop conditions**
  Also define a loop, as a sequence of states visited exactly once and whose first and last elements coincide. Then the likelihood ratio of the initial to the new measure for all loops along a path where $\{\tau_F < \tau_0\}$ is realized is constant and equal to 1.
- **Stochastic matrix conditions**

$$\sum_{y \in E} P^*(x,y) = 1, \qquad \forall x \in E.$$

Note that the cycle (loop) conditions are trivially satisfied when the importance sampling distribution is the original one. However, the first two optimality conditions are not satisfied.

By also defining $\gamma(x)$ as the probability that the system reaches F starting from state x without visiting state 0, Kuruganti and Strickland [24], [25] obtained the situation where the elements of the new transition matrix are given by

$$P^*(x,y) = \frac{P(x,y)\gamma(y)}{\gamma(x)}. \tag{12.7}$$

In this case, $P(x,y)\gamma(y)$ denotes the contribution to $\gamma(x)$ of any path starting from state x with the transition (x,y) and realizing the event $\{\tau_F < \tau_0\}$. Under the optimal change of measure the probability of the (x,y) transition must be in proportion to $P(x,y)\gamma(y)$. Moreover, the constant of proportionality must be exactly $\gamma(x) = \sum_{y \in F(x) \cup R(x)} P(x,y)\gamma(y)$, the probability that the system reaches F starting from x without visiting state 0. Using this representation for the new transition matrix, it is trivial to see that the optimality conditions are verified. It is necessary, however, to underline the importance of the elimination of all transitions going to state 0: *any arbitrary simulation scheme for the estimation of* $\mathbf{P}(\tau_F < \tau_0)$ *can be improved by eliminating such transitions.* Moreover, Kuruganti and Strickland have also proposed a way of calculating the optimal change of measure, as well as the variance of the corresponding estimator, recursively [26].

Juneja, in [10], selected importance sampling distributions for Markov-additive processes, satisfying at least one of the following two conditions for the corresponding likelihood ratio:

- Equi-probable cycle condition (EPC), where the likelihood ratio is one for all cycles (loops);
- Dominant probability cycle condition (DPC), where the likelihood ratio is bounded by one for all cycles (loops);

Juneja demonstrated that with importance sampling measures satisfying one of the previous two conditions, a large variance reduction can be obtained in estimating rare event probabilities. Furthermore, using large deviations arguments he showed that in the case of stochastic systems formed by interaction of independent Markov additive processes, only exponentially twisted distributions can satisfy the EPC condition.

Example 12.4.1 Consider the state diagram given on the left-hand side of figure 12.1, representing a Markov chain with state space $E = \{0, 1, \ldots, 5\}$, $F = \{5\}$ and suppose that we are interested in estimating $\mathbf{P}(\tau_F < \tau_0)$ using simulation. By eliminating all the repair transitions going to state 0, we obtain the transition diagram given on the right-hand side of the same figure. In order to evaluate the new transition probabilities $p^*_{ij} = P^*(i,j)$, we have to make use of the following system of equations :

- **Unfeasibility conditions**;

$$p^*_{10} = 0 = p^*_{40};$$

- **Direct path conditions**;

$$\frac{p_{01}p_{13}p_{35}}{p^*_{01}p^*_{13}p^*_{35}} = \frac{p_{02}p_{24}p_{45}}{p^*_{02}p^*_{24}p^*_{45}} = \text{constant} = \mathbf{P}(\tau_F < \tau_0);$$

- **Loop conditions**;

$$p_{13}p_{31} = p^*_{13}p^*_{31} \quad \text{and} \quad p_{24}p_{42} = p^*_{24}p^*_{42};$$

- **Stochastic matrix conditions**;

$$p^*_{01} + p^*_{02} = p^*_{13} = p^*_{24} = p^*_{31} + p^*_{35} = p^*_{42} + p^*_{45} = 1.$$

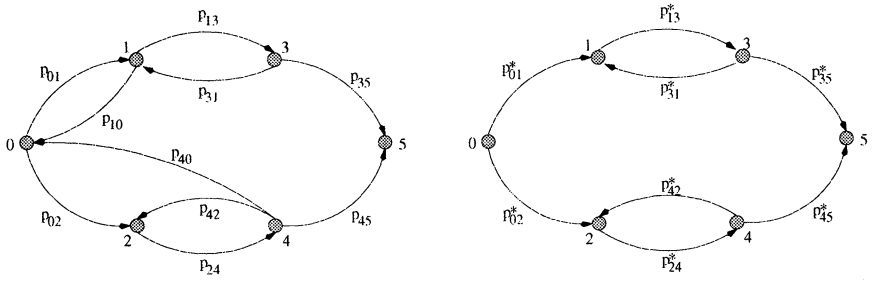


Fig. 12.1. System's state diagram before and after the transformation

$$P = \begin{bmatrix} \cdot & 0.1 & 0.9 & \cdot & \cdot & \cdot \\ 0.8 & \cdot & \cdot & 0.2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1.0 & \cdot \\ \cdot & 0.9 & \cdot & \cdot & \cdot & 0.1 \\ 0.6 & \cdot & 0.3 & \cdot & \cdot & 0.1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad \text{and } P^* = \begin{bmatrix} \cdot & 0.017 & 0.983 & \cdot & \cdot & \cdot \\ 0.0 & \cdot & \cdot & 1.0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1.0 & \cdot \\ \cdot & 0.18 & \cdot & \cdot & \cdot & 0.82 \\ 0.0 & \cdot & 0.3 & \cdot & \cdot & 0.7 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

The original (P) and the new transition probability matrix (P^*) are given above. Thus, in order to find the exact value of $\gamma = \mathbf{P}(\tau_F < \tau_0)$, we just have to calculate the likelihood ratio for any (direct) path going from state 0 to state 5. For the direct path $0 \to 2 \to 4 \to 5$, we obtain

$$\gamma = \frac{p_{02}p_{24}p_{45}}{p^*_{02}p^*_{24}p^*_{45}} = 0.130793.$$

We obtain the same result as expected by using the direct path $0 \to 1 \to 3 \to 5$:

$$\gamma = \frac{p_{01}p_{13}p_{35}}{p^*_{01}p^*_{13}p^*_{35}} = 0.130793.$$

See also [25] for an application of the optimal importance sampling to tandem queues.

12.4.4 Exact calculation of $\gamma(x)$

In order to calculate the probabilities $\gamma(x)$ for any $x \in E$, it is sufficient to note that:

$$\gamma(x) = \sum_{y \in E - \{0\}} P(x,y)\gamma(y), \qquad \forall x \in E, \tag{12.8}$$

and consequently solve the system of s-type (12.8) equations. In particular, we are interested in finding $\gamma(0) = \mathbf{P}(\tau_F < \tau_0)$.

12.5 Cases of Application of the Recursive Approach

The simplest case in which the recursive approach can be applied is the case of Birth-Death Markov chains. Consider such a chain $Y = \{Y_n, n \geq 0\}$, with state space $E = \{0, 1, \ldots, s\}$, $F = \{s\}$ and whose state-diagram is given in figure 12.2. Let us also denote as p_i the transition probability $\mathbf{P}(Y_n = i+1 | Y_{n-1} = i)$, with $p_0 = 1$ and $p_s = 0$.

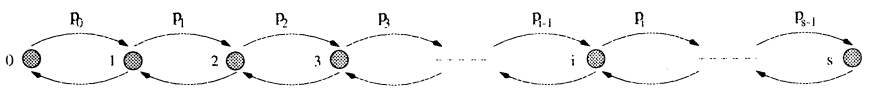


Fig. 12.2. State diagram for a birth-death process

In order to find the optimal change of measure, the following system of equations has to be solved ($\forall i = 1, \ldots, s-2$) (loop conditions)

$$p_i(1-p_{i+1}) = p_i^*(1-p_{i+1}^*) \Rightarrow$$
$$p_{i+1}^* = 1 - \frac{p_i(1-p_{i+1})}{p_i^*}. \tag{12.9}$$

Given that the transition to state 0 will be eliminated (unfeasibility condition), we obtain $p_1^* = 1$, and using equation (12.9) we can therefore recursively calculate all consecutive p_i^* values, for $i = 2, \ldots, s-1$. The procedure remains the same when we have more than one transitions to state 0, from different states, as illustrated in figure 12.3. These transitions will be eliminated and will not affect the system of equations to be solved.

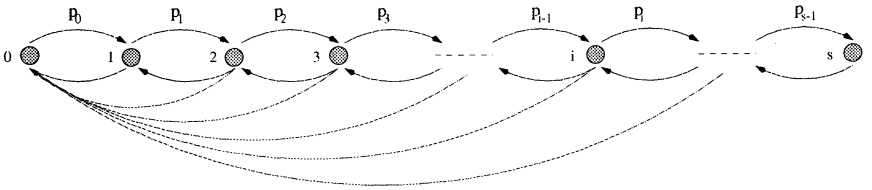


Fig. 12.3. Multiple transitions to state 0

For systems having more than one repair transition at any state i (but exactly one failure transition), the procedure is still simple and efficient, since it remains recursive. Consider, for instance, the system whose state diagram is illustrated in figure 12.4. In order to find the optimal change of measure, we first have to eliminate all transitions going to state 0. We will thus obtain $p_0^* = 1$ and $p_1^* = 1$. Using the optimality condition concerning the cycle $1 \rightarrow 2 \rightarrow 1$, we can calculate the value of p_2^*. Again, using the condition for the cycle $2 \rightarrow 3 \rightarrow 2$, we obtain the value of p_{32}^*, while the corresponding condition for the cycle $2 \rightarrow 3 \rightarrow 1 \rightarrow 2$, will give us the value of p_{31}^* and thus $p_3^* = 1 - p_{31}^* - p_{32}^*$. In a similar fashion, all transition probabilities can be calculated.

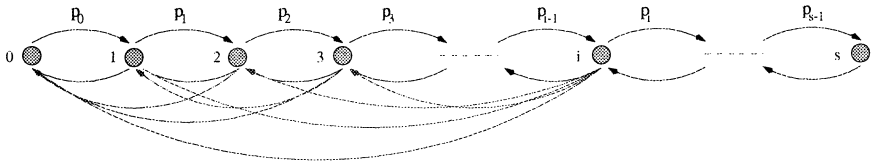


Fig. 12.4. Multiple repairs at each state

Consider now the case of a system whose state space is $E = \{0, 1, 2, \ldots, s\}$, with $F = \{s\}$, as in figure 12.5. The subsets G_i, $i = 1, \ldots, k$ are as in figure 12.4. From any state of these subsets, we may have a transition to state 0. Furthermore, from state 0, multiple transitions to different states in the subsets G_i, $i = 1, \ldots, k$, are allowed. However, only the last state of these sets communicates with state s and, moreover, there are no transitions between these sets.

For this system, we can easily obtain the new transition probabilities - in the way described previously - except from p_{0i}^*, for $i = 1, \ldots, k$. In order to

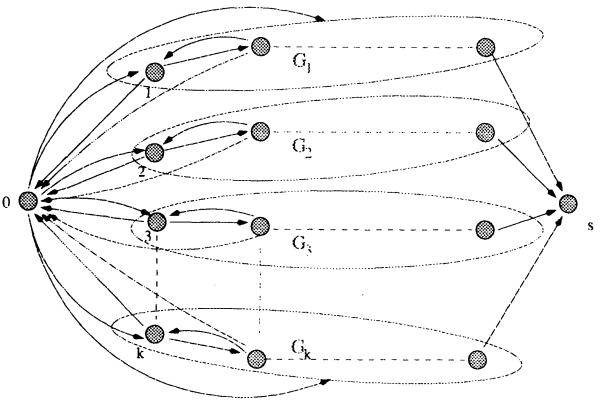


Fig. 12.5. State diagram with k subsets G_i

find them, we have to use the conditions associated with the k direct paths and also the fact that $\sum_{j=1}^{k} p_{0j}^{*} = 1$. Note, however, that from the $\frac{k(k-1)}{2}$ equations associated with the direct paths that we will obtain, only the $(k-1)$ - those combining the i-th direct path with the $(k-1)$ remaining direct paths - are independent.

In the simple case of a birth and death process, Juneja in [10] has obtained the optimal change of measure, using its dominant probability cycle condition. Using similar arguments, he also obtained the optimal change of measure for the M/M/1 queue that consists in interchanging the arrival rate to the service rate of customers. The same result has also been demonstrated by Cottrell *et al.* in [6], using a large deviations approach.

Stationary distribution

Suppose that the initial chain has a stationary law denoted by π and also that under P^* the new stationary law is π^*. Since π is the stationary law of the initial chain, then it must satisfy $\pi P = \pi$ and thus

$$\sum_{x \in E} \pi(x) P(x,y) = \pi(y) \Rightarrow \text{(from (12.7))}$$

$$\sum_{x \in E} \pi(x) \frac{P^*(x,y)\gamma(x)}{\gamma(y)} = \pi(y) \Rightarrow \sum_{x \in E} \pi(x)\gamma(x) P^*(x,y) = \pi(y)\gamma(y),$$

implying that

$$\pi^*(x) = \frac{\pi(x)\gamma(x)}{\sum_{x \in E} \pi(x)\gamma(x)}.$$

12.6 System Model

We consider a system having C different types of components, $0 < C < +\infty$, with n_i, $0 < n_i < +\infty$ components of each type. Let $N = \sum_{i=1}^{C} n_i$ be the total number of components of the system. The components may fail and consequently get repaired. The repair discipline is arbitrary, but it follows the

"work conserving rule," where the repairmen are occupied as long as there are unfinished repairs. This system can be described by the following process [14]

$$X_t = (X_1(t), X_2(t), \ldots, X_C(t)), \qquad t \geq 0,$$

where $X_i(t)$ is the number of type i components that are operational at time t. For a more general class of highly reliable systems, we have to add some more state descriptors in the state of the system, for instance the list of components waiting to be repaired, etc.

An interpretation of the term "highly reliable" for a Markovian system is the fact that the failure rates of individual components are much smaller than the corresponding repair rates. Let μ_{min} be the minimum repair rate of a component and λ_{max} the maximum failure rate. Without loss of generality we can suppose that $\mu_{min} = 1$, while we note $\varepsilon = \lambda_{max}$ for the parameter that reflects the highly reliable nature of the system, which is called the "rarity parameter" of the system. We define also a function $f(\varepsilon)$ to be $o(\varepsilon^d)$ if $f(\varepsilon)/\varepsilon^d \to 0$ when $\varepsilon \to 0$, where d is a constant. In a similar fashion, $f(\varepsilon) = O(\varepsilon^d)$ if $|f(\varepsilon)| \leq c_1\varepsilon^d$ for a constant $c_1 > 0$ and for all ε that are sufficiently small. Moreover, $f(\varepsilon) = \underline{O}(\varepsilon^d)$ if $|f(\varepsilon)| \geq c_2\varepsilon^d$ for a constant $c_2 > 0$. Finally, $f(\varepsilon) = \Theta(\varepsilon^d)$ if $f(\varepsilon) = O(\varepsilon^d)$ and $f(\varepsilon) = \underline{O}(\varepsilon^d)$. For a more detailed description of the system model the reader is referred to [29], [30], [36].

Shahabuddin [36] has made the following assumptions concerning the system model :

Assumption 12.6.1
- The Markov chain is irreducible over the set E.
- All states in $E - \{0\}$ have at least one repair transition.
- All states in U have at least one repair transition.

Using the previous assumptions, Shahabuddin [36] has demonstrated that the elements of the transition matrix P corresponding to a failure transition may be written as $c\varepsilon^d + o(\varepsilon^d)$, $c \in (0, 1]$, $d \geq 1$, while those that correspond to a repair transition can be written as $c + o(1)$, $c \in (0, 1]$. Consequently, if all failure probabilities are of the same ε-order then the system is called *balanced*, and otherwise it is called *unbalanced*. For the transition from state 0 - whose probabilities are written as $c\varepsilon^d + o(\varepsilon^d)$, $c > 0$, $d \geq 0$ - he has made a supplementary assumption, namely

Assumption 12.6.2
- For all $y \in F$, $P(0, y)$ is either 0, or it has the form $c\varepsilon^d + o(\varepsilon^d)$, $c > 0$, $d > 0$.

The last assumption means that the system cannot go directly to a failed state from state 0, or if such a transition exists its probability is much smaller than all other failure transitions. If this assumption is not satisfied then the event $\{\tau_F < \tau_0\}$ (as well as system failure) is no longer a rare event since in this case we will obtain $\gamma = c + o(1)$, $c > 0$.

Shahabuddin has also demonstrated that the probability $\gamma = \mathbf{P}(\tau_F < \tau_0)$ and the variance of the estimator $\hat{\gamma}_n$ may be represented as $a_0\varepsilon^r + o(\varepsilon^r)$, where a_0 and r are positive constants.

12.7 Regenerative Simulation

Consider τ a stopping time for $\{Y_n : n \geq 0\}$. That means that the realization or not of the event $\{\tau = n\}$ may be determined by $Y^n \equiv (Y_0, \ldots, Y_n)$. Let us note E_n the set of all possible paths of the chain Y until time n

$$E_n \equiv \{y^n = (y_0, y_1, \ldots, y_n) : y_j \in E\}.$$

Then, for any $y^n \in E_n$, we have

$$\mathbf{P}(y^n) \equiv \mu(y_0)P(y_0, y_1)\ldots P(y_{n-1}, y_n),$$

where $\mu(y_0) = \mathbf{P}(Y_0 = y_0)$, the initial law of the chain. Also let B_n stand for the set of paths for which $\{\tau = n\}$. It is then obvious that $B_n \subset \Omega_n$.

Proposition 12.7.1 (Goyal *et al.* [16])
Consider a discrete time Markov chain with transition probability matrix P. Let $\mathbf{P}$ be the probability measure associated with the different trajectories of the chain and τ a stopping time which is finite under $\mathbf{P}$, with probability 1. Note also Z, a measurable function of Y^τ for which $\mathbf{E_P}[|Z(Y^\tau)|] < \infty$. Let $\mathbf{P}'$ be a new probability measure for which τ is also finite with probability 1 and for any $y^n \in B_n$, $\mathbf{P}'(y^n) \neq 0$ whenever $Z(y^n)\mathbf{P}(y^n) \neq 0$. Then $\mathbf{E_P}[Z(Y^\tau)] = \mathbf{E_{P'}}[Z(Y^\tau)L(Y^\tau)]$, with $L(Y^\tau) = \frac{\mathbf{P}(y^n)}{\mathbf{P}'(y^n)}$, for any $y^n \in B_n$. □

We have to note here that the new importance sampling measure may not necessarily correspond to a time-homogeneous Markov chain. We can use, for example, a measure $\mathbf{P}'$

$$\mathbf{P}'(y^n) = \mathbf{P}'(y_0)\mathbf{P}'(y_1|y_0)\ldots\mathbf{P}'(y_n|y_0\ldots y_{n-1}),$$

where $\mathbf{P}'(y_n|y_0\ldots y_{n-1})$ represents the likelihood of the path $Y_n = y_n$ given that $Y^{n-1} = (y_0, \ldots, y_{n-1})$. Such a measure used in importance sampling is called a "dynamic importance sampling measure" (DIS, see [36] and references therein).

The steady state unavailability of the system α represents the fraction of time for which the system is considered failed. Let $h(y) = 1/q(y)$ be the mean sojourn time in state y and let $g(y) = \mathbf{1}_F(y)h(y)$. We can then write [7]

$$\alpha = \frac{\mathbf{E_P}\left[\sum_{k=0}^{\tau_0-1} g(Y_k)\right]}{\mathbf{E_P}\left[\sum_{k=0}^{\tau_0-1} h(Y_k)\right]}. \tag{12.10}$$

For the estimation of the MTTF (mean time to failure) of the system, we have the following representation [16]

$$\begin{aligned}\text{MTTF} = \mathbf{E_P}[T_F] &= \frac{\mathbf{E_P}[\min(T_0, T_F)]}{\mathbf{P}(T_F < T_0)} \\ &= \frac{\mathbf{E_P}\left[\sum_{k=0}^{\min(\tau_0,\tau_{F'})-1} h(Y_k)\right]}{\mathbf{E_P}\left[\mathbf{1}_{\{\tau_{F'}<\tau_0\}}\right]}\end{aligned} \tag{12.11}$$

which is the key relation for the MTTF estimation using Monte Carlo. Consequently, the general problem of estimation can be formulated as the ratio of two expectations

$$\eta = \frac{\mathbf{E}_{\mathbf{P}}[G]}{\mathbf{E}_{\mathbf{P}}[H]},$$

where G and H for the case of steady state unavailability and MTTF estimation have been given above. In the case of MTTF estimation, the main problem of the simulation lies in our difficulty in estimating the denominator that corresponds to a rare event, while for the numerator we can simply use direct simulation. In the case of unavailability estimation, we use importance sampling to estimate the numerator and direct simulation for the denominator.

Simulation of the two parts of the ratio can be carried out independently, and so we can use a different number of samples to estimate $\mathbf{E}_{\mathbf{P}}[G]$ and $\mathbf{E}_{\mathbf{P}}[H]$. This method is called "measure-specific dynamic importance sampling" (MSDIS; see [16] and [15]). Suppose that a total number of n regenerative cycles will be used for the simulation. Suppose also that the first ζn cycles,[1] with $0 < \zeta < 1$, will be generated using $\mathbf{P}_1$, while the remaining $(1-\zeta)n$ cycles will use $\mathbf{P}_2$ ($\mathbf{P}_1 \neq \mathbf{P}_2$). Let L represent the likelihood ratio between the original and the new measure and note as G_j, L_j, and H_j the samples of G, L, and H respectively from the i-th regenerative cycle. We can then construct the following estimator for η

$$\hat{\eta}_{(n,\zeta)} = \frac{\frac{\sum_{j=1}^{\zeta n} G_j L_j}{\zeta n}}{\frac{\sum_{j=\zeta n+1}^{n} H_j L_j}{(1-\zeta)n}}.$$

This is a consistent estimator for η, and we have [16]

$$\lim_{n \to +\infty} \hat{\eta}_{(n,\zeta)} = \eta$$

with probability 1. Moreover,

$$\sqrt{n}(\hat{\eta}_{(n,\zeta)} - \eta) \xrightarrow{d} N(0, \sigma^2(\mathbf{P}_1, \mathbf{P}_2)/\mathbf{E}^2_{\mathbf{P}_2}[H]), \tag{12.12}$$

where "$\xrightarrow{d}$" denotes the convergence in distribution and

$$\sigma^2(\mathbf{P}_1, \mathbf{P}_2) = \frac{Var_{\mathbf{P}_1}[GL]}{\zeta} + \eta^2 \frac{Var_{\mathbf{P}_2}[HL]}{(1-\zeta)}. \tag{12.13}$$

Thus, in the case of steady-state unavailability estimation we can take $\mathbf{P}_1 \neq \mathbf{P}$ and $\mathbf{P}_2 = \mathbf{P}$, while in the case of MTTF estimation we can choose $\mathbf{P}_1 = \mathbf{P}$ and $\mathbf{P}_2 \neq \mathbf{P}$.

12.8 Failure Biasing Methods

12.8.1 Simple failure biasing (SFB)

This method is also known as "Bias1". Introduced for the first time by Lewis and Böhm in [27], it has since been modified by Shahabuddin [16], [39]. Simple

[1] Suppose that $\zeta n \in \mathbb{N}$.

failure biasing increases the probability of failure at any state $x \in U$, by allocating ρ $(0 < \rho < 1)$ as the new total failure probability, and consequently decreases the corresponding repair probability. New transition probabilities are attributed proportionally to the original ones, preserving in this way the probabilistic structure of the system. Let us note $P' = \{P'(x,y) : x,y \in E\}$ for the new transition probability matrix. Then the change of measure obtained by SFB is described by :

$$P'(x,y) = \begin{cases} P(0,y), & \text{for } x = 0,\ y \in F(0), \\ \rho \frac{P(x,y)}{\sum_{y \in F(x)} P(x,y)}, & \text{for } y \in F(x),\ x \in U - \{0\}, \\ (1-\rho) \frac{P(x,y)}{\sum_{y \in R(x)} P(x,y)}, & \text{for } y \in R(x),\ x \in U - \{0\}, \\ P(x,y), & \text{for } x \in F, \\ 0, & \text{otherwise.} \end{cases}$$

12.8.2 Balanced failure biasing (BFB)

A method slightly different from SFB is the method called balanced failure biasing. Introduced by Shahabuddin in [35] this method is also known as "Bias1/Balancing." It differs from SFB in the way the new probabilities are attributed to different states. This time they are attributed uniformly. The change of measure is described by :

$$P'(x,y) = \begin{cases} \rho \frac{1}{|F(x)|}, & \text{for } y \in F(x), x \in U, \\ (1-\rho) \frac{P(x,y)}{\sum_{y \in R(x)} P(x,y)}, & \text{for } y \in R(x), x \in U - \{0\}, \\ P(x,y), & \text{for } x \in F, \\ 0, & \text{otherwise.} \end{cases}$$

Both of the above techniques increase the failure transition probabilities, and thus the event $\{\tau_F < \tau_0\}$ becomes more likely to occur. Simple failure biasing is the most natural way to accelerate failures of individual components, since it preserves the system's original underlying probabilistic structure. It involves the inconvenience, however, of allocating a new transition probability that may still depend on the rarity parameter of the system. This happens when the system is unbalanced and some failure transitions that lie along the most likely path to failure have an $O(\varepsilon)$ probability. Using SFB, these paths will not be significantly emphasized. In order to circumvent this difficulty and thus have transition probabilities of the same order of magnitude, BFB has to be used. Using this method, all transition probabilities are $\Theta(1)$ and therefore independent of the rarity parameter ε [19], [30], [29], [36].

Shahabuddin [36] gave an example of an unbalanced system for which simple failure biasing cannot give bounded relative error, while in [29], Nakayama constructed an unbalanced system for which simple failure biasing gives bounded relative error. It has been demonstrated that simple failure biasing has the property of bounded relative error only for balanced systems, while when balanced failure biasing is used the property holds for both balanced and unbalanced systems [35], [36], [10]. However, it may be the case that in some systems SFB has a better performance than BFB.

Concerning the failure biasing constant ρ, Goyal *et al.*, in [16], suggested using values between 0.5 and 0.9, since in general the best value of ρ is difficult to obtain.

12.8.3 Bias2 failure biasing

This method developed by Goyal *et al.* [16] changes the transition matrix in a way similar to the SFB method. The new failure probability at any state $x \in U - \{0\}$ becomes ρ_0 and consequently the new repair probability becomes equal to $1 - \rho_0$. Moreover, this method gives preference to failures of those type of components that have some components already failed. This is done via the constant ρ_1. In order to describe this method, we need first to define

$$F_2(x) = \{y \in F(x) : n_i(y) < n_i(x) < n_i(0) \text{ for at least one type } i\}.$$

This set contains all states y for which the transition (x, y) has at least a type i component that has already failed at state x, and at least one component of this type fails on this transition. Then

$$F_1(x) = F(x) - F_2(x),$$

will be the set of all the other failure transitions, from state x. Moreover, let

$$P_{F_i}(x) = \sum_{z \in F_i(x)} P(x, z), \quad \text{for } i = 1, 2,$$

be the total probability of taking a failure transition in $F_i(x)$ from state x. The new transition probability matrix is described by:

- for $(x, y) \notin \Gamma$,
$$P'(x, y) = 0;$$
- at state 0, $P'(0, y) = P(0, y)$, $\forall y \in F(0)$ (no changes);
- at state $x \in U - \{0\}$, we have :
  - if $F_1(x) \neq \emptyset$ and $F_2(x) \neq \emptyset$, then :

$$P'(x, y) = \begin{cases} \rho_0 \rho_1 \frac{P(x,y)}{P_{F_2}(x)}, & \text{for } y \in F_2(x), \\ \rho_0(1 - \rho_1)\frac{P(x,y)}{P_{F_1}(x)}, & \text{for } y \in F_1(x), \\ (1 - \rho_0)\frac{P(x,y)}{P_R(x)}, & \text{for } y \in R(x), \\ 0, & \text{otherwise;} \end{cases}$$

  - if $F_1(x) = \emptyset$, or $F_2(x) = \emptyset$, then :

$$P'(x, y) = \begin{cases} \rho_0 \frac{P(x,y)}{P_F(x)}, & \text{for } y \in F(x), \\ (1 - \rho_0)\frac{P(x,y)}{P_R(x)}, & \text{for } y \in R(x), \\ 0, & \text{otherwise;} \end{cases}$$

- finally for $(x, y) \in \Gamma$ and $x \in F$,
$$P'(x, y) = P(x, y).$$

This method has the bounded relative error (BRE) property for balanced systems, but can give unbounded relative error in the case of unbalanced systems [30]. A slight modification of this method, called Bias2 Balanced Failure Biasing (where the probabilities are allocated to individual transitions in a uniform fashion) gives rise to bounded relative error for any type of system (see [30] for details).
The empirical values for the constants ρ_0 and ρ_1 - proposed by Goyal *et al.* [16] - are 0.8 and 0.7.

12.8.4 Failure distance biasing (FDB)

This method is based on the notion of failure distance. Introduced and developed by Carrasco in [3] and [4], distance biasing involves the inconvenience of necessitating a lot of information about the system. This information is not always easy to obtain (for instance the minimal cut sets, whose calculus is extremely long). In order to describe the method, let us first define for any state $x \in U$, the failure distance as

$$d(x) = \min_{y \in F} \left(\sum_{i=1}^{C} (n_i(x) - n_i(y)) \right)$$

and $d(x) = 0$ for $x \in F$. The failure distance represents the minimal number of components whose failure in state x would take the system down.

Using this definition, transitions can be classified in two different categories. We say that a failure transition $(x, y) \in \Gamma$ is *dominant* if $d(y) < d(x)$. In the opposite case, it is called *non-dominant*. Moreover, (x, y) is *critical* when $d(y) < d(x) - 1$, where the criticality of the transition is defined as $c(x, y) = d(x) - d(y)$. Consider also two constants ρ_d, ρ_c, $0 < \rho_d, \rho_c < 1$, which are independent of the rarity parameter ε. Distance biasing does not change the transition probabilities from a failed state. Take a state $x \in U - \{0\}$. The new transition probabilities will be allocated proportionally to the original ones. The total failure probability at state x becomes ρ_0, while the corresponding repair probability becomes $(1 - \rho_0)$. After that, the set $F(x)$ is divided into two sets, the first one containing the dominant transitions while the second one contains the non-dominant transitions. A conditional probability $(1 - \rho_d)$ is then allocated to the non-dominant transitions, and consequently we allocate ρ_d to the dominant transitions.[2] Furthermore, the set of dominant transitions is divided into a set containing the transitions of minimal criticality, where we allocate $(1 - \rho_c)$ and all the other transitions where we allocate ρ_c. This last step of the procedure can be repeated recursively until the time we obtain transitions of the same criticality in every set. The values proposed empirically by Carrasco [3], [4] are : $\rho_0 = 0.8$, $\rho_d = 0.7$ and $\rho_c = 0.2$.

Introduced informally by Shahabuddin [36], the following two methods are slight modifications of the BFB. They both have the bounded relative error property for any type of system.

12.8.5 Balanced 1 failure biasing (B1FB)

This method is similar to SFB, but this time transition probabilities are distributed uniformly. The new transition matrix is described by

$$P'(x, y) = \begin{cases} \frac{1.0}{|F(0)|}, & \text{for } x = 0,\ y \in F(0), \\ \frac{1.0}{|F(x)+R(x)|}, & \text{for } y \in F(x) \cup R(x),\ x \in U - \{0\}, \\ P(x, y), & \text{for } x \in F, \\ 0, & \text{otherwise.} \end{cases}$$

This method can be modified by introducing the failure biasing constant once more to give preference to failures of individual components.

[2] If one of these two sets is empty the allocation does not take place.

12.8.6 Balanced 2 failure biasing (B2FB)

$$P'(x,y) = \begin{cases} \rho \frac{1}{|F(x)|}, & \text{for } y \in F(x), \\ (1-\rho)\frac{1}{|R(x)|}, & \text{for } y \in R(x), \\ P(x,y), & \text{for } x \in F, \\ 0, & \text{otherwise.} \end{cases}$$

Remark : As we have already mentioned, the optimal change of measure concerning the estimation of the probability $\mathbf{P}(\tau_F < \tau_0)$, implies the elimination of all transitions of the Markov chain going to state 0. Consequently, any failure biasing method used to estimate this probability can be further improved by eliminating such transitions.

12.8.7 Bounded relative error and failure biasing

Bounded relative error is a guarantee for the efficiency of a method. However, a method that does not have this property is not necessarily less powerful than any other that satisfies it.

In [30], Nakayama established a necessary and sufficient condition for a failure biasing method to have bounded relative error. This condition, however, is very difficult to verify in practice, since one has to examine a large number of sample paths of the Markov chain. He also gave a sufficient condition for bounded relative error to hold: *the elements of the new probability matrix must be independent of the rarity parameter, ε, for all feasible transitions of the chain.* This result was established before by Shahabuddin in [36], using an approach based on matrix calculus.

In the following table, we summarize when the bounded relative error property is verified by the different failure biasing methods and we also give the suggested empirical values for the failure biasing constants.

Method	Balanced system	Unbalanced system	Suggested values
SFB	BRE	not always	$\rho \in [0.5, 0.9]$
BFB	BRE	BRE	$\rho \in [0.5, 0.9]$
Bias2	BRE	not always	$(\rho_0, \rho_1) = (0.8, 0.7)$
FDB	BRE	not always	$(\rho_0, \rho_d, \rho_c) = (0.8, 0.7, 0.2)$
B1FB	BRE	BRE	$\rho \in [0.5, 0.9]$
B2FB	BRE	BRE	$\rho \in [0.5, 0.9]$

12.9 Unreliability Estimation

12.9.1 One-component system

Consider a one-component system that may, at a given time t, be either operational or failed. When it has failed a repairman repairs the component and the system is as good as new again. This system may be described by a two-state Markov process having state space $E = \{0, 1\}$. State 0 is the up state of the system while state 1 is the down state ($F = \{1\}$). Let λ and μ be the failure and repair rates of the component and suppose that we are interested

in estimating the unreliability of the system at time t, denoted by $\bar{R}(t)$. This quantity stands for the probability that the system fails before t

$$\bar{R}(t) = \mathbf{P}(T_0 \leq t) = F_0(t) = 1 - e^{-\lambda t}, \tag{12.14}$$

where T_0 is the sojourn time in state 0 and $F_0(\cdot)$ the corresponding cdf. For the system to fail before t the sojourn time in state 0 must be obviously less than t (see (12.14)) and so if we want to use importance sampling in order to estimate $\bar{R}(t)$, then the optimal change of measure, in a way analogous to relation (12.6), will be given by

$$f_X^*(x) = \frac{f_X(x)}{F_0(t)} \mathbf{1}_{[0,t]}(x) = \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda t}} \mathbf{1}_{[0,t]}(x).$$

The optimal change of measure, which is conditional on the fact that the component will finally fail before t, can help us to develop efficient changes of measure for the unreliability estimation of a Markovian system, as shown below.

12.9.2 General case

Consider now the Markov chain defined in Section 12.6. Our goal is to use importance sampling in order to estimate the quantity $\xi = \mathbf{E}_Q[h(X)]$, where $h(\cdot)$: $E \to \mathbf{R}$ and $\mathbf{E}_Q$ implies that the expected value is taken with respect to the measure induced by the generator Q (supposed conservative). Suppose also that the function may be expressed as $h(Y_0, V_0, \ldots, Y_n, V_n)$, with Y_i the state visited at the i-th jump of the process and V_i the corresponding sojourn time. Moreover, conditionally on $\{Y_n : n \geq 0\}$, $\{V_n : n \geq 0\}$ are i.i.d. exponential r.v. with parameter $\frac{1}{-Q(y_n, y_n)} = \frac{1}{q(y_n)}$.

For the unreliability estimation of the system we can change both the transition matrix as well as the laws governing the sojourn times of the system in different states. In the first case, we can use a "failure biasing" method that will accelerate the failures of individual components, while for the sojourn times we can use the change of measure corresponding to the optimal change of measure for the one component system. Suppose that at time s the system is in a state $y \in U$. For the system to fail before time t, it must make at least one transition in the interval (s, t), meaning that the sojourn time at state y must be less than $t - s$. Using the conditional laws, we are sure that the system will finally fail before the observation period expires [11], and so the event $\{T_F \leq t\}$, representing the failure of the system before t, will be realized at every time during the simulation.

The realizations of the chain will have the form $(y_0, v_0, \ldots, y_n, v_n)$, with $y_0, \ldots, y_{n-1} \in U$, $y_n \in F$ and $v_0, \ldots, v_{n-1}, v_n$ such that $v_i > 0$ for $i = 0, \ldots, n$ and $v_0 + \ldots + v_{n-1} \leq t$. The set Γ represents the set of all possible transitions, while

$$\begin{aligned} \Delta = \{&(y_0, v_0, y_1, v_1, \ldots, y_n, v_n) : n \geq 1, y_0 = 0, y_n \in F, y_i \notin F, \\ &1 \leq i < n, (y_i, y_{i+1}) \in \Gamma \text{ for } 0 \leq i < n, v_0 + v_1 + \ldots + v_{n-1} \leq t \\ &\text{and } v_i > 0 \text{ for } i = 0, \ldots, n\} \end{aligned}$$

will represent the set of paths for which the event $\{T_F \leq t\}$ has been realized. The likelihood of such a path will be :

$$\mu(y_0)\prod_{i=0}^{n-1}P(y_i,y_{i+1})\prod_{i=0}^{n}q(y_i)e^{-q(y_i)v_i}.$$

Using an importance sampling estimator corresponding to the new generator Q', verifying

$$Q(i,j)\neq 0 \text{ implying that } Q'(i,j)\neq 0,$$

the corresponding likelihood ratio will be [13]

$$L(\cdot)=\prod_{i=0}^{n-1}\frac{P(y_i,y_{i+1})}{P'(y_i,y_{i+1})}\prod_{i=0}^{n}\frac{q(y_i)e^{-q(y_i)v_i}}{q'(y_i)e^{-q'(y_i)v_i}},$$

where we suppose that the initial laws were the same.
The variance of this importance sampling estimator for $\xi=\mathbf{P}(T_F\leq t)$, will be given by

$$\begin{aligned}Var_{Q'}[\mathbf{1}_{\{T_F\leq t\}}L(\cdot)] &= \mathbf{E}_{Q'}[\mathbf{1}^2_{\{T_F\leq t\}}L^2(\cdot)]-\left(\mathbf{E}_{Q'}[\mathbf{1}_{\{T_F\leq t\}}L(\cdot)]\right)^2\\ &= \mathbf{E}_{Q'}[\mathbf{1}_{\{T_F\leq t\}}L^2(\cdot)]-\xi^2,\end{aligned}$$

and we are rather interested in its first part

$$\mathbf{E}_{Q'}[\mathbf{1}_{\{T_F\leq t\}}L^2(\cdot)]=\mathbf{E}_{Q}[\mathbf{1}_{\{T_F\leq t\}}L(\cdot)]$$

$$=\sum_{(y_0,v_0,y_1,\ldots,y_n,v_n)\in\Delta}\prod_{i=0}^{n-1}\frac{P^2(y_i,y_{i+1})}{P'(y_i,y_{i+1})}\cdot$$

$$\cdot\int_0^t\int_0^{t-v_0}\cdots\int_0^{t-v_0-\ldots-v_{n-1}}\int_0^{\infty}\frac{q_{y_0}e^{-q_{y_0}v_0}}{\frac{q_{y_0}e^{-q_{y_0}v_0}}{1-e^{-q_{y_0}t}}}\cdots\frac{q_{y_{n-1}}e^{-q_{y_{n-1}}v_{n-1}}}{\frac{q_{y_{n-1}}e^{-q_{y_{n-1}}v_{n-1}}}{1-e^{-q_{y_{n-1}}(t-v_0-\ldots-v_{n-1})}}}$$

$$\cdot\, q_{y_0}e^{-q_{y_0}v_0}q_{y_1}e^{-q_{y_1}v_1}\ldots q_{y_{n-1}}e^{-q_{y_{n-1}}v_{n-1}}dv_0dv_1\ldots dv_{n-1}=$$

$$=\sum_{(y_0,v_0,y_1,\ldots,y_n,v_n)\in\Delta}\prod_{i=0}^{n-1}\frac{P^2(y_i,y_{i+1})}{P'(y_i,y_{i+1})}\cdot$$

$$\cdot\int_0^t\cdots\int_0^{t-v_0-\ldots-v_{n-1}}\int_0^{\infty}(1-e^{-q_{y_0}t})\ldots(1-e^{-q_{y_{n-1}}(t-v_0-\ldots-v_{n-1})})$$

$$\cdot\, q_{y_0}e^{-q_{y_0}v_0}\ldots q_{y_n}e^{-\lambda_{y_n}v_n}dv_0\ldots dv_n.$$

Note that $(1-e^{-q_{y_0}t})<1,\ldots,(1-e^{-q_{y_{n-1}}(t-v_0-\ldots-v_{n-1})})<1$. Then:

$$B=(1-e^{-q_{y_0}t})(1-e^{-q_{y_1}(t-v_0)})\ldots(1-e^{-q_{y_{n-1}}(t-v_0-\ldots-v_{n-1})})<1.$$

Then, given the fact that $P'=P$, we obtain

$$\mathbf{E}_{Q'}[\mathbf{1}^2_{\{T_F\leq t\}}L^2(\cdot)]<\sum_{(y_0,v_0,y_1,\ldots,y_n,v_n)\in\Delta}\prod_{i=0}^{n-1}P(y_i,y_{i+1})$$

$$\cdot\int_0^t\int_0^{t-v_0}\cdots\int_0^{t-\ldots-v_{n-1}}\int_0^{\infty}q_{y_0}e^{-q_{y_0}v_0}\ldots q_{y_n}e^{-q_{y_{n-1}}v_n}dv_0dv_1\ldots dv_n$$

$$=\mathbf{E}_Q[\mathbf{1}_{\{T_F\leq t\}}]=\xi,$$

which means that

$$Var_{Q'}[\mathbf{1}_{\{T_F\leq t\}}L(\cdot)]<Var_Q[\mathbf{1}_{\{T_F\leq t\}}]=\xi-\xi^2,$$

and so the variance of the new estimator is less than the variance of the direct estimator.

12.9.3 Example

Consider the state diagram given in figure 12.6, representing a system consisting of two types of components, with two components of each type [32]. The state space $E = \{0, \ldots, 8\}$ contains 9 states and the system is considered operational if at least one component of each type is operational, while otherwise it is considered failed ($F = \{4, \ldots, 8\}$). Failure times follow an exponential distribution with parameter $\lambda = 10^{-4}$, while the corresponding repair rate is $\mu = 1.0$.

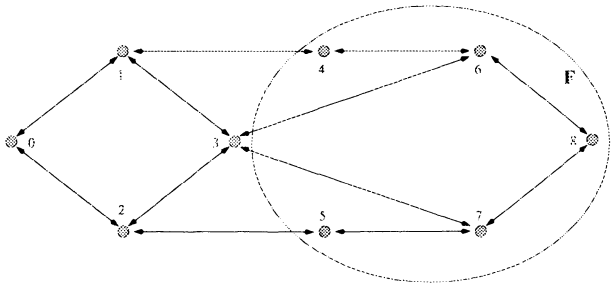


Fig. 12.6. System's state diagram

In order to accelerate the failures of individual components a failure biasing method has to be used. This will make the system move towards the failed set of states with a relatively high probability. Moreover, in order to estimate its unreliability we need to observe a system failure before the observation period, and thus the conditioning method can be appropriate. The associated estimation results are illustrated in figure 12.7 together with the exact value of the system's unreliability obtained by an analytical method. Note that we used a time step of 0.2 and we observe the system until time $t = 10$. Another approach would be to use the conditioning method only for the holding times in state 0, since the time spent in this state represents the majority of the time spent by the process. This method is known as *forcing* and it was introduced by Lewis and Böhn in [27]. Another alternative could be to use the condition that the system will fail before time T, with $T > t$ the actual observation period of the system (in the example we have used $T = 20$ time units). Note also that both the methods give very good results for small values of time. Similar arguments can be used for the estimation of measures such as the expected interval availability and the steady state availability [16].

12.10 Analytical-Statistical Methods

The idea behind an analytical-statistical method is efficient combination of Monte Carlo simulation and analytical solutions. To estimate a given quantity of interest using this method, simulation and analytical solutions have to be used interchangeably.

The first analytical-statistical method is due to Kovalenko, but in recent years a lot of interesting results have been obtained for a variety of problems

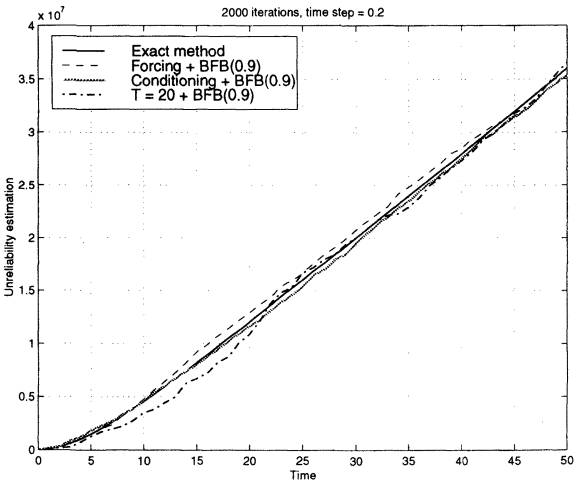


Fig. 12.7. Estimation results for 2000 iterations

(see Chapter 6 and 7 of [23] and references therein). Thus, quick simulation algorithms have been constructed for efficient estimation of the unreliability and unavailability of a repairable system, the evaluation of non-stationary state probabilities of alternating renewal processes, the evaluation of the sojourn times distribution of Markov processes, etc.

What it is interesting to note is that using an analytical-statistical method we do not modify the system's underlying probabilistic structure. The variance reduction obtained stems from the fact that a significant part of the algorithm uses analytical solutions to calculate rare event probabilities. It is therefore different from importance sampling where the system is simulated using a new probability measure.

Below, we briefly review the basic principles of the analytical-statistical method for the unreliability estimation of k-out-of-n structures, which is the most suitable for the illustration of this quick simulation method.

Simulation of k-out-of-n structures

Consider a system having n independent components and let $F_i(\cdot)$, $G_i(\cdot)$ $(i = 1, \ldots, n)$ represent the cumulative distribution functions associated with the failure and repair of individual components respectively. A system is defined as a k-out-of-n : F system, if it is failed when at least k of its n components are failed.

Suppose that we are interested in estimating the unreliability of the system at time t, which stands for the probability that the system fails before t, where $[0, t]$ is our observation period. This is denoted by $\bar{R}(t)$. Let T_F be the hitting time representing the entrance of the system to the failed subset of states, denoted by F and containing states with at least k components failed. Then, the unreliability of the system at time t may be written as $\bar{R}(t) = \mathbf{P}(T_F \leq t)$. Let T_i also be a sequence of positive random variables $0 < T_1 < T_2 \ldots < T_k$, where T_i stands for the time the system has i failed components for the first time. Since for the system to fail by time t at least k components have to be failed by the same instant, then by defining $B_i(t) = \{T_i \leq t\}$, for $1 \leq i \leq k$, we can write

$$\begin{aligned}\bar{R}(t) &= \mathbf{P}(T_F \leq t) = \mathbf{P}[B_k(t)] = \mathbf{P}[\cap_{i=1}^{k} B_i(t)] \\ &= \mathbf{P}[B_k(t)|B_{k-1}(t)]\mathbf{P}[B_{k-1}(t)|B_{k-2}(t)]\ldots\mathbf{P}[B_1(t)]. \end{aligned} \tag{12.15}$$

The analytical-statistical method provides an efficient way of estimating the conditional probabilities in (12.15) and thus estimating the unreliability of the system at time t. The details of the exact calculation of these probabilities are given in [23].

No doubt one of the basic advantage and the main interest of the method is that it can also be used in the case of non-Markovian systems. In such systems, we get rid of the assumption on the exponentiality of the failures and repairs of individual components. Moreover, under some reasonable conditions for the distribution functions $F_i(\cdot)$, $i = 1, \ldots, n$, the analytical-statistical method has the bounded relative error property (see again [23]), which is a guarantee for the efficiency of the method.

Consider for example the system illustrated in figure 12.8, where we sup-

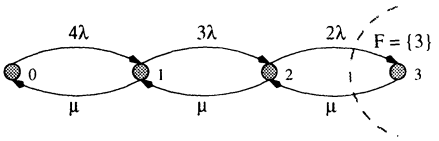


Fig. 12.8. State diagram for a 2-out-of-4 system

pose that the failure and repair distributions of individual components are exponential with parameters $\lambda_i = \lambda = 10^{-3}$, $\mu_i = \mu = 1$, $i = 1, \ldots, 4$ and the system has only one repairman. The estimation results for this small system are given in figure 12.9, where the unreliability of the system together with the associated 95% confidence interval is illustrated. The exact value is also represented in the same figure for comparison purposes. Note that the estimates obtained using the analytical-statistical method are quite stable even for a small number of iterations.

Similar remarks can be made for the estimation of the unreliability of a 2-out-of-3 system with $\lambda_1 = 10^{-3}$, $\lambda_2 = 10^{-4}$, $\lambda_3 = 10^{-5}$, $\mu_i = 1$, $i = 1, 2, 3$. These results are illustrated in figure 12.10. Note that we are estimating a probability of the order of 10^{-8}, using only 10^4 iterations, meaning that a large variance reduction is obtained.

Consider now a consecutive k-out-of-$n : F$ system, denoted by $C(k, n : F)$. This system is consisting of n statistically independent components, connected in a linear fashion and it is considered failed if at least k consecutive components are failed. Systems of this kind have become very popular in recent years, since they are associated with many important applications (see [5] for a survey in this topic).

A modification of the previous algorithm concerning simple k-out-of-n systems can also be used in this case. Let $A_k(t)$ represent the event $\{k$ consecutive components are failed by time $t\}$. Then the unreliability of a $C(k, n : F)$ system at time t, denoted by $\bar{R}_c(t)$, in a way similar to equation (12.15), can be estimated by

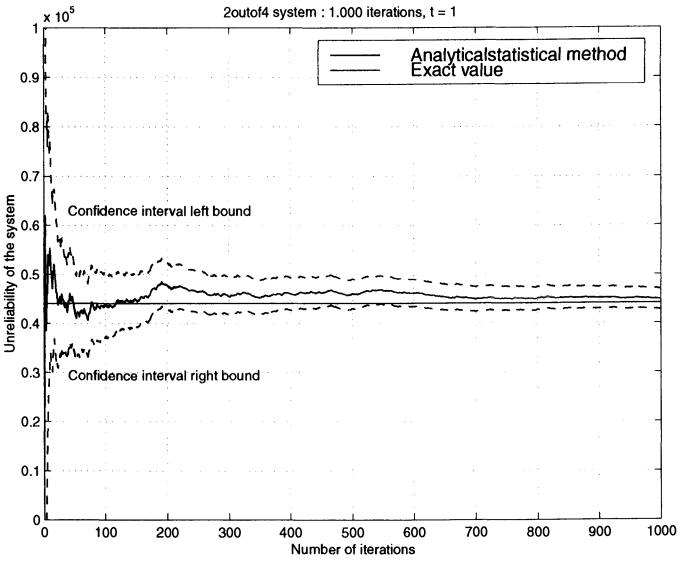


Fig. 12.9. Estimation results for the 2-out-of-4 system

$$\begin{aligned}\bar{R}_c(t) &= \mathbf{P}[A_k(t)] = \mathbf{P}[A_k(t) \cap B_k(t)] \\ &= \mathbf{P}\left[A_k(t) \cap \left\{\cap_{i=1}^{k} B_i(t)\right\}\right] \\ &= \mathbf{P}[A_k(t)|B_k(t)]\mathbf{P}[B_k(t)|B_{k-1}(t)] \ldots \mathbf{P}[B_1(t)]. \end{aligned} \tag{12.16}$$

The analytical-statistical method can be used to estimate the conditional probabilities $\mathbf{P}[B_i(t)|B_{i-1}(t)]$ for $i = 2, \ldots, k$, while direct simulation can be effectively used to estimate $\mathbf{P}[A_k(t)|B_k(t)]$, since the system will already have k components failed. Moreover, a splitting technique can be used to obtain better estimates of the conditional probability $\mathbf{P}[A_k(t)|B_k(t)]$. Thus, at any time when the system has k components failed before time t, we perform repeated simulations to check whether the system will finally fail before the time horizon. The average value of the results of these simulations will then be our estimate for the probability $\mathbf{P}[A_k(t)|B_k(t)]$. Given that all the other conditional probabilities in equation (12.16) have already been calculated using the analytical-statistical method, we can thus obtain one estimate of the system's unreliability. By repeating the same procedure for a number of times, depending of course on the chosen confidence interval, an estimate of $\bar{R}_c(t)$ can be obtained.

12.11 Concluding Remarks

In this paper, we have discussed some quick simulation methods currently used in reliability systems.

Simulation is a good alternative for systems having a large state space and/or many interdependencies between individual components, since in such

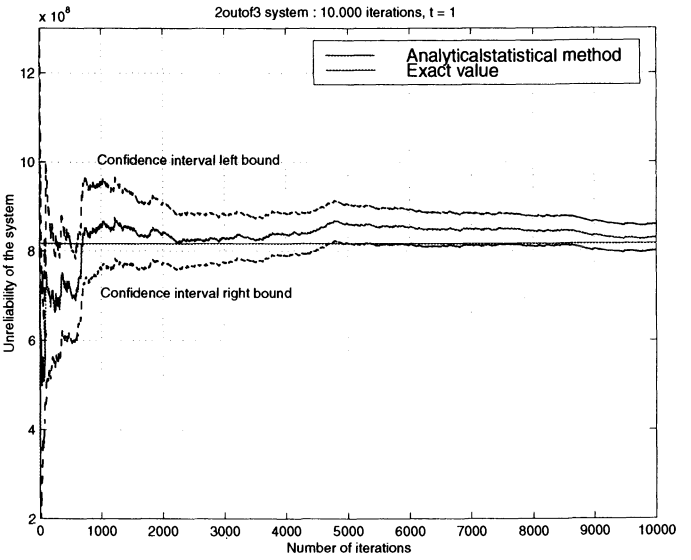


Fig. 12.10. Estimation results for the 2-out-of-3 system

cases the existing analytical solutions are not efficient. However, direct simulation suffers from the drawback that the effort needed to estimate up to a certain level the probability of an event increases as the event becomes rarer and rarer, and thus it may be inappropriate in such settings. Therefore, modifications of the direct simulation scheme have to be carried out.

The associated methods, called variance reduction methods, reduce the variance of the corresponding estimator, thus accelerating the estimation procedure. Importance sampling and analytical-statistical methods are quite appropriate for quick simulation of reliability systems, where the rare event under question is usually associated with the failure of the system or the entrance of the system to a subset of states and it may happen when only a few events, each of which are rare, happen.

However, we have to note the principal difference between importance sampling and analytical statistical methods. In the first case, we change the probabilistic dynamics of the original system S and simulate a new one, say S_1. The goal of this change of measure is to make the rare event under consideration more likely to occur. A compensatory factor called the likelihood ratio is introduced into the estimator in order to eliminate the bias resulting from this change of measure. Unbiased estimates of the parameter are thus obtained. In the second case, the system is simulated using the original failure and repair distributions for the individual components, and the variance reduction stems from the fact that rare event probabilities are explicitly calculated using analytical expressions.

Furthermore, an analytical-statistical method does not make many restrictive assumptions for the system model, thus enabling simulation of systems that may be non-Markovian in nature. On the other hand, importance sampling is principally used in Markovian systems, where the regenerative structure of the system can be exploited.

References

1. Bobbio, A. and Trivedi, K. S. (1986), "An aggregation technique for the transient analysis of stiff Markov chains," *IEEE Transactions on Computers*, **C-35**, 1291-1298
2. Bratley, P., Fox, B.L. and Schrage, L.E. (1987), A Guide to Simulation, Springer-Verlag, Berlin
3. Carrasco, J. A. (1991), "Failure distance-based simulation of repairable fault-tolerant systems," *Proceedings of the Fifth International Conference on Modeling Techniques and Tools for Computer Performance Evaluation*, 337-351
4. Carrasco, J. A. (1991), "Efficient transient simulation of failure/repair Markovian models," *Proceedings of the Tenth Symposium on Reliable and Distributed Computing*, 152-161, IEEE Computer Society Press, Pisa, Italy
5. Chao, M. T., Fu, J. C. and Koutras, M. V. (1995), "Survey of reliability studies of consecutive-k-out-of-n:F & related systems," *IEEE Transactions on Reliability*, **44**, 120-127
6. Cottrell, M., Fort, J. C. and Malgouyres, G. (1983), "Large deviations and rare events in the study of stochastic algorithms," *IEEE Transactions on Automatic Control*, **AC-28**, 907-920
7. Crane, M. A. and Iglehart, D. L. (1975), "Simulating stable stochastic systems III, regenerative processes and discrete event simulation," *Operations Research*, **23**, 33-45
8. Fishman, G. S. (1996), Monte-Carlo : Concepts Algorithms and Applications, Springer Series in Operations Research, Springer-Verlag, New York
9. Fox, B. L. and Glynn, P. W. (1986), "Discrete time conversion for simulating semi-Markov processes," *Operations Research Letters*, **5**, 191-196
10. Juneja, S. K. (1993), Rare Events Simulation of Stochastic Systems, PhD Thesis, Stanford University
11. Gertsbakh, I. and Spungin, I. (1999), Product-Type Estimator of Convolutions, in Semi-Markov Models and Applications (Janssen, J., Limnios, N. eds.). 201-206, Kluwer Dordrecht, The Netherlands
12. Glynn, P. W., Heidelberger, P., Nicola, V. F. and Shahabuddin, P. (1993), "Efficient estimation of the mean time between failures in non-regenerative dependability models," *Proceedings of the 1993 Winter Simulation Conference*, IEEE Computer Society Press, 311-316
13. Glynn, P. W. and Iglehart, D. L. (1989), "Importance sampling for stochastic simulations," *Management Science*, **35**, 1367-1392
14. Goyal, A. and Lavenberg, S. S. (1987), "Modeling and analysis of computer system availability," *IBM Journal of Research and Development*, **31**, 651-664
15. Goyal, A., Heidelberger, P. and Shahabuddin, P. (1987), "Measure specific dynamic importance sampling for availability simulations," *Proceedings of*

the 1987 Winter Simulation Conference , IEEE Computer Society Press, 351-357

16. Goyal, A., Shahabuddin, P., Heidelberger, P., Nicola, V. F. and Glynn, P. W. (1992), "A unified framework for simulating Markovian models of highly reliable systems," *IEEE Transactions on Computers*, **41**, 36-51
17. Goyal, A., Lavenberg, S. S. and Trivedi, K. S. (1987), "Probabilistic modeling of computer system availability," *Annals of Operations Research*, **8**, 285-306
18. Hammersley, J. M. and Handscomb, D. C. (1964), Monte Carlo Methods. Methuen, London
19. Heidelberger, P. (1995), "Fast simulation of rare events in queueing and reliability models," *ACM Transactions on Modeling and Computer Simulation*, **5**, 43-85
20. Heidelberger, P. (1988), "Discrete event simulations and parallel processing, statistical properties," *SIAM Journal on Scientific and Statistical Computing*, **9**, 1114-1132
21. Heidelberger, P., Nicola, V. F. and Shahabuddin, P. (1992), "Simultaneous and efficient simulation of highly dependable systems with different underlying distributions," *Proceedings of the 1992 Winter Simulation Conference*, IEEE Computer Society Press, 458-465
22. Heidelberger, P, Shahabuddin, P. and Nicola, V. F. (1994), "Bounded relative error in estimating transient measures of highly dependable non-Markovian systems," *ACM Transactions on Modeling and Computer Simulation*, **4**, 137-164
23. Kovalenko, I., Kuznetzov, N. Y. and Pegg, P. A. (1997), Mathematical Theory of Reliability of Time Dependent Systems with Practical Applications. John Wiley & Sons,
24. Kuruganti, I. and Strickland, S. G. "Optimal importance sampling for Markovian systems," *Proceedings of the 1995 IEEE Systems, Man & Cybernetics Conference*
25. Kuruganti, I. and Strickland, S. G. (1997), "Optimal importance sampling for Markovian systems with applications to tandem queues," *Mathematics and Computers in Simulation*, **44**, 61-80
26. Kuruganti, I. and Strickland, S. G. "Importance sampling for Markov chains: computing variance and determining optimal measures," *Proceedings of the 1996 Winter Simulation Conference*
27. Lewis, E. E. and Bohm, F. (1984), "Monte Carlo simulation of Markov unreliability Models," *Nuclear Engineering and Design*, **77**, 49-62
28. Nakayama, M. K. (1995), "Asymptotics for likelihood ratio derivative estimators in simulations of highly reliable Markovian systems," *Management Science*, **41**, 524-554
29. Nakayama, M. K. (1994), "A characterization of the simple failure biasing method for simulations of highly reliable Markovian systems," *ACM Transactions on Modeling and Computer Simulation*, **4**, 52-88
30. Nakayama, M. K. (1996), "General conditions for bounded relative error in simulations of highly reliable Markovian systems," *Advances in Applied Probability*, **28**, 687-727
31. Nakayama, M. K., Goyal, A. and Glynn P. W. (1994), "Likelihood ratio sensitivity analysis for Markovian models of highly dependable systems," *Operations Research*, **42**, 137-157
32. Nicola, V. F., Nakayama, M. K., Heidelberger, P. and Goyal, A. (1993), "Fast simulation of highly dependable systems with general failure and repair processes," *IEEE Transactions on Computers*, **42**, 1440-1452

33. Ripley, B. D. (1987), Stochastic Simulation. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, NY
34. Ross, S. M. (1990), A Course in Simulation. Maxwell MacMillan International Editions, NY
35. Shahabuddin, P. (1990), Simulation and Analysis of Highly Reliable Systems, Ph.D. Thesis, Stanford University, California
36. Shahabuddin, P. (1994), "Importance sampling for the simulation of highly reliable Markovian systems," *Management Science*, **40**, 333-352
37. Shahabuddin P. (1994), "Fast transient simulation of Markovian models of highly dependable systems," *Performance Evaluation*, **20**, 267-286
38. Shahabuddin, P. and Nakayama, M. K. (1993), "Estimation of reliability and its derivatives for large time horizons in Markovian systems," *Proceedings of the 1993 Winter Simulation Conference*, IEEE Computer Society Press, 422-429
39. Shahabuddin, P., Nicola, V. F., Heidelberger, P., Goyal, A. and Glynn, P. W. (1988), "Variance reduction in mean time to failure simulations," *Proceedings of the 1998 Winter Simulation Conference*, IEEE Computer Society Press, 491-499
40. Shahabuddin, P. (1993), "Efficient estimation of the mean time between failures in non regenerative dependability models," *Proceedings of the 1993 Winter Simulation Conference*, IEEE Computer Society Press, 311-316
41. Strickland, S. G. (1993), "Optimal importance sampling for quick simulation of highly reliable Markovian systems," *Proceedings of the 1993 Winter Simulation Conference*, IEEE Computer Society Press, 437-444

Index