OPTIMAL POLICIES IN RELIABILITY MODELLING OF SYSTEMS SUBJECT TO SPORADIC SHOCKS AND CONTINUOUS HEALING

by

Debolina Chatterjee

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Mathematical Sciences

Indianapolis, Indiana

December 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Jyotirmoy Sarkar, Chair

Department of Mathematical Sciences

Dr. Benzion Boukai

Department of Mathematical Sciences

Dr. Fang Li

Department of Mathematical Sciences

Dr. Honglang Wang

Department of Mathematical Sciences

Approved by:

Dr. Jeffrey Watt

I dedicate this thesis to Baba for his immeasurable sacrifices, to Maa for her purest love, and to Pratim for his unfeigned support.

ACKNOWLEDGMENTS

This dissertation is a result of the love and support of many important people in my professional and personal life, and I would like to take this opportunity to thank everyone who has been so kind and helpful throughout this journey.

First and foremost, I thank my advisor Dr. Jyotirmoy Sarkar for choosing to guide me. I could not have asked for a better mentor and guide. I thank him for his warmth, care and attention to detail and for being patient and kind with me even at times when I was not my most productive self. I am thankful to the other members of my dissertation committee, Drs. Benzion Boukai, Fang Li, and Honglang Wang for giving me insightful comments and suggestions that helped improve the quality of this dissertation. Dr. Boukai has not only taught wonderful courses like sequential analysis and applied Bayesian analysis with care, but has also been a kind mentor who nurtured my teaching interests by providing me with ample opportunities to strengthen my teaching capability, has always been considerate towards me, and helped me with professional connections. I thank Dr. Honglang Wang for teaching statistical computing so nicely from where I learned useful computing skills and for organizing good talks at the statistics seminar due to which I was exposed to the contemporary research.

I also thank Drs. Hanxiang Peng, Fei Tan, Fang Li, Pavel Bleher, and Jared Barber for teaching useful statistics and mathematics courses, which were very good learning experiences. Words cannot express my gratitude to Dr. Roland Roeder for his immense help toward me, a non-pure math background student, in cracking the qualifying exam that granted my candidature in this doctoral study.

I sincerely thank Dr. Evgeny Mukhin, the then graduate director; the Chair of the math department, Dr. Jeffrey Watt; and other members of the graduate students' selection committee for Fall 2017 for giving me the opportunity to study here and for providing me with full financial support throughout my doctoral education, without which carrying out education in the USA would not have been possible for me. Furthermore, I am extremely grateful to the entire Department of Mathematical Sciences, IUPUI, for providing a respectful, inclusive, and encouraging environment. In addition to the faculty in the department, I thank

Tina Carmichael and Cindy Ritchie for being kind and fun and Virginia Ranly for taking care of various paperwork for me.

I am very lucky to have received the care and hospitality of Dr. Sarkar and his wife, Falguni Sarkar, from the very first day in this country. Mrs. Sarkar provided me with many basic necessities during my initial days and all throughout which I very much needed as an international student; and the various amazing meals that she invited me to with genuine care are among my most loved experiences. I spent some of my best days with many of my good friends in Indianapolis, especially with Oindrila Bhattacharya, who was the first friend I made here, whom I thank for various meaningful conversations and for making it easier to glide through lonely days away from home.

Looking back to the very beginning of my journey, I thank my undergraduate and graduate professors back in India for teaching me statistics. I am proud to have had the opportunity to earn my Master's degree from the Department of Statistics at the University of Calcutta, which was the nursery for statistics education in India, and I am indebted to all the faculty of this department, especially Drs. Aditya Chattopadhyay, Nripes Kumar Mandal, Manisha Pal, Asish Kumar Chattopadhyay, Bhashwati Ganguli, Gaurangadeb Chattopadhyay and Rahul Bhattacharya for the amazing learning environment and encouragement to pursue research.

My family has been my support and strength throughout my life. I thank my father Debasis Chatterjee for his sacrifices and choices to make my education a priority. He inculcated curiosity and discipline in me, was my first mathematics teacher, and his encouragement helped grow the interest of logical thinking in me. I thank my mother Anupama Chatterjee, who raised me with honor and rich values almost alone. She has inspired me to stand up for myself and become an independent self-made woman. I consider myself lucky to have met my husband Pratim Guha Niyogi, who has been a true friend, philosopher and guide and from whom I learned humility and perseverance. I thank him for believing in me when I couldn't recognize my own abilities. I also take this moment to remember my grandmother Late Ila Chatterjee and grandfather Late Haraprasad Kanjilal, whom I lost during this journey, but I know they would be proud of me from wherever they are. I thank all my well-wishers for keeping me in their thoughts. Finally, I thank the Almighty for giving me the perfect blend of highs and lows that I needed and for keeping me grounded throughout the journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbb{R}$	Set of real numbers
$\mathbb{N}$	Set of natural numbers
$\forall$	For all
$\in$	Belongs to
$\rightarrow$	Approaches to
∞	Infinity (or extremely large value)
$\sim$	Distributed as
$\mid$	Given that
$<<$	Much less than
$exp\{x\}$	The exponential function where the argument x is written as an exponent
$\mathbf{1}\{x \in A\}$	Indicator function that takes value 1 when $x \in A$ and zero otherwise
$\boldsymbol{\pi}_j$	Stationary probability in State j
$\mu_{(i,j)}$	Expected sojourn time in State (i, j)
θ_j	Limiting probability for the stochastic process to be found in State j
A_{av}	Limiting average availability
μ	Mean lifetime
ν	Mean repair time
r	Number of repair facilities
s	Number of spare units
Δ	Increments of time at which a system is observed
δ	Duration of time elapsing after arrival of an external shock
k	The prespecified number of positive interventions required to induce self-healing effect
λ	Proportionality constant which is used to find the distribution of Stage 1 and the system lifetime from adjusted convolution of inter-arrival cumulative distribution functions
σ_F	Standard deviation of the inter-arrival time distribution of valid shocks
σ_G	Standard deviation of the inter-arrival time distribution of positive interventions

c_0	Cost of initial installation of a system
c_{p_1}	Cost of maintenance in Stage 1
c_{p_2}	Cost of maintenance in Stage 2
c_f	Cost of failure replacement
c_I	Cost of inspection
C	Overall cost
CT	Cycle time
t_u	The u-th inspection interval
v_u	Time at the u-th inspection
C_u	Overall cost after the u-th inspection
CT_u	Cycle time at the u-th inspection
$D(t)$	Cumulative damage to the system at time t
$B(t)$	Time dependent boundary threshold, a system fails as soon as cumulative damage $D(t)$ crosses $B(t)$
$N(t)$	Number of shocks to the system at time t
κ	The rate of self-healing
d	The distance by which when $D(t)$ comes closer to $B(t)$, an alarm goes off signalling that the system is at risk
γ	Survival probabilities
$\Gamma(s)$	The incomplete gamma function for parameter s
$t^*_\gamma(d)$	Survival percentiles after an alarm is set off indicating that the system is at risk
c_{op}	Cost due to operation of the system
c_{rev}	Revenue earned by the system
I_f	Indicator whether a system is healable or non-healable

ABBREVIATIONS

CDF	Cumulative Distribution Function
PDF	Probability Density Function
PMF	Probability Mass Function
IID	Independent and Identically Distributed
HPP	Homogeneous Poisson Process
MSUT	Mean System Up Time
MSDT	Mean System Down Time
CPUT	Cost Per Unit Time
VS	Valid Shock
PI	Positive Intervention
PM	Preventive Maintenance
CM	Corrective Maintenance
FT	Failure Time
RL	Residual Lifetime

ABSTRACT

Recent years have seen a growth in research on system reliability and maintenance. Various studies in the scientific fields of reliability engineering, quality and productivity analyses, risk assessment, software reliability, and probabilistic machine learning are being undertaken in the present era. The dependency of human life on technology has made it more important to maintain such systems and maximize their potential. In this dissertation, some methodologies are presented that maximize certain measures of system reliability, explain the underlying stochastic behavior of certain systems, and prevent the risk of system failure.

An overview of the dissertation is provided in Chapter 1, where we briefly discuss some useful definitions and concepts in probability theory and stochastic processes and present some mathematical results required in later chapters. Thereafter, we present the motivation and outline of each subsequent chapter.

In Chapter 2 we compute the limiting average availability of a one-unit repairable system subject to repair facilities and spare units. Formulas for finding the limiting average availability of a repairable system exist only for some special cases: (1) either the lifetime or the repair-time is exponential; or (2) there is one spare unit and one repair facility. In contrast, we consider a more general setting involving several spare units and several repair facilities; and we allow arbitrary life- and repair-time distributions. Under periodic monitoring, which essentially discretizes the time variable, we compute the limiting average availability. The discretization approach closely approximates the existing results in the special cases; and demonstrates as anticipated that the limiting average availability increases with additional spare unit and/or repair facility.

In Chapter 3, the system experiences two types of sporadic impact: valid shocks that cause damage instantaneously and positive interventions that induce partial healing. Whereas each shock inflicts a fixed magnitude of damage, the accumulated effect of k positive interventions nullifies the damaging effect of one shock. The system is said to be in Stage 1, when it can possibly heal, until the net count of impacts (valid shocks registered *minus* valid shocks nullified) reaches a threshold m_1. The system then enters Stage 2, where no further healing is possible. The system fails when the net count of valid shocks reaches another threshold

m_2 ($> m_1$). The inter-arrival times between successive valid shocks and those between successive positive interventions are independent and follow *arbitrary* distributions. Thus, we remove the restrictive assumption of an exponential distribution, often found in the literature. We find the distributions of the sojourn time in Stage 1 and the failure time of the system. Finally, we find the optimal values of the choice variables that minimize the expected maintenance cost per unit time for three different maintenance policies.

In Chapter 4 the above defined Stage 1 is further subdivided into two parts: In the early part, called Stage 1A, healing happens faster than in the later stage, called Stage 1B. The system stays in Stage 1A until the net count of impacts reaches a predetermined threshold m_A; then the system enters Stage 1B and stays there until the net count reaches another predetermined threshold m_1 ($> m_A$). Subsequently, the system enters Stage 2 where it can no longer heal. The system fails when the net count of valid shocks reaches another predetermined higher threshold m_2 ($> m_1$). All other assumptions are the same as those in Chapter 3. We calculate the percentage improvement in the lifetime of the system due to the subdivision of Stage 1. Finally, we make optimal choices to minimize the expected maintenance cost per unit time for two maintenance policies.

Next, we eliminate the restrictive assumption that all valid shocks and all positive interventions have equal magnitude, and the boundary threshold is a preset constant value. In Chapter 5, we study a system that experiences damaging external shocks of random magnitude at stochastic intervals, continuous degradation, and self-healing. The system fails if cumulative damage exceeds a time-dependent threshold. We develop a preventive maintenance policy to replace the system such that its lifetime is utilized prudently. Further, we consider three variations on the healing pattern: (1) shocks heal for a fixed finite duration τ; (2) a fixed proportion of shocks are non-healable (that is, $\tau = 0$); (3) there are two types of shocks—self healable shocks heal for a finite duration, and non-healable shocks. We implement a proposed preventive maintenance policy and compare the optimal replacement times in these new cases with those in the original case, where all shocks heal indefinitely.

Finally, in Chapter 6, we present a summary of the dissertation with conclusions and future research potential.

1. INTRODUCTION

Reliability theory has become an important field of study in the last few decades because of advancement of technology and the dependence of human life on it. There is an ever increasing need for reliable systems and to study their functioning. By definition, a set of things that work together as part of a mechanism or an interconnected network is called a *system*. It can be biological, environmental, economic, or industrial. Research on reliable systems often spans across various disciplines including, but not limited to, engineering, statistics, applied probability, economics, medical science, warfare, actuaries, and survival analysis. In this dissertation, we look at industrial systems and explore some concepts in industrial statistics. A major part of reliability analysis is studying life distributions and system failure. Great emphasis is placed on studies of the system lifetime and on the measures taken to maximize benefits from a system while monitoring the system's condition. In the next three sections of this chapter, we discuss some useful results in probability theory and stochastic processes (Section 1.1); some useful mathematical results and concepts (Section 1.2); and some concepts in reliability theory (Section 1.3), which lay the foundation for the upcoming chapters and will be referred to time and again. Subsequently, in Section 1.4 we briefly discuss the main essence of each chapter and the motivation behind the current research.

1.1 Concepts in probability theory and stochastic processes

Let us discuss some basic concepts in probability theory which are necessary in the upcoming chapters. The definitions in this section are taken mainly from Ross *et al.* (1996) and Ross (2014).

For a discrete random variable X, the probability mass function (PMF) is defined as $p(x) = P(X = x)$, where $p(x) \geq 0$ for at most a countable number of values of x such that for all values $x_1, x_2, x_3, \cdots$ assumed by X, $\sum_{i=1}^{\infty} p(x_i) = 1$. The cumulative distribution function (CDF) is defined as $F(x) = \sum_{i|x_i \leq x} p(x_i)$.

For a continuous random variable X, the probability density function (PDF) is a non-negative function $f(x)$ such that $P(X \in B) = \int_B f(x)dx$, where $x \in \mathbb{R}$ and B is a set of

real numbers. The set $\{x : f(x) > 0\}$ is known as the support of the distribution. In particular, $\int_{-\infty}^{\infty} f(x)dx = 1$. The CDF $F(x)$ for a continuous random variable X is defined as $F(x) = P(X \in (-\infty, x]) = \int_{-\infty}^{x} f(u)du$. By the fundamental theorem of calculus, we have $\frac{d}{dx}F(x) = f(x)$.

For a discrete random variable X with PMF $p(x)$ for $x \in S$, where S is a countable set, the expected value or mean of the random variable is $E[X] = \sum_x x\ p(x)$, and for a continuous random variable X with PDF $f(x)$ for $x \in \mathbb{R}$, $E[X] = \int_{-\infty}^{\infty} x\ f(x)\ dx$, provided the right hand side is finite.

A *convolution* of two CDFs F_1 and F_2 is defined as the CDF F such that for all $x \in \mathbb{R}$, $F(x) = \int_{-\infty}^{\infty} F_1(x-y)dF_2(y)$ and is written as $F = F_1 * F_2$ (Chung, 2001) [Page 152]. For CDFs $F_1, F_2, \ldots, F_n$ of n independent continuous random variables $X_1, X_2, \ldots, X_n$; the CDF of $Z = X_1 + X_2 + \cdots + X_n$ can be represented as $F = F_1 * F_2 * \cdots * F_n$. In particular, the PDF of Z is

$$f_Z(z) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1 \text{ times}} f_{X_1}(z - x_2 - \cdots - x_n) f_{X_2}(x_2) \ldots f_{X_n}(x_n)\ dx_2\ dx_3 \ldots\ dx_n \tag{1.1.1}$$

In reliability theory, probability distributions with positive support (i.e; $B \subseteq \mathbb{R}^+$) are especially useful. Some such notable distributions are mentioned below. We also mention the form of their expectations, as these will be useful in Chapters 2-5.

- **Exponential distribution** For some $\beta > 0$, the PDF of an exponentially distributed random variable X is given by

$$f(x) = \beta \exp\{-\beta x\} \mathbf{1}\{x \geq 0\} \tag{1.1.2}$$

  where $\mathbf{1}\{x \in A\}$ is the indicator function that takes the value 1 when $x \in A$ and zero otherwise. The value β is called the rate parameter of the exponential distribution. The expectation of an exponential distribution is $E[X] = 1/\beta$. In the upcoming chapters, an exponential distribution with rate β is denoted as *exponential(β)*.

- **Gamma distribution** For some $\alpha > 0, \beta > 0$, a random variable X is said to be gamma distributed if the PDF is given by

$$f(x) = \frac{\exp\{-x/\beta\}\ x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} \mathbf{1}\{x \geq 0\} \tag{1.1.3}$$

  The parameter α is called the shape parameter, and β is called the scale parameter. The quantity $\Gamma(\alpha) = \int_0^{\infty} \exp(-x) x^{\alpha-1} dx$ is called the gamma function. The expectation of a gamma distribution is $E[X] = \alpha\beta$. In the upcoming chapters, a gamma distribution with shape parameter α and scale parameter β is denoted as *gamma(α, β)*.

- **Weibull distribution** For some $\alpha > 0, \beta > 0$, a random variable X is said to follow a Weibull distribution if the PDF is given by

$$f(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} \exp\{-(x/\beta)^{\alpha}\} \mathbf{1}\{x \geq 0\} \tag{1.1.4}$$

  The parameter α is called the shape parameter and β is called the scale parameter. The expectation of a Weibull distribution is $E[X] = \beta\ \Gamma(1 + 1/\alpha)$. In the upcoming chapters, a Weibull distribution with shape parameter α and scale parameter β is denoted as *Weibull(α, β)*.

- **Log-normal distribution** For mean parameter $\mu \in \mathbb{R}$ and the scale parameter $\sigma > 0$, the PDF of a log-normal distribution is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right\} \mathbf{1}\{x \geq 0\} \tag{1.1.5}$$

  The expectation of a log-normal distribution is $E[X] = \exp(\mu + \sigma^2/2)$ In the upcoming chapters, a log-normal distribution with mean parameter μ and scale parameter σ is denoted as *lognormal(μ, σ)*.

- **Inverse-Gaussian distribution** For mean $\mu > 0$ and shape parameter $\beta > 0$, the PDF of an inverse-Gaussian distribution is

$$f(x) = \sqrt{\frac{\beta}{2\pi x^3}} \exp\left\{-\frac{\beta(x-\mu)^2}{2\mu^2 x}\right\} \mathbf{1}\{x \geq 0\} \tag{1.1.6}$$

  The expectation of an inverse-Gaussian distribution is $E[X] = \mu$. In the upcoming chapters, an inverse-Gaussian distribution with mean parameter μ and shape parameter β is denoted as *inverse-Gaussian(μ, β)*.

Although there are many other important parametric life-distributions such as logistic, generalized gamma, Gompertz, Pareto, generalized inverse-Gaussian among others, in the upcoming chapters we will frequently refer to the above mentioned distributions. Figure 1.1 shows the density curves of the exponential, Weibull, gamma, log-normal and inverse Gaussian distributions with choice of parameters such that each has mean 1. These choices will be frequently referenced in Chapters 2-5. For any distribution, its parameter(s) can be estimated from the data using one or more techniques such as the method of moments, the maximum likelihood method, etc.

Next, let us discuss some ideas of stochastic processes that will be useful in this dissertation. The concepts below are cited from Ross *et al.* (1996) and Ross (2014).

A *stochastic process* $\mathbf{X} = \{X(t), t \in T\}$ is a collection of random variables. The set T is called the *index set* of the process and for each $t \in T$, $X(t)$ is a random variable called the *state* of the process at time t. The collection of all possible states of a stochastic process is called *state space*. For countable T, we call the stochastic process $X(t)$ a *discrete-time stochastic process* and for continuum T, it is called a continuous-time stochastic process. A realization $\mathbf{X}$ of a stochastic process is called a *sample path*.

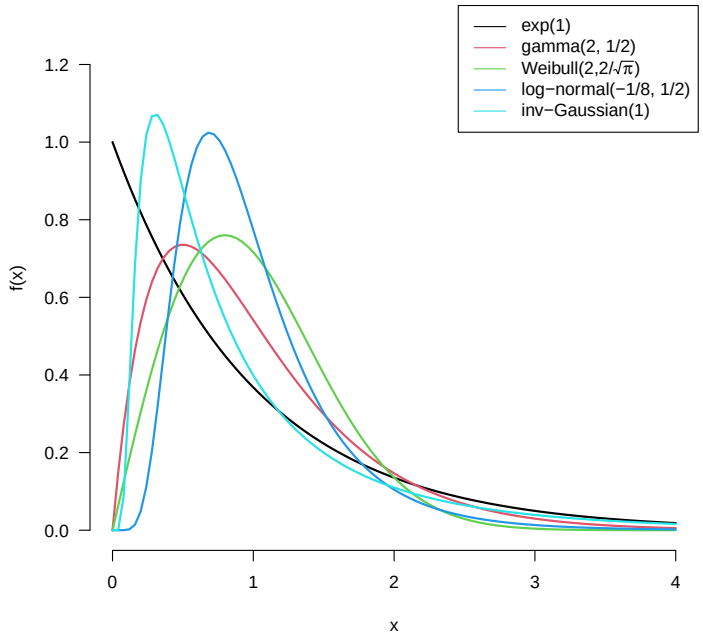


Figure 1.1. Densities of various distributions mentioned above each with mean 1.

Two processes $\{X_t\}_{t\in T}$ and $\{Y_t\}_{t\in T}$ are said to be *stochastically independent* if for all $n \in \mathbb{N}$, and for all $t_1, t_2, \ldots, t_n \in T$, the random vectors $(X(t_1), \ldots, X(t_n))$ and $(Y(t_1), \ldots, Y(t_n))$ are independent, or their joint CDF can be expressed as

$$F_{X_{t_1},\ldots,X_{t_n},Y_{t_1},\ldots,Y_{t_n}}(x_1, \ldots, x_n, y_1, \ldots, y_n)$$
$$= F_{X_{t_1},\ldots,X_{t_n}}(x_1, \ldots, x_n) \times F_{Y_{t_1},\ldots,Y_{t_n}}(y_1, \ldots, y_n) \;\; \forall\, x_1, \ldots, x_n, y_1, \ldots, y_n; \;\forall\, n \in \mathbb{N} \quad (1.1.7)$$

There can be different types of stochastic processes. If for all $n \in \mathbb{N}$, and for times $t_1 < t_2 < \cdots < t_n$, $P(X(t_n) \leq x|X(t_1), \ldots, X(t_{n-1})) = P(X(t_n) \leq x|X(t_{n-1}))$, then the stochastic process $\mathbf{X}$ is called a *Markov process*, and for such a process, the future probabilities are determined by the most recent values and not by the previous values.

A discrete-time stochastic process that takes on a finite or countable number of possible values can be expressed as $\{X_n, n = 0, 1, 2, \ldots\}$. If $X_n = i$, the process is said to be in *state*

i with a fixed probability P_{ij} that it will next be in state j. If $P(X_{n+1} = j|X_n = i, X_{n-1} = i_{n-1}, \cdots, X_1 = i_1, X_0 = i_0) = P_{ij}$ for all states $i_0, i_1, \ldots, i_{n-1}, i, j$ and all $n \geq 0$, then such a stochastic process is said to be a *Markov chain* where the probability of moving from state i in the n-th step to state j in the $(n+1)$-st step only depends on the fact that the process was in state i in the n-th step. The conditional distribution of any future state X_{n+1} is independent of past states $X_0, X_1, \ldots, X_{n-1}$ and depends only on the current state X_n. In other words, a continuous-time Markov chain is also called a Markov process. The P_{ij}'s are called *state transition probabilities* and have the properties $P_{ij} \geq 0$ for all $i, j \geq 0$; and $\sum_{j=0}^{\infty} P_{ij} = 1, i = 0, 1, \cdots$. The one-step transition probabilities P_{ij} for all observed states $i = 0, 1, 2, \ldots$ and $j = 0, 1, 2 \ldots$ are represented by the following transition probability matrix.

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \ldots \\ P_{10} & P_{11} & P_{12} & \ldots \\ \vdots & \vdots & \vdots & \\ P_{i0} & P_{i1} & P_{i2} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{1.1.8}$$

The *n-step transition probabilities* P_{ij}^n that a stochastic process which in the k-th transition $(k = 0, 1, 2, \ldots)$ is in state i will be in state j after n additional transitions is given by $P_{ij}^n = P(X_{n+k} = j|X_k = i),\ n \geq 0;\ i, j \geq 0$ and is calculated by raising P to the n-th power and taking the (i, j)-th element.

A stochastic process $\{N(t), t \geq 0\}$ is said to be a *counting process* if integer-valued $N(t)$ represents the total number of events that have occurred up to time t, where $N(t) \geq 0$; for $s < t, N(s) \leq N(t)$; and $N(t) - N(s)$ equals the number of events that occurred in the interval $(s, t]$. The arrival of an external impetus into a system is often characterized by counting processes. A common counting process is the Poisson process (also known as homogeneous poisson process (HPP)). The counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate β (> 0) if the number of events in any interval of length t is Poisson distributed with mean βt, that is, for all $s, t \geq 0$,

$$P(N(t+s) - N(s) = n) = exp\{-\beta t\}\frac{(\beta t)^n}{n!},\ n = 0, 1, \ldots \tag{1.1.9}$$

In particular, $E[N(t)] = \beta t$. Let the time to the first event be denoted by T_1 and that between the first and the second event be denoted by T_2. For time t, no event of the Poisson process occurs in the interval $[0, t]$, then $P(T_1 > t) = P(N(t) = 0) = \exp\{-\beta t\}$. That is, T_1 follows an exponential distribution with rate β. Therefore, an important property of a Poisson process is that the inter-arrival time distribution of events in such a process is exponential. Next, note that

$$\begin{aligned} P(T_2 > t | T_1 = s) &= P(0 \text{ events in } (s, s+t] | T_1 = s) \\ &= P(0 \text{ events in } (s, s+t]) \\ &= P(T_2 > t) \end{aligned} \tag{1.1.10}$$

which is free of s. Therefore, $P(T_2 > t) = E[P(T_2 > t|T_1)] = exp\{-\beta t\}$. Among continuous probability distributions, the above observed property is uniquely owned by the exponential distribution. It is called the *lack of memory* or *memorylessness* property and is the reason for the widespread use of exponential distribution in reliability models. Thus, for the exponential distribution, the waiting time for the next event at any moment does not depend on how long the system has been in the current state; rather, it depends on when we observe the system.

A state is called an *absorbing state* if once the system enters that state, it never leaves. The state j is said to be *accessible* from the state i if $P_{ij}^n > 0$ for some $n > 0$. If, moreover, state i is accessible from state j, states i and j are said to *communicate* with each other. A Markov chain is said to be *irreducible* if all states communicate with each other. Let f_{ij} denote the probability that starting in state i the system ever goes to state j. The state i is said to be *recurrent* if $f_{ii} = 1$ and *transient* if $f_{ii} < 1$. When it is possible to eventually get from every state to every other state with a positive probability, a Markov chain is called a *ergodic* or *irreducible* Markov chain.

For an irreducible ergodic Markov chain, $\lim_{n\to\infty} P_{ij}^n$ exists and is independent of i. Then letting $\boldsymbol{\pi}_j = \lim_{n\to\infty} P_{ij}^n$, for $j \geq 0$; $\boldsymbol{\pi}_j$ is the unique non-negative solution of

$$\boldsymbol{\pi}_j = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i P_{ij}, \text{ for } j \geq 0, \text{ and } \sum_{j=0}^{\infty} \boldsymbol{\pi}_j = 1 \tag{1.1.11}$$

Therefore, $\boldsymbol{\pi}_j$ represents the *long-run proportion* of time that the Markov chain is in state j. Such a $\boldsymbol{\pi}_j$ is called a *steady state probability* and $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)$ is called the *stationary distribution* of the Markov chain.

Let $S_n = \sum_{i=1}^{n} X_i,\ n \geq 1$ denote the arrival time of the n-th event, also called the *waiting time* of the n-th event. The n-th event occurs before or at time t if and only if the number of events occuring by time t is at least n; that is, $N(t) \geq n \Leftrightarrow S_n \leq t$. If the sequence of non-negative random variables $\{X_1, X_2, \dots, X_n\}$ is independent and identically distributed with mean $1/\beta$, then the resulting counting process $\{N(t), t \geq 0\}$ is a Poisson process with rate β (see Proposition 2.2.1, Ross *et al.* (1996)) and is a *renewal process*. Essentially, in a renewal process, the time between the first and second events has the same distribution as the time until the first event, and these times are independent of each other. The occurrence of an event indicates a renewal. We say that a cycle is completed every time a renewal occurs, and the expected time to complete a cycle is called *expected cycle time*.

If a process can be in any one of the states $1, 2, \dots, N$ and at each time it enters state i it remains there for a random amount of time having mean μ_i and then makes a transition into state j with probability P_{ij}, it is called a *semi-Markov process* because the probability distribution of the future state depends only on the current state (and not on the history of states visited so far); and the system stays in any state for a random duration whose distribution depends on the current state and the immediately next state. The time spent in a given state i is commonly termed *sojourn time*. Here, μ_i is the mean sojourn time in state i.

For a semi-Markov process, the *long-run probability* that a system will be found in state i, denoted by θ_i is independent of the initial state, and according to Proposition 4.8.1 of Ross *et al.* (1996) and the theorem that follows,

$$\theta_i = \frac{\boldsymbol{\pi}_i \mu_i}{\sum_{j=1}^{N} \boldsymbol{\pi}_j \mu_j} \tag{1.1.12}$$

where the denominator $\sum_{j=1}^{N} \boldsymbol{\pi}_j \mu_j$ is the expected cycle time.

1.2 Some useful mathematical results and concepts

Now, let us briefly present some mathematical concepts that will be useful in obtaining certain results in the following chapters.

- **Laplace transformation** For continuous random variable X that has a density function denoted by f with the positive half line as support, the Laplace transformation of f exists and is given by $\mathcal{L}(f,s) = \int_0^\infty exp\{-sx\}f(x)\ dx$.

- **Wald's first identity** (Wald, 1944) Let $\{X_n, n \in \mathbb{N}\}$ be a sequence of independent and identically distributed random variables and let $N \geq 0$ be an integer-value random variable independent of the sequence $\{X_n, n \in \mathbb{N}\}$. Suppose that N and X_n' have finite expectations. Then, for the random sum $S_N := \sum_{n=1}^{N} X_n$, the following holds.

$$E[S_N] = E[N] \times E[X_1] \tag{1.2.1}$$

- **Kullback-Leibler Divergence (KLD) measure** For continuous random variables X and Y with CDF F and G, respectively, and the corresponding PDFs $f(x)$ and $g(x)$ for $x \in \mathbb{R}$, the Kullback-Leibler divergence (KLD) of f and g is defined as

$$D_{KL}(F||G) = \int_{-\infty}^{\infty} f(x)\ log\left(\frac{f(x)}{g(x)}\right)\ dx \tag{1.2.2}$$

  Note that equation (1.2.2) is not symmetric in f and g. Here F is called the reference CDF, and the divergence measure gives an idea of how similar the two distributions are and hence is a measure of the goodness of fit of the two distributions. Therefore, the KLD is a test statistic which can be used to test the hypothesis $H_0 : F = G$ vs $H_1 : F \neq G$. If the parametric forms of the distributions F and G are unknown, the hypothesis can be tested using a simulated p-value as follows: Observations on F and G are randomly generated N times (where N is preferably large), the KLD is calculated each time and then the proportion of time the KLD values exceed the observed KLD is the simulated p-value. A similar approach has been described to calculate a simulated p-value for composite null models in Robins *et al.* (2000).

1.3 Relevant concepts in reliability theory

In this section, we present some relevant definitions and concepts in reliability theory that will be used in the upcoming chapters.

The ability of an item to perform a required function under given environmental and operational conditions and for a specified period of time is defined as its *reliability* (Hoyland and Rausand, 2009). The reliability function, denoted by $R(t)$ is theoretically defined as the probability of success (or being in the functional state) at time t. Mathematically, if T is a random variable that denotes time, then $R(t) = P(T \geq t) = 1 - F(t)$, where $F(t)$ is the CDF of the random variable T at time t. Note that $R(t) \in [0, 1]$ and is also commonly known as the survival function at time t. The reliability of a system is described by its ability to function under specified conditions for a given period of time. The opposite of being in a functioning state is called *failure*. The duration for which a system operates without failure is defined as *system lifetime*. Figure 1.2 shows the survival (or reliability function) curves for the choices of probability distributions mentioned in Section 1.1 each with mean 1.

A system is said to be *repairable* when we can renew or replace either the entire system or its components so that the system becomes functional again. Each component of a system is commonly termed a *unit*. After repair or replacement, a unit can sometimes regain its original form (such a repair is called a *perfect repair* and it renders the unit *as good as new*) or can be restored to a diminished quality than its previous state (such a repair is called *imperfect repair*). When a system works at its full potential without failure, we call the system is in an "up" state and, on the other hand, if the system is not in the working condition, we refer to that as a "down" state. Hence, the average time the system is in the up state is called the mean system up time (MSUT), and the average time that the system is in the down state is called the mean system down time (MSDT). Ideally, we want to minimize MSDT to maximize the lifetime of the system. Therefore, a system needs to be monitored to detect any failure or potential failure at a future time.

There are various monitoring policies that can either take place continuously (continuous monitoring) or at some specific intervals of time (periodic monitoring). Repair and replacement are essential maintenance operations that aim at maximizing the lifetime of a system.

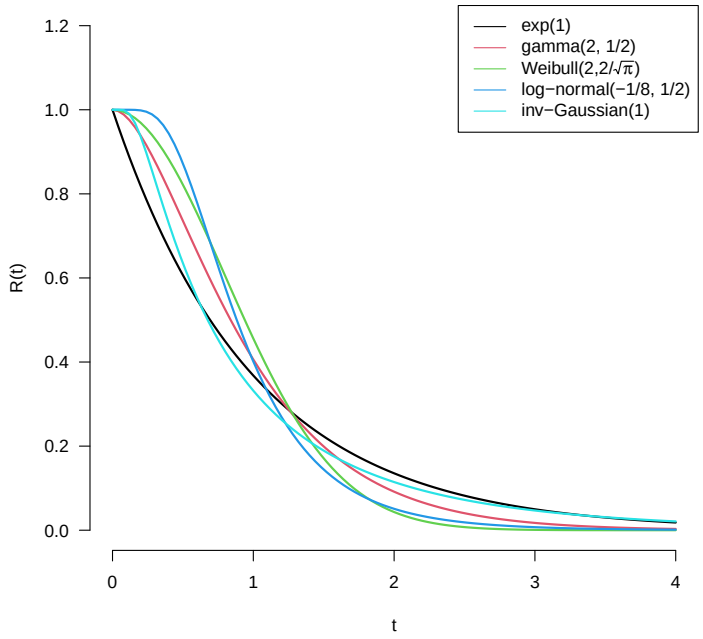


Figure 1.2. The reliability function $R(t)$ over time t for various lifetime distributions mentioned in Section 1.1 each with mean 1.

Typically, there are two types of *maintenance* operations in reliability theory: corrective maintenance (CM) and preventive maintenance (PM). In CM operations, components are repaired and then reattached to the system so that the system is restored to satisfactory operation in the shortest possible time. It includes diagnosis of the problem, repair and/or replacement of faulty components, and verification of the repair action. Sometimes spare units are used so that the system is not idle or down while repair is ongoing at the given moment, and once the repair is complete and the spare component fails, the repaired component is reattached to the system. If spare units may fail while kept as spare, they are said to be in hot standby and otherwise in *cold standby*. In this research, we will only consider the case where the spares are kept in cold standby. The time between such failure of a unit and the beginning of a maintenance action or between failure of a spare unit and reattachment of the repaired unit can be either sufficiently long during which the system will be once again in a down state, or it may be instantaneous. In Chapter 2 we will mostly encounter a CM.

For preventive maintenance (PM), the system is regularly inspected to detect any kind of depreciation and risk of failure. If a system is replaced because of an alarm that, with a high probability, it might fail within the next t units of time, then the time at which the replacement occurs is called *replacement time*, and the lifetime unused as a result of early replacement is defined as the *residual lifetime*. We will see such an example of time-based preventive maintenance in Chapter 5. The time taken for a unit to undergo repair is called *repair time*, which is a random variable, and similar to lifetime distributions, repair-time distributions also have positive supports. Often, the exponential distribution is the commonly used repair-time distribution due to its lack of memory property, which ensures that the probability of a repair continuing after a certain additional time depends on the observation epoch rather than how long the repair has been going on.

An important measure of the reliability of a system is the ability to perform its required function at a specified time or over a specified period of time and is defined as *availability*. The availability of the system at time t is $A(t) = P(\text{item is functioning at time } t)$. The mean value of the instantaneous availability function over the period $(0, T]$ is called the average availability and is denoted by $\overline{A(T)} = \frac{1}{T}\int_0^T A(t)dt$. The long-run probability that the system is functioning is called limiting availability, which is denoted by $A(\infty) = \lim_{t\to\infty} A(t)$, provided that the limit exists. Oftentimes, under continuous life- and repair-time distributions and continuous monitoring, the limiting availability exists; and equals *limiting average availability*, or the limiting proportion of time the system is up; and is given by

$$A_{av} = \lim_{t\to\infty} \overline{A(T)} = \frac{MSUT}{MSUT + MSDT} \tag{1.3.1}$$

When a down system returns to a working state, we say that a *renewal* happens. The average time between successive renewals is called the expected cycle time (ECT). This type of renewal process is known as an alternating renewal process because the state of the component(s) of the system alternates between a functioning state and a repair state. If a system is observed at certain intervals, at each observation epoch, the system is found to be in a certain state, and one can find probabilities of transitioning from one state to another due to failure or repair. This leads to the calculation of the long-term probability that a

system is in a given state using the renewal theorem (see equation (1.1.12)). For further reference, see Barlow and Proschan (1996); Marshall and Olkin (2007); Hoyland and Rausand (2009).

Sometimes, a working system can be affected by external impacts. Such examples in an industrial system can be sudden power failures, inventory shortages, delay in repair, internal mechanical faults, etc. Such impacts are widely studied in the literature and are commonly known as "shocks". As a result of a shock, a system can either fail instantaneously upon which repair and replacement actions need to be taken, or it can accumulate some *damage* (for example, damage in an electric circuit due to sudden voltage changes) that cumulatively makes the system weak and prone to eventual failure. Sometimes, systems may recover from the damages by maintenance operations or naturally by themselves. This phenomenon is called "healing". In our research, we have widely considered the type of healing that occurs naturally without the need for any impetus. Such healing is defined as *self-healing*. While in Chapters 3 and 4 we consider a type of self-healing where the net count of some of the damaging impacts is nullified by the healing effect; in Chapter 5 we consider continuous self-healing where healing is according to a continuous function of time.

Now, let us look at the factors that cause the system to slowly deteriorate in its working conditions and eventually fail; see Gorjian *et al.* (2010). Over time, the fault tolerance capacity of the system deteriorates, which is called *degradation*. Degradation can be categorized as a diminishing boundary strength for tolerating the accumulation of damage in a system. We use a non-increasing function of time t, denoted by $B(t)$ as the decreasing boundary threshold due to degradation in Chapter 5. Age is another important factor that causes a system to eventually lose its potential to function in the same way as compared to when it was younger. During such situations, before the system deteriorates further, an *age-replacement* is undertaken (Nakagawa, 2006). The effect of age on a system is incorporated in Chapters 3 and 4.

To determine the optimal time for a preventive maintenance action, we need to mathematically formulate a model that describes the associated costs and risks. Usually, it is assumed that if the unit fails before a certain time, a corrective action will occur, and if it

does not fail by that time, a preventive action may be taken. Thus, the optimal replacement time can be found by minimizing the cost per unit time ($CPUT$) that is given by

$$CPUT = \frac{\text{Total expected replacement cost per cycle}}{\text{Expected cycle time}} \tag{1.3.2}$$

The various associated costs are initial costs to install a system, cost of maintenance per unit time, revenue earned from the system per unit time (it is a negative cost), cost of inspection, cost of failure replacement, etc.

1.4 Outline of the dissertation

Let us now talk about the chapter-wise description of the current work. Each chapter has its own detailed introduction and literature review sections.

The availability of a maintained system is an important measure of reliability that results in improving its functionality, quality, efficiency, and ease of use. Heavy industries such as power plants, metal casting, chemical production, space administration, etc. rely on expensive technology for production and maintenance. The failure of such systems is detrimental to industry and may lead to economic and logistic challenges. Therefore, the system should be actively maintained by setting up a few repair facilities, and also storing another auxiliary spare unit to serve as replacement when a system/damaged unit needs to be repaired. There arise many logistical issues to address. For example, the system must be continuously monitored to detect failure and immediately switch the operation to the spare unit. Additionally, one must determine the optimal number of repair facilities to be created and the optimal number of spare units to be reserved so that the overall availability of the system is not compromised and the total cost is managed.

In Chapter 2, we examine a one-unit repairable system with identical spare and repair units in cold standby. If the system is continuously monitored and the repair time Y distribution is exponential, let $Y_{k|i}(1 \leq k \leq i)$ denote the time when k out of i failed units are repaired. Due to the lack of memory property of the exponential distribution, the successive differences $Y_{1|i}; \cdots ; (Y_{k+1|i} - Y_{k|i}); \cdots ; (Y_{i|i} - Y_{i-1|i})$ have independent exponential distributions with parameters $i\beta; \cdots ; (i-k)\beta; \cdots ; \beta$, respectively. Equivalently, $Y_{k|i}$ is the k-

th order statistics among i IID exponential(β) random variables. The PMF of N_i, the number of repairs completed during the lifetime of the current operating unit, can be calculated using the Laplace transform technique. However, when the repair time distribution is other than exponential, the lack of memory property does not apply and the successive differences $Y_{1|i}; \cdots ; (Y_{k+1|i} - Y_{k|i}); \cdots ; (Y_{i|i} - Y_{i-1|i})$ are not independent exponential distributions (Sarkar and Li, 2006). Therefore, except in the situation of having one spare and one repair facility, it is necessary to keep track of the time to repair all failed units at all times. Thus, identifying an embedded discrete-time Markov chain becomes difficult, and deriving the limiting average availability requires a different technique. Therefore, in Chapter 2 we introduce a computational technique such that we can incorporate non-exponential repair time distributions.

In Chapters 3 and 4, we look at systems which are exposed to external impacts. Here, we are not limited to a one-unit system. It can be any type of system commonly found in the reliability literature. We particularly categorize impacts according to their attributes with respect to the deterioration in the state of the system. Impacts that have a damaging effect are called valid shocks (VS), and those that may have a positive impact on the system that leads to healing are called positive interventions (PI).

The literature discusses various types of shock model, as described in (Nakagawa, 2007; Zhao *et al.*, 2018b; Gong *et al.*, 2020):

- In a *cumulative shock model*, the system is considered to have failed when the cumulative magnitude of the shocks exceeds a given threshold.
- In an *extreme shock model*, the system fails as soon as a massive catastrophic shock occurs.
- In a *run shock model*, the system is considered failed when there is a series of shocks whose magnitude is greater than a threshold.
- In a *δ-shock model* the system fails when the time lag between two adjacent shocks is less than a given critical value δ.

- In a *mixed shock model* two different shock models can be combined: For example, the system fails as soon as the cumulative magnitude of the shocks exceeds a critical level or r consecutive critical shocks occur, whichever occurs first.

In the current research, we are particularly interested in the cumulative shock model. We consider that the damaging and healing-inducing impacts arrive independently, and these define the corresponding stochastic paths. The system fails when cumulative damage at a given time t exceeds a fixed predetermined threshold. While in Chapter 3 we divide the lifetime of the system into two stages based on its age and healing ability, in Chapter 4 the healable stage is further divided into substages. Note that the arrival of VS and PI are according to independent point processes. The number of VS is a discrete random variable with a given PMF, which we can obtain from the observed data, and the inter-arrival processes of the VS and PI are stochastically independent. We present two approaches to describe the underlying stochastic process and compute the distributions of healable stage duration and the system lifetime. In Chapters 3 and 4 also, we remove the restrictive assumption on the inter-arrival time distributions. In the literature, either shocks/impacts are explicitly assumed to have exponential inter-arrival times, or even if they mention non-exponential inter-arrival times, the illustrations involve exponential examples only. We generalize the inter-arrival time distribution to be arbitrary. We also find optimal policies to replace the system before its failure by minimizing the costs per unit of time. This approach of using cost per unit time or sometimes revenue per unit time is a common optimization technique.

A major restriction of the setup in Chapters 3 and 4 is that we have limited ourselves to counts of shocks and not their magnitude. We consider equal magnitude of VS, a fixed threshold value, and a restrictive healing ability where healing is defined as nullification of the effect of one VS. Therefore, in Chapter 5 we consider a more comprehensive stochastic model of a system. In this chapter, we consider the varying magnitudes of shocks and their varying behavior. We also consider that the system is experiencing internal degradation (Gorjian *et al.*, 2010). We look into cases where (i) some shocks can heal up to a specific duration, (ii) some shocks are healable while others are not, and (iii) there are two types of shocks: healable shocks heal up to a fixed duration τ and non-healable shocks lead to sudden

degradation of the system. Here, we design a new time-based replacement policy to replace the system before it fails, while maximizing the lifetime and minimizing costs per unit time.

In this dissertation, we have defined phenomena such as accumulated system damage, self-healing behavior, and system degradation by some particular mathematical models with certain choices of parameters and hyperparameters. The following excerpt from Hoyland and Rausand (2009) [Page 12] justifies the importance of building mathematical models to study the reliability of the system.

> *A mathematical model is necessary in order to be able to bring in data and use mathematical and statistical methods to estimate reliability, safety, or risk parameters. For such models, two conflicting interests always apply:*
>
> - *The model should be sufficiently simple to be handled by available mathematical and statistical methods.*
> - *The model should be sufficiently "realistic" such that the deduced results are of practical relevance.*

The choices of parameters that are used to define the systems in this dissertation are not exclusive and may be modified as the practitioner deems appropriate.

To summarize, the main focus of this dissertation is to incorporate different approaches to model systems which are supported by several spare units and/or repair facilities and are exposed to external stress factors such as shocks having a deteriorating impact, which sometimes can heal. In all the approaches that we take, we essentially figure out how the system lifetime can be used to its utmost capacity. We illustrate our findings by considering various choices of the necessary parameters that define the system. Each chapter has separate sections that show detailed computational results to support our conclusions.

2. COMPUTING LIMITING AVERAGE AVAILABILITY OF A REPAIRABLE SYSTEM THROUGH DISCRETIZATION

Content in the following chapter was previously published by Reliability Engineering & Systems Safety. 2020, January: doi.org/10.1016/j.ress.2019.106616.
Debolina Chatterjee and Jyotirmoy Sarkar are co-authors of the published work.

2.1 Introduction

We recall a well-studied model of a repairable system and some known results under that model. However, several restrictive assumptions in this otherwise attractive model severely limit its applicability. Here, we remove these restrictive assumptions by devising a discretization approach, which reduces the burden of monitoring the system continuously, reproduces the results in the known special cases, and extends to the most general setting.

2.1.1 Formulation of the problem

Consider a continuously monitored one-unit repairable system supported by several identical spare units and several identical repair facilities. Initially, one unit is put on operation; and all spare units remain on cold standby (that is, spare units cannot fail). Upon failure of the operating unit, instantaneously a spare unit, if available, is put on operation (this is called instantaneous installation to operation) and the failed unit is sent to a repair facility (this is called instantaneous commencement of repair). Repair takes a random amount of time; and after repair the unit is restored back to a level equivalent to a new unit (this is called the perfect repair policy), which becomes a spare unit. We assume that lifetimes and repair-times are stochastically independent. The system fails (and enters a down state) when the operating unit fails and there is no spare unit on standby to take over the operation. Thereafter, when at least one of the repairs is completed, the repaired unit is immediately put into operation; and the system is revived.

The most important measure of success of a repairable system is the long run probability that the system is functioning, or the limiting availability of the system. Oftentimes,

under continuous life- and repair-time distributions and continuous monitoring, the limiting availability exists; and then it equals the limiting average availability, or the limiting proportion of time the system is up; and is given by

$$A_{av} = \frac{MSUT}{MSUT + MSDT} \tag{2.1.1}$$

where MSUT denotes the mean system up time, and MSDT denotes the mean system down time.

In the very special case of exponential lifetime and exponential repair-time distributions with means μ and ν, respectively, Barlow and Proschan (1996) [Page 206], provided the limiting average availability for the case of one repair facility ($r = 1$) and either no or one spare unit ($s = 0$ or $s = 1$). More specifically,

$$A_{av}(r = 1, s = 0) = \frac{\mu}{\mu + \nu} = \frac{1/\nu}{1/\nu + 1/\mu} \tag{2.1.2}$$

since, in this case, in equation (2.1.1) MSUT equals the mean time to failure and MSDT equals the mean time to repair; and

$$A_{av}(r = 1, s = 1) = \frac{\mu(\mu + \nu)}{\mu^2 + \mu\nu + \nu^2} = \frac{1/\nu}{1/\nu + 1/\mu - 1/(\mu + \nu)} \tag{2.1.3}$$

2.1.2 Availability in some other models

Allowing arbitrary distributions for the lifetime X and the repair-time Y, Sen and Bhattacharjee (1984) [Page 283], derived the limiting average availability of a one-unit system supported by one repair facility and one spare unit as

$$A_{av}(r = 1, s = 1) = \frac{E[X]}{E[\max\{X, Y\}]} \tag{2.1.4}$$

Indeed, when equation (2.1.4) is specialized to exponential life- and repair-time distributions, one can recover equation (2.1.3).

In Sarkar and Chaudhuri (1999), for a maintained system under continuous monitoring and perfect repair policy, the instantaneous availability is determined using the Fourier transform technique. Here repair-time is restricted to exponential, but lifetime is allowed to be either gamma or exponential. Further, using the same technique but incorporating several imperfect repairs before a replacement or a perfect repair, the availability is obtained for exponential lifetime and repair-time distributions (with possibly different parameters) in Biswas and Sarkar (2000).

Assuming periodic inspection, in Sarkar and Sarkar (2000), the system availability is determined when repair is perfect, lifetime is either gamma or exponential and repair-time is constant. The work is extended in Biswas *et al.* (2003) by allowing an imperfect repair policy and a random repair-time (specifically, exponential). Further in Sarkar and Sarkar (2001), a periodically inspected system supported by a spare unit and maintained with perfect repair or upgrade is considered; and both the instantaneous availability and the limiting average availability are determined for arbitrary lifetime, degenerate upgrade time and exponential repair-time. The paper Cui and Xie (2005) adds to the results of Sarkar and Sarkar (2000) by assuming that the periodic inspections take place at fixed time points after repair or replacement in case of failure.

Allowing arbitrary continuous lifetime, but restricting to exponential repair-times only, Sarkar and Li (2006) derived the limiting average availability of a one-unit system under continuous monitoring when there are $s \geq 1$ spare units and $r \geq 1$ repair facilities, by studying the embedded Markov chain (tracked at selected observation times), which is said to be in *State* i where $(i = 0, 1, ..., s, s + 1)$, if there are i failed units undergoing or awaiting repair by that observation time.

Apart from a one-unit system, availability has been studied also for a k-out-of-N system. For example, the authors of de Smidt-Destombes *et al.* (2004) study the interactions among several control variables such as preventive maintenance policy, spare part inventories, and repair capacity while they affect the system availability. They present an exact as well as an approximate method to develop a trade-off among these control variables. These authors also advocate in de Smidt-Destombes *et al.* (2007) a block replacement policy in which all failed and degraded components are repaired by a single repair shop while spare units take

over the operation. They provide two approximate methods to analyze the relationship between system availability and control variables. In both papers, they assume the component lifetimes and repair-times are exponentially distributed.

For a k-out-of-$N : G$ system, Wu *et al.* (2014) and Wu *et al.* (2018) allow the repair-time to have a general distribution, but assume the lifetime to be exponential. The former paper considered one repairman with a single vacation, while the latter considered a replaceable repair equipment that may fail during the repair period and then be replaced by a new one. Both papers used the supplementary variable technique and the Laplace transform to calculate the availability. The supplementary variable technique is implemented in Wang *et al.* (2019) to derive state equations by defining the system state space and sojourn time in each state to calculate the availability of the system.

2.1.3 Overcoming the challenge

Let us highlight a serious drawback in the models mentioned above to set the stage for our current research. Although not realistic, researchers often assume exponential life- or repair-time distribution to simplify mathematical derivations. They exploit the lack of memory property of the exponential distribution to ensure that the successive differences between life- or repair-times are independent exponential variables (with different rates), and thereby they obtain closed form expressions for the limiting average availability.

Can we make the model more realistic by allowing arbitrary lifetime and arbitrary repair-time distributions for any number of spare units and repair facilities? The challenge of obtaining the limiting average availability under this general setting is expressed in Sarkar and Li (2006) as follows:

> *When repair-time distribution is other than exponential, except for the case of $(r = 1, s = 1)$, one must keep track of the time on repair of all failed units at all times. Therefore, there is no hope of identifying an embedded discrete-time Markov chain, and the derivation of the limiting average availability will require a technique different from the one presented in this paper.*

Some recent papers allow arbitrary life- and repair-time distributions: In Levitin *et al.* (2015), the authors studied single-component repairable systems supporting different levels of workloads. They provide a numerical algorithm to evaluate the probability that the system will perform a specified amount of work within a specified mission time, and the associated conditional expected cost. The paper Levitin *et al.* (2017) models dynamic performance of multi-state series parallel systems with repairable elements that can function at different load levels and employs a universal generating function technique to assess system performance. Here the instantaneous availability is evaluated at different load levels. Further, in Levitin *et al.* (2019), the authors proposed a discrete-state continuous-time stochastic process to evaluate instantaneous availability for a common bus performance sharing (CBPS) system. The technique involves integration with respect to the joint distribution of (T_j, X_j) (where T_j denotes the detection time of the failure of the j-th component and X_j denotes the operation time).

The current chapter responds to the challenge posed in Sarkar and Li (2006) by adopting a discretization approach: We inspect the system only at discrete time points; and we intervene only when during inspection we find a unit has failed or the failed system is ready for revival because at least one repair has been completed. In particular, if a repair has been completed, but the operating unit has not failed, we do not intervene at all! Thus, this approach essentially discretizes the time variable. Moreover, it relaxes the burden of monitoring the system continuously to monitoring it periodically (at inspection times only); hence, it is logistically preferable.

In Section 2.2, we revisit the case of $(r = 1, s = 1)$; model the stochastic process through discretization as a semi-Markov process; derive the limiting average availability; and exhibit its closeness to the analytic result equation (2.1.4) of Sen and Bhattacharjee (1984). In Section 2.3, we extend the discretization method to the case of $(r = 2, s = 1)$; that is, we permit a second repair facility. Finally, Section 2.4 concludes the chapter with a summary.

2.2 Discretization approach for $(r = 1, s = 1)$

We assume the following:

(A1) Lifetimes of the units are independent and identically distributed (IID) continuous random variables with arbitrary cumulative distribution function (CDF) F on a positive support.

(A2) Repair-times are IID continuous random variables with arbitrary CDF G on a positive support.

(A3) Lifetimes and repair-times are stochastically independent.

(A4) Repair is perfect; that is, a repaired unit is as good as new.

(A5) The system is under periodic monitoring; that is, it is inspected at regular intervals.

(A6) Interventions are made only at observation epochs when an operating unit is found to have failed or when the down system is ready for revival because at least one failed unit has been repaired.

(A7) Whenever at inspection a unit is found to have failed, it is sent to the repair facility. Repair commences instantaneously if the facility is free. Otherwise, the failed unit awaits repair until the facility is free.

(A8) Installation to operation happens immediately when a failed unit is sent to repair (at an inspection epoch) and there is a spare unit (as a result of an already completed repair), or when the failed system is ready for revival at an inspection epoch because one of the failed units has been repaired.

2.2.1 States of the system

Figure 2.1 depicts the states of the system (with explanations below), transition between them and the random variables determining such transitions. We label the states of the system to indicate the number of failed units:

(S0) *State 0* means there is no failed unit.

(S1) *State 1* means there is one failed unit.

(S2) *State 2* means there are two failed units. Additionally, we must use a second index to indicate how long the repair on the first failed unit has been going on when the system enters *State 2*, because that will determine how long the system will stay in *State 2*. This second index splits *State 2* into sub-states: We say the system is in *State* $(2, k)$ for $k = 1, 2, \ldots, N-1$, if repair on the first failed unit has been going on for a duration $k\Delta$ when the other unit was detected to have failed. This is because we monitor the system only at epochs that are multiples of Δ from the start (or from system revival). Note that by the time the system is detected to have failed, repair on the first failed

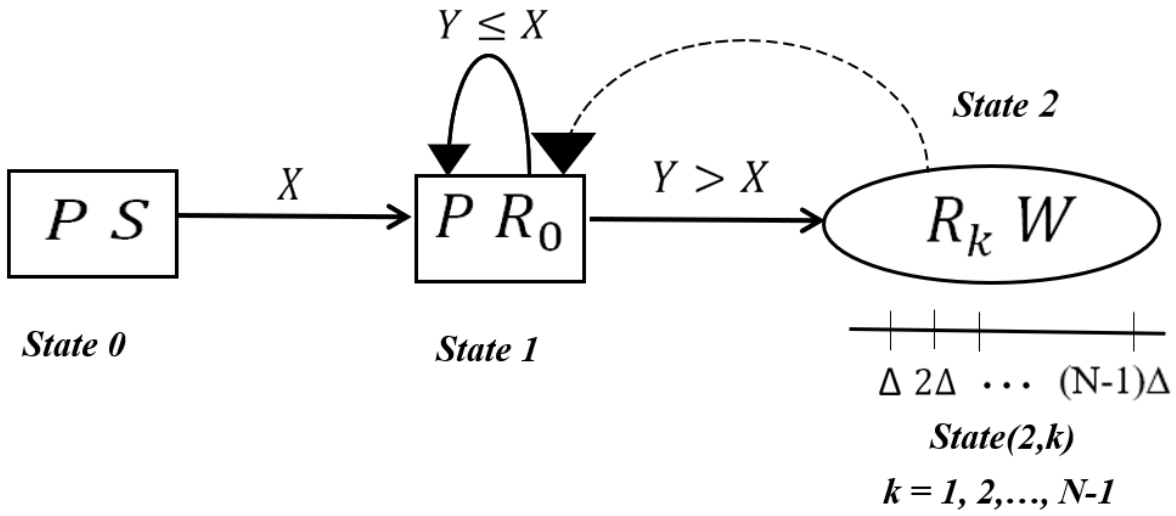


Figure 2.1. The state transition diagram for the $(r = 1, s = 1)$ case. A rectangle denotes an up state, and an oval denotes a down state. The status of each unit is denoted as follows: P for operation; S for standby; R for repair (with subscript indicating for how many inspection periods the repair has been going on); and W for waiting for repair.

unit has been going on for a positive duration. Hence, there is no *State* $(2, 0)$. Also, repair is surely completed in $N\Delta$ duration. Hence, there is no *State* $(2, N)$.

Let F and G denote the CDFs of the discretized lifetime and repair-time X and Y respectively. Let p and q denote the corresponding probability mass functions (PMFs) calculated by taking successive differences $p_k = F(k\,\Delta) - F((k-1)\,\Delta)$ and $q_k = G(k\,\Delta) - G((k-1)\,\Delta)$ respectively, for $k = 1, 2, \ldots, N$. Let R denote the CDF of $\max\{X, Y\}$ calculated

by taking product $R(k\,\Delta) = F(k\,\Delta)\,G(k\,\Delta)$, and let r denote the corresponding PMF of $\max\{X, Y\}$ obtained by successive differences $r_k = R(k\,\Delta) - R((k-1)\,\Delta)$ for $k = 1, 2, \ldots, N$.

We describe the transition probabilities between states of the system:

- At time $t = 0$, the system is in *State 0*, where one unit begins to operate and the other spare unit is on cold standby. The system goes from *State 0* to *State 1* when the operating unit is detected to have failed, repair starts on it and the spare unit is put on operation instantaneously. Hence,

$$P_{0\to 1} = 1 \tag{2.2.1}$$

  The system never returns to *State 0*.

- From *State 1*, after an intervention, the system can go to two places:

  (i) If repair on the failed unit is completed before the operating unit is detected to have failed, then we do not record this transition at all. Instead, we wait until the operating unit is detected to have failed at epoch $k\Delta$. Then we interchange the roles of the two units; and say that the the system has re-entered *State 1*. This happens with probability

$$P_{1\to 1} = \sum_{k=1}^{N} p_k\; G(k\Delta) \tag{2.2.2}$$

  (ii) On the other hand, if the operating unit is detected to have failed at epoch $k\Delta$, before the repair on the previously failed unit is completed, then the system goes to *State (2,k)* with probability

$$P_{1\to(2,k)} = p_k\{1 - G(k\Delta)\} \tag{2.2.3}$$

  In this case, the freshly failed unit awaits repair to commence on it only after the repair on the previously failed unit is found to be completed at an inspection

epoch. While the system is in *State 2* (that is, in any of the *States (2,k)*), no unit is operating; and the system is down.

- From *State (2,k)* the system surely goes to *State 1* when the ongoing repair on the first failed unit is found to be completed at an inspection time and the repair on the second failed unit begins. This happens with probability

$$P_{(2,k)\to 1} = 1 \tag{2.2.4}$$

In the proposed discretization approach, we split the repair-time into N (to be determined momentarily) intervals each of length Δ; and observe the system at epochs $k\Delta$ for $k = 1, 2, \ldots, N$. For all practical purposes, we assume that repair is completed only at epochs $k\Delta$, since those are the observation epochs (and possible installation epochs).

We choose N large enough so that the probability that the larger of lifetime and repair-time (hence, either lifetime or repair-time) exceeds $N\Delta$ is very small (preferably under 0.001, say); that is, $\{1 - R(N\Delta)\} \approx 0.001$.

The continuous-time stochastic process, after discretization, can be described as a Semi-Markov Process since the probability distribution of the future state depends only on the current state (and not on the history of states visited so far); and the system stays in any state for a random duration whose distribution depends on the current state and the immediately next state.

Moreover, from the above discussion of transitions and associated probabilities, we note that the embedded discrete-time Markov chain is irreducible (that is, all states communicate with one another); and since the state space is finite, the chain is recurrent.

Using the theory of semi-Markov processes (see Ross *et al.* (1996)), we can find the limiting proportion of time the system spends in each state. First, we find the stationary probabilities $\{\boldsymbol{\pi}_j, j \in S\}$ of the discrete-time Markov chain by solving the following state equations:

$$\boldsymbol{\pi}_j = \sum_{i \in S} \boldsymbol{\pi}_i P_{ij}, \text{ for all } j \in S; \text{ and } \sum_{j \in S} \boldsymbol{\pi}_j = 1 \tag{2.2.5}$$

where P_{ij} denotes the transition probability from *State* $i \in S$ to *State* $j \in S$ and the transition probability matrix P, which is of dimension $(N+1) \times (N+1)$, is as follows:

$$P = \begin{array}{c} \begin{matrix} 0 & 1 & (2,1) & \dots & (2,N-1) \end{matrix} \\ \begin{pmatrix} P_{0,0} & P_{0,1} & P_{0,(2,1)} & \cdots & P_{0,(2,N-1)} \\ P_{1,0} & P_{1,1} & P_{1,(2,1)} & \cdots & P_{1,(2,N-1)} \\ P_{(2,1),0} & P_{(2,1),1} & P_{(2,1),(2,1)} & \cdots & P_{(2,1),(2,N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_{(2,N-1),0} & P_{(2,N-1),1} & P_{(2,N-1),(2,1)} & \cdots & P_{(2,N-1),(2,N-1)} \end{pmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ (2,1) \\ \vdots \\ (2,N-1) \end{matrix}$$

$$= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \tag{2.2.6}$$

In the P-matrix above, the row and the column labels stand for the corresponding states. Note that although the transition matrix P is $(N+1) \times (N+1)$, it has non-zero entries (denoted by *) only in the second row corresponding to transition out of *State 1* and in the second column corresponding to transition into *State 1*. Therefore, it is straight-forward to solve equation (2.2.5).

Second, we find the expected sojourn time in each state; that is, the expected time the system stays in that state before it moves to a new state. If a unit is found to have failed at inspection time $k\Delta$, it must have failed during the interval $((k-1)\Delta, k\Delta]$. For simplicity, we assume that it has failed at the midpoint of the interval; that is, it was operating for the initial $\Delta/2$ period in the interval and was in failed state during the last $\Delta/2$ period (but was undetected). Although this is a rather crude assumption, it serves our purpose as far as computation of limiting average availability is concerned.

The expected sojourn times μ_0 and μ_1 in *State 0* and *State 1* respectively, both equal $E[X] - \Delta/2 = \sum_{k=1}^{N} p_k \, k\Delta - \Delta/2$, since we do not record a repair until after the operating

unit fails. We subtract $\Delta/2$ from the expected discretized lifetime to account for the fact that the system is actually down during the last $\Delta/2$ duration within each *State 0* and *State 1*.

The expected sojourn time $\mu_{(2,k)}$ in any *State* $(2,k)$ (a down state), is the expected additional repair-time, given that the previously failed item has been undergoing repair for $k\Delta$ time. For $k = 1, 2, \ldots, N$, we have,

$$\mu_{(2,k)} = E[Y|Y > k] = \sum_{j=1}^{N-k} \frac{q_{j+k}\ j\Delta}{1 - G(k\Delta)} \tag{2.2.7}$$

There is no need to make a further adjustment of $\Delta/2$ in equation (2.2.7) as the system is down the whole time while in *State (2,k)*.

Next, using Corollary to Proposition 4.8.1 of Ross *et al.* (1996), the limiting probability that the stochastic process will be found in *State* j (where j runs over all N States $1, (2,1), (2,2), \ldots (2, N-1)$) is independent of the initial state and is given by

$$\theta_j = \frac{\boldsymbol{\pi}_j \mu_j}{\sum_{i=1}^{N} \boldsymbol{\pi}_i \mu_i} \tag{2.2.8}$$

The denominator $\sum_{i=1}^{N} \boldsymbol{\pi}_i \mu_i$ in equation (2.2.8) is called the *expected cycle time*; and it is the expected time between successive renewals (or entry into *State* 1). Having calculated all θ_j's, we define $\theta_2 = \theta_{(2,1)} + \cdots + \theta_{(2,N-1)} = 1 - \theta_1$, since *State 2* is the aggregate of *States* $(2,1), (2,2), \ldots, (2, N-1)$.

Since the system is up in *States* 0 *and* 1, and down in *State* 2, but the system never returns to *States* 0, the limiting average availability of the system is given by

$$A_{av} = 1 - \theta_2 = \theta_1 \tag{2.2.9}$$

2.2.2 Computation and comparison

We want to compare the limiting average availability computed by equation (2.2.9) under discretization approach to the value computed by equation (2.1.4) under continuous monitoring. As a test case, let us assume a Weibull(shape=3, scale=1.12) lifetime distribution

with mean lifetime 1, and a Weibull(shape=2, scale=2) repair-time distribution with mean repair-time 1.77.

Under discretization approach, since $F(12)G(12) < 0.001$, we decompose the time range $(0, 12]$ into $N = 120$ intervals of length $\Delta = 0.1$ each. We construct the CDFs of discretized life- and repair-times, F and G, from the above mentioned Weibull distributions evaluated at $k\,\Delta$ for $k = 1, 2, \ldots, 120$. We construct the PMFs p, q, r as defined above by successive differences.

Using equations (2.2.1 - 2.2.4), we construct the transition probability matrix P, which in this case is of dimension 121×121. Recall from above that P has non-zero entries only in row 2 and column 2. Below we partially display the second row rounding each entry to 3 decimal places; all other entries of the second column are 1.

$$P = \begin{array}{c} \begin{matrix} 0 & 1 & (2,1) & (2,2) & (2,3) & (2,4) & \ldots & (2,N-1) \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & .252 & .001 & .005 & .013 & .024 & \cdots & * \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ (2,1) \\ \vdots \\ (2,N-2) \\ (2,N-1) \end{matrix} \tag{2.2.10}$$

Next, we calculate the stationary probabilities using equation (2.2.5):
We find $\boldsymbol{\pi}_0 = 0, \boldsymbol{\pi}_1 = 0.572$; and for *State* $(2,k)$*'s* (for $k = 1, 2, \ldots, N-1$), the stationary probabilities, rounded to 4 decimal places, are: $\{\boldsymbol{\pi}_{(2,1)}, \boldsymbol{\pi}_{(2,2)}, \ldots, \boldsymbol{\pi}_{(2,N-1)}\} = \{0.0004, 0.0028, 0.0075, 0.0140, 0.0218, 0.0300, 0.0375, 0.0433, 0.0464, 0.0465, 0.0435, 0.0381, 0.0311, 0.0237, 0.0167, 0.0110, 0.0066, 0.0037, 0.0019, 0.0009, 0.0004, 0.0001, 0.0001, 0, 0, 0, \cdots, 0\}$.

Lastly, the expected sojourn times in *State* 0 and *State* 1 are both obtained from $E[X] - \Delta/2 = \sum_{k=1}^{N} p_k\, k\Delta - \Delta/2$ as $\mu_0 = \mu_1 = 10.0014$. Likewise, for *State* $(2,k)$'s (for $k = 1, 2, \ldots, N-1$), we get the expected sojourn times (rounded to 4 decimal places) as $\{\mu_{(2,1)}, \mu_{(2,2)}, \cdots, \mu_{(2,N-2)}, \mu_{(2,N-1)}\} = \{17.2677, 16.3901, 15.5837, \cdots, 1.4000, 1.000\}$.

Therefore, $\theta_0 = 0$ and θ_j's for $j = 1, (2,1), \ldots, (2, N-1)$ are calculated using equation (2.2.8). In particular, $\theta_2 = 0.4665872$, and the expected cycle time $\sum_{i=1}^{N} \boldsymbol{\pi}_i \mu_i = 10.72444$. Moreover, using equation (2.2.9), the limiting availability to be $\theta_1 = 1 - \theta_2 = 0.5334128$.

Two comments follow:

(1) The exact analytic result, given in equation (2.1.4), yields the limiting availability to be 0.5334131. Thus, our discretization approach closely approximates the analytic result previously derived by Sen and Bhattacharjee (1984).

(2) For the case $(r = 1, s = 1)$, the limiting average availability is 0.53341, while for the case $(r = 1, s = 0)$, using equation (2.1.1), the limiting average availability is only $1/2.77 = 0.361$. Thus, there is a significant increase (47.76%) in A_{av} with the introduction of a spare unit.

For $(r = 1, s = 1)$, having established the test case of Weibull life- and Weibull repair-times, we carry out a more comprehensive study of various combinations of life- and repair-time distributions, always ensuring mean lifetime=1 and mean repair-time=1.77. We report in Table 2.1 the limiting average availability using both the analytical formula and the discretization approach. We extend the time range to $(0, 20]$ so that $F(20)G(20) < 0.001$, but we keep $\Delta = 0.1$, implying that there are 201 states.

Highlighted in the table is the special case when both life- and repair-time distributions are exponential. The analytic result for this case is already given in Barlow and Proschan (1996)[Page 206], Sen and Bhattacharjee (1984)[Page 283] and Sarkar and Li (2006)(Corollary 2.2). Here we demonstrate that the result of the discretization approach (0.46971) closely approximates the analytic result (0.46926). The slight discrepancy is due to crudely subtracting $\Delta/2$ from the expected sojourn times of the system up states; *State 0* and *State 1*.

To increase limiting average availability we have allowed a spare unit to take over the operation when the main unit has failed and is under repair. Of course, when there is only one repair facility (that is, $r = 1$), then when the system is down only the first failed unit is under repair while the other failed unit is awaiting repair. In order to improve the limiting average availability of the system, one strategy is to introduce one more repair facility to expedite the repair of the second failed unit. However, when there are multiple repair facilities, no

Table 2.1. Availability under different life- and repair-time distributions for the $(r = 1, s = 1)$ case. The top entry of each cell is the availability computed through discretization and the bottom entry using equation (2.1.4).

Lifetime \ Repair-time	Exponential (1/1.77)	Gamma (2, 0.855)	Weibull (2, 2)
Weibull (3, 1.12)	0.49341 *0.49335*	0.52055 *0.52055*	0.53341 *0.53341*
Gamma (2, 0.5)	0.48172 *0.48167*	0.50413 *0.50413*	0.51515 *0.51515*
Inverse-Gauss (1 , 1)	0.47221 *0.47215*	0.49058 *0.49057*	0.49867 *0.49904*
Exponential (1)	**0.46971** ***0.46926***	0.48787 *0.48746*	0.49677 *0.49638*
Lognormal (-0.5 , 1)	0.46263 *0.46452*	0.47865 *0.48062*	0.48946 *0.48902*

analytic result exists in the literature to allow both life- and repair-time distributions to be arbitrary. The close agreement between the values obtained from equations (2.1.4) and (2.2.9) gives us confidence to proceed with the discretization approach in case $r > 1$.

2.3 Discretization approach for $(r = 2, s = 1)$

Having justified the discretization approach when $(r = 1, s = 1)$, we proceed to apply it to the case of a second repair facility, where no analytic result is available. Here, $(r = 2, s = 1)$; that is, there are one operating unit, one identical spare unit and two identical repair facilities.

2.3.1 States of the system

Figure 2.2 shows the states of the system (with explanations below), transitions between them and the random variables determining the transitions.

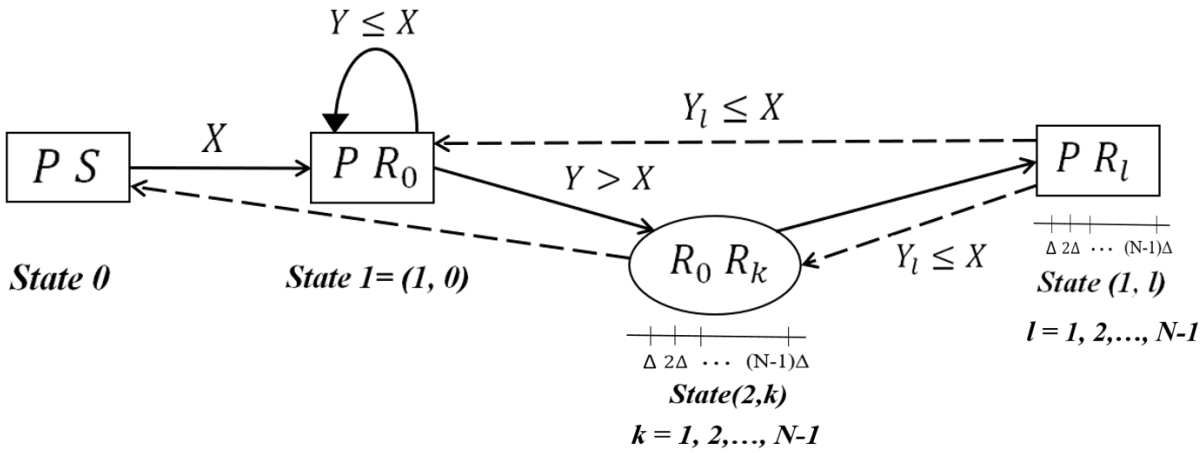


Figure 2.2. The state transition diagram for the $(r = 2, s = 1)$ case. The notations are the same as in Figure 2.1.

Initially, the system is in *State 0*, where one unit begins to operate and the other unit is on cold standby. We write the state-space of the system in two different notation—using one or two indices—depending on the level of details required for the analysis:

$$S = \{0; 1; 2^+; 1^+\} = \{0; (1,0); (2,1), \ldots, (2, N-1); (1,1), \ldots, (1, N-1)\}$$

where the first index i denotes how many units have been detected to have failed and are under repair, and the second index j tells us how long the repair on the first failed unit has been going on when the repair on the second failed unit just starts.

Let us explain the state space notation in terms of several examples:

- *State* $1 = (1, 0)$ means that one unit has been detected to have failed; it has been placed on repair just now, so that its repair duration so far is 0; and the other unit has just been placed on operation.

- Note that there is no *State* $(2, 0)$ because by the time failure on the second unit is detected, the repair on the first failed unit has already started and it has been going on for a positive multiple of Δ. Also, there is no *State* $(2, N)$ because if repair has been

going on for duration $N\Delta$, it must have been completed. Likewise, there is no *State* $(1, N)$.

- *State* $(2, 5)$ (provided, of course, $N > 5$) means that the system just entered *State 2* (that is, both units are known to have failed); repair on the first failed unit has been going on for 5Δ periods; and repair on the second failed unit has just started.

- *State* $(1, 7)$ (provided, of course, $N > 7$) means that repair on the only failed unit has been going on for 7Δ periods when the other unit is just put on operation (hence, there is only one failed unit).

Recall that we only record those inspection epochs when a failure is detected or when a down system is ready for revival because at least one unit has been repaired. In particular, we do not record epochs when a repair is completed, but the other unit is still operating.

Next, let us write down the recorded transitions between states and the associated transition probabilities. Recall that we monitor the system only at epochs $\Delta, 2\Delta, 3\Delta, \ldots$. As in the case of $(r = 1, s = 1)$, we assume that X is the discretized lifetime with CDF F and PMF p; and Y is the discretized repair-time having CDF G and PMF q. Also, we choose N such that the larger of life- and repair-times exceeds $N\Delta$ with probability at most 0.001.

- From *State 0*, the system surely goes to *State 1=(1,0)* after a random lifetime having PMF p. Therefore,

$$P_{0\to(1,0)} = 1 \tag{2.3.1}$$

- From *State 1=(1,0)*, if the operating unit is still functioning at epoch $k\,\Delta$, we do nothing. But if the operating unit is found to have failed at epoch $k\Delta$, then it must have failed in the interval $((k-1)\Delta, k\Delta]$, which happens with probability $p_k = F(k\,\Delta) - F((k-1)\,\Delta)$. There are two distinct cases to consider:

  (i) Repair is already completed by epoch $k\Delta$ (that is, repair is finished sometime during $(0, k\,\Delta]$), which happens with probability $P(Y \le k\Delta)$=$G(k\Delta)$. In this case, interchange the roles of the two units—the repaired unit takes over the

operation and the failed unit is put on repair. Hence, the system re-enters *State 1=(1,0)*. Hence,

$$P_{(1,0)\to(1,0)} = \sum_{k=1}^{N} p_k \, G(k\,\Delta) \tag{2.3.2}$$

(ii) Repair is not completed by epoch $k\Delta$, which happens with probability $P(Y \leq k\Delta) = G(k\Delta)$. In this case, the system goes down, since both units have failed and there is no other spare unit to take over operation. More specifically, the system enters *State* $(2, k)$. Hence,

$$P_{(1,0)\to(2,k)} = p_k \, \{1 - G(k\,\Delta)\} \tag{2.3.3}$$

- When the system enters *State* $(2, k)$, we continue to observe the system at regular intervals of Δ, labeling those epochs as $(k+1)\Delta, (k+2)\Delta, \ldots$. Two distinct cases are possible:

  (i) Both failed units are repaired during the same time interval, say, $((k+j-1)\Delta, (k+j)\Delta]$, where $j = 1, 2, \ldots, N-k$. To find the probability of this case happening, add over all j the product of two independent probabilities: Given that the repair of the first failed unit was not completed by time $k\Delta$, the conditional probability that it is completed during $((k+j-1)\Delta, (k+j)\Delta]$ is $\frac{q_{k+j}}{1-G(k\Delta)}$. The probability that the second failed unit on which repair started at epoch $k\Delta$ is repaired during the same time interval as the first failed unit is q_j. Finally, note that in this case, one of the repaired units (it does not matter which one, since the two units are identical) is put on operation and the other becomes a standby spare; that is, the system enters *State 0*. Therefore,

$$P_{(2,k)\to 0} = \sum_{j=1}^{N-k} q_j \, \frac{q_{k+j}}{1 - G(k\Delta)} \tag{2.3.4}$$

  (ii) One of the repairs is completed, but not the other. In this case, the repaired unit is put on operation; and the repair on the other unit, which has been going on for

$l\Delta$ time, continues on, causing the system to enter *State* $(1, l)$. The meaning of l is explained below in two sub-cases depending on which repair is completed—repair on the first failed unit, or repair on the second failed unit.

(a) Suppose that the first failed unit, on which the repair has been going on for $k\Delta$ time, is repaired earlier; and it happens during interval $((k+l-1)\Delta, (k+l)\Delta]$. The conditional probability of this event is $\frac{q_{k+l}}{1-G(k\Delta)}$. The probability that the second failed unit, on which repair had started freshly at epoch $k\Delta$, will not be repaired within the additional $l\Delta$ duration is $P(Y > l\Delta) = 1 - G(l\Delta)$.

(b) Suppose that the second failed unit, on which repair started at epoch $k\Delta$, gets repaired earlier; and it happens during interval $((l-1)\Delta, l\Delta]$, which has probability q_{l-k}. Then the conditional probability that the first failed unit will not be repaired by epoch $l\Delta$, given that the repair was not completed by epoch $k\Delta$, is $\frac{1-G(l\Delta)}{1-G(k\Delta)}$.

Combining the two sub-cases (a) and (b), we have

$$P_{(2,k)\to(1,l)} = [q_{k+l} + q_{l-k}] \left\{ \frac{1 - G(l\Delta)}{1 - G(k\Delta)} \right\} \tag{2.3.5}$$

where we interpret $q_t = 0$, unless $1 \leq t \leq N$.

- From *State (1,l)*, the system can go to one of two directions:

(i) If repair is completed before the operating unit fails, we do not record that transition; instead, we wait until the operating unit fails, say during interval $((j-1)\Delta, j\Delta]$ (for $j = 1, 2, \ldots, N$), with probability p_j, and the system goes to *State (1,0)*. The conditional probability that repair is completed before this additional time $j\Delta$, given that the repair was not completed by time $l\Delta$, is $\frac{G((l+j)\Delta)-G(l\Delta)}{1-G(l\Delta)}$. Hence,

$$P_{(1,l)\to(1,0)} = \sum_{j=1}^{N} p_j \left\{ \frac{G((l+j)\Delta) - G(l\Delta)}{1 - G(l\Delta)} \right\} \tag{2.3.6}$$

where we interpret $G(t\Delta) = 1$, whenever $t \geq N$.

(ii) If the operating unit fails during interval $((k-l-1)\Delta, (k-l)\Delta]$, which happens with probability p_{k-l}, before repair of the failed unit is completed, then the system goes down and enters *State* $(2,k)$, where $k > l$. Given that the ongoing repair is not completed by time $l\,\Delta$, the conditional probability that the repair will not be completed in additional time $(k-l)\,\Delta$ (that is, by epoch $k\,\Delta$) is $\frac{1-G(k\Delta)}{1-G(l\Delta)}$. Hence, for $k > l$,

$$P_{(1,l)\to(2,k)} = p_{k-l}\left\{\frac{1-G(k\Delta)}{1-G(l\Delta)}\right\} \tag{2.3.7}$$

Considering all the above state transition, the transition probability matrix P is of dimension $2N \times 2N$ and has the following structure:

$$P = \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & (2,1) & \dots & (2,N-1) & (1,1) & \dots & (1,N-1) \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * & 0 & \cdots & 0 \\ * & 0 & 0 & \cdots & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ * & 0 & 0 & \cdots & 0 & * & \cdots & * \\ 0 & * & * & \cdots & * & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * & 0 & \cdots & 0 \end{pmatrix} \end{array} \begin{array}{c} 0 \\ 1 \\ (2,1) \\ \vdots \\ (2,N-1) \\ (1,1) \\ \vdots \\ (1,N-1) \end{array} \tag{2.3.8}$$

The row and column labels in above matrix represent the corresponding states. As in the case of $(r=1, s=1)$, here also the continuous-time stochastic process, after discretization, is a semi-Markov process. Hence, the analysis follows along similar lines.

First, we find the stationary probabilities $\{\boldsymbol{\pi}_j, j \in S\}$ of the discrete-time Markov chain by solving the state equations that are similar in structure to equation (2.2.5), but involve many more states.

Second, we find the expected sojourn time in each state. In fact, the expected sojourn times $\mu_0, \mu_{(1,0)}$ and $\mu_{(1,l)}$ in *States* $0, (1,0), (1,l)$, for $1 \le l \le N-1$, are all equal to $E[X] - \Delta/2 = \sum_{k=1}^{N} p_k\, k\Delta - \Delta/2$. [The subtraction of $\Delta/2$ accounts for the system being down during the last $\Delta/2$ duration within each state $0, (1,0), (1,l)$.] The expected sojourn time $\mu_{(2,k)}$ in *State*

$(2,k)$ (a down state) is the expected value of the minimum of the two repair-times Y_0 and Y_k having CDFs $G(j)$ and $\frac{G(k+j)-G(k)}{1-G(k)}$ for $0 \leq j \leq N$ (with $G(t) = 1$ for $t > N$) respectively. Using Problem 1.1 of Ross *et al.* (1996), this expectation can be found as the sum of the survival function evaluated at non-negative integers. That is, for $k = 1, 2, \ldots, N$, we have

$$\begin{aligned} \mu_{(2,k)} = E[\min\{Y_0, Y_k\}] &= \sum_{j=0}^{N} P(Y_0 \geq j, Y_k \geq j) \\ &= \sum_{j=0}^{N-k} \frac{[1 - G(j\Delta)][1 - G((k+j)\Delta)]}{1 - G(k\Delta)} \end{aligned} \tag{2.3.9}$$

Here, there is no need to make an additional adjustment of $\Delta/2$ as the system is down throughout the time it is in *State (2,k)*.

Next, using Corollary to Proposition 4.8.1 of Ross *et al.* (1996), the limiting probability that the stochastic process will be found in *State* j is independent of the initial state and is given by expressions of the form equation (2.2.8), but with many more states. Let us define *State* 1^+ as aggregate of *States* $(1,1), (1,2), \ldots, (1, N-1)$ and *State* 2 as aggregate of *States* $(2,1), (2,2), \ldots, (2, N-1)$.

Having calculated all θ_j's, we define $\theta_2 = \theta_{(2,1)} + \cdots + \theta_{(2,N-1)}$. Since the system is up in *States* $0, 1, 1^+$, and down in *State* 2, all states being recurrent, the limiting average availability of the system is given by

$$A_{av} = 1 - \theta_2. \tag{2.3.10}$$

2.3.2 Computations and comparison

We compute the limiting average availability for various life- and repair-time distributions, always choosing mean lifetime 1 and mean repair-time 1.77. We have truncated all distributions to have support $[0, 12]$, which we have partitioned into 120 equal sub-intervals; that is, we choose $\Delta = 0.1$. Consequently, there are 240 states in the state space S.

The transition probability matrix P is 240×240, whose entries, using equations (2.3.1 - 2.3.7) and rounded to 4 decimal places, are partially displayed:

$$P = \begin{array}{c} \begin{array}{ccccccccc} 0 & 1 & (2,1) & \dots & (2,N-1) & (1,1) & \dots & (1,N-1) & \end{array} \\ \left(\begin{array}{ccccccccc} 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0.2517 & 0.0007 & \cdots & * & 0 & \cdots & 0 \\ 0.3213 & 0 & 0 & \cdots & 0 & 0.0075 & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0.0025 & 0 & 0 & \cdots & 0 & 0.9975 & \cdots & 0 \\ 0 & 0.2875 & * & \cdots & * & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0.9997 & 0 & \cdots & 0.0003 & 0 & \cdots & 0 \end{array} \right) \begin{array}{c} 0 \\ 1 \\ (2,1) \\ \vdots \\ (2,N-1) \\ (1,1) \\ \vdots \\ (1,N-1) \end{array} \end{array} \tag{2.3.11}$$

The stationary probabilities are obtained by using equation (2.2.5). They are $\boldsymbol{\pi}_0 = 0.010$, $\boldsymbol{\pi}_{(1,0)} = 0.265$, and $\{\boldsymbol{\pi}_{(2,1)}, \boldsymbol{\pi}_{(2,2)}, \boldsymbol{\pi}_{(2,3)}, \dots, \boldsymbol{\pi}_{(2,N-2)}, \boldsymbol{\pi}_{(2,N-1)}\} = \{0.0002, 0.0013, 0.0035, \dots, 0, 0\}$. The expected sojourn times in *State* 0, *State* $(1,0)$ and *State* $(1,l)$ for $l = 1, 2, \dots, N-1$ are all equal to 10.0016. Using equation (2.3.9), $\{\mu_{(2,1)}, \mu_{(2,2)}, \mu_{(2,3)} \dots, \mu_{(2,N-2)}, \mu_{(2,N-1)}\} = \{12.549, 12.093, 11.665, \cdots, 1.399, 1\}$.

Next, using equation (2.2.8), we see that the limiting probabilities that the stochastic process will stay in a *State* j, for $j \in S$ are respectively $\theta_0 = 0.0106$, $\theta_{(1,0)} = 0.2794$, $\theta_{1^+} = 0.3764$ and $\theta_2 = 0.4666$. Also, the expected cycle time is 9.493. Finally, using equation (2.3.10), the limiting average availability is obtained as 0.66650.

Furthermore, in Table 2.2, we display the limiting average availability calculated for the same set of life- and repair-times as in the case $(r = 1, s = 1)$ and the percentage improvement when $(r = 2, s = 1)$. Table 2.2 exhibits about 25-35% increase in limiting average availability when a second repair facility is included in the presence of one spare unit.

Table 2.2. We compare the limiting average availability between cases $(r = 1, s = 1)$ and $(r = 2, s = 1)$. The top entry in each cell is the computed availability for $(r = 2, s = 1)$; and the bottom entry is the percentage increase in availability compared to the $(r = 1, s = 1)$ case given in Table 2.1.

Life-time \ Repair-time	Exponential (1/1.77)	Gamma (2, 0.855)	Weibull (2 , 2)
Weibull (3, 1.12)	0.65807 *33.37*	0.66392 *27.54*	0.66650 *24.94*
Gamma (2, 0.5)	0.64764 *34.44*	0.65057 *29.05*	0.65171 *26.51*
Inverse-Gauss (1 , 1)	0.63903 *35.33*	0.63992 *30.44*	0.63943 *28.23*
Exponential (1)	**0.63676** ***35.56***	0.63718 *30.61*	0.63693 *28.21*
Lognormal (-0.5 , 1)	0.63024 *36.23*	0.63009 *31.64*	0.62537 *29.88*

2.4 Summary

Recall from Section 2.2 that our discretization approach closely approximates the analytic result for the $(r = 1, s = 1)$ case. Also, from Section 2.3 we note that for the $(r = 2, s = 1)$ case under exponential life- and exponential repair-times, the analytic result of Sarkar and Li (2006), yields a limiting average availability of 0.63871, while our discretization approach using equation (2.2.9) gives a limiting average availability of 0.63676. Hence, we claim that the discretization approach works reasonably well; and it can be used to compute the limiting average availability for *any* life- and repair-time distributions. We also find that as we increase an additional spare unit from $(r = 1, s = 0)$ to $(r = 1, s = 1)$ or as we add an additional repair facility from $(r = 1, s = 1)$ to $(r = 2, s = 1)$ there is a significant increase in the limiting average availability of the system. We anticipate a further increase in limiting average availability when r and s are increased, but that will also lead to increase in the number of states and become computationally burdensome. Nevertheless, the discretization approach presented in this chapter can yield the limiting average availability under any arbitrary continuous life- and repair-time distributions for other systems as well.

3. OPTIMAL REPLACEMENT POLICIES FOR SYSTEMS UNDER SPORADIC SHOCKS AND HEALING IMPETUS

Content in the following chapter was previously published by Quality Technology and Quantitative Management. 2022, April: doi.org/10.1080/16843703.2022.2051846.
Debolina Chatterjee and Jyotirmoy Sarkar are co-authors of the published work.

3.1 Introduction

Shocks are sudden external stimuli that cause changes to the normal functioning of a system. In the last two decades, shock models have been studied quite extensively by researchers across different domains. The threat posed by shocks to a system and the associated costs have motivated researchers to find optimal solutions to problems of choosing monitoring intervals, applying preventive repairs, and timing system replacement. The primary purpose in this chapter is to incorporate arbitrary inter-arrival distributions of shocks and healing impetus in describing the evolution of a system and in deciding optimally when to replace a system. Let us review some existing papers in shock model literature which can be broadly classified based on arrival times of shocks, types of shocks, aging and degradation, amount of damage inflicted by the shocks, and healing. We also mention papers that discuss different maintenance policies and give real life examples.

The arrival process of the shocks play an important role in reliability studies of a system. Most common choices are homogeneous Poisson process (HPP) (Rafiee *et al.*, 2015), non-homogeneous Poisson process (NHPP) (Sheu and Chien, 2004; Chien and Sheu, 2006; Chien *et al.*, 2006), compound Poisson process (Wang *et al.*, 2017), and "phase-type" distribution (Zhao *et al.*, 2018a; Eryilmaz and Kan, 2019).

Shocks can differ in types. Eryilmaz and Kan (2019) assumes two types of shocks that arrive with certain probabilities: "fatal" shocks cause significant damage, but "non-fatal" shocks do not destroy any component. Similarly, Chien and Sheu (2006) and Chien *et al.* (2006) consider type I (minor) and type II (catastrophic) shocks.

Some papers consider system aging and degradation according to some well-defined processes such as the Weiner process in Cui *et al.* (2016); Kong *et al.* (2017); Dong *et al.* (2020a); Jia *et al.* (2020); Gao *et al.* (2020); Huang *et al.* (2021); linear degradation in Rafiee *et al.* (2015); and gamma process in Wang *et al.* (2017).

Some shock models include a change point where the system behavior changes depending on the degradation process, the count of shocks and the cumulative load from the shocks. Accordingly, the lifetime is split into stages. Gao *et al.* (2020) incorporates two types of phase change patterns (PCPs): PCP I occurs where the number of shocks reaches a threshold (prefixed); and PCP II occurs when the cumulative load of shocks reaches a specified threshold. Zhao *et al.* (2019) divides the component states into stages according to the degree of damage.

The concept of "self healing" is mentioned in Cui *et al.* (2018) and Shen *et al.* (2018). The former paper derives bounds for reliability for systems experiencing shocks that arrive according to some stochastic process, are classified into different types, and inflict different magnitude of damage. In the latter paper, external shocks of varying intensities arrive according to a Poisson process, and once the system has accumulated enough damage, it loses its self healing ability. Zhao *et al.* (2018b) study a two-stage mixture shock model where all shocks inflict equal damage and self-healing occurs only in Stage 1. Other notable work on self-healing include Dong *et al.* (2020a) and Dong *et al.* (2020b).

Several maintenance policies are studied to determine the optimal replacement rule. The paper Rafiee *et al.* (2015) finds optimal inspection intervals that minimize the long run average cost of system maintenance. Dong *et al.* (2020b) uses Nelder-Mead downhill simplex method to find the optimal age to replace the system by minimizing the long run average cost per unit time. Here the system is also able to self-heal from external shocks. In Tekin and Eryilmaz (2019), the system is replaced either on failure or at the optimal replacement time (determined by minimizing the long-run average cost), whichever happens first.

Some applications of shock models in real life are provided in Usynin and Hines (2007); Lafont *et al.* (2012); Keedy and Feng (2013); Bhuyan and Dewanji (2017).

Taking a cue from the above-mentioned papers, we consider two types of impacts: "valid shocks" (VS) that cause equal damage; and "positive interventions" (PI) whose accumulation triggers a healing effect. The main objective of this chapter is to remove the restrictive

assumption on the inter-arrival time distributions. The above-mentioned papers either explicitly assume shocks/impacts have exponential inter-arrival times, or even if they mention non-exponential inter-arrival times, they illustrate their methodologies with only exponential examples. In this chapter, we generalize the inter-arrival time distribution to be arbitrary. Whenever we can count the number of VS and PI, we can compute the distribution of Stage 1 duration and the system lifetime. We also allow system aging, which causes loss of healing capability in Stage 2. Furthermore, we study three replacement policies, adapted from those of Zhao *et al.* (2018b), each of which optimize the average cost per unit time to operate the system under different scenarios.

This chapter is organized as follows: Section 3.2 describes the evolution of the system under shock and healing, together with a real life example; Section 3.3 illustrates two approaches to compute the distributions of Stage 1 duration and lifetime, and displays the results; Section 3.4 describes the maintenance policies, and finally Section 3.5 summarizes the main findings of this research.

3.2 The system set up and assumptions

We consider a system that is exposed to external impacts which can be broadly classified into two types: impacts which cause damage to the system are called *valid shocks (VS)* and those which do not cause any damage, rather induce a healing effect, sufficient accumulation of which nullifies the damage due to one VS, are called *positive interventions (PI)*. We assume arbitrary inter-arrival times of VS and PI such that the system can experience long enough lifetime before it fails or becomes severely damaged requiring replacement. We assume each VS causes an equal amount of damage; and likewise each PI contributes equally towards healing.

Let $\{X_i; i \in \mathbb{N}\}$ denote the inter-arrival times of VS; and let them be independent and identically distributed (IID) with an arbitrary distribution function F. Similarly, let $\{Y_j; j \in \mathbb{N}\}$ denote the inter-arrival times of PI; and let them be IID with another distribution function G. Further, assume that the arrival process of PI is stochastically independent of the arrival process of the VS.

The PIs help the system heal from the damages inflicted by the VS. The accrual of k *cumulative* PIs nullifies the damage caused by one VS, thereby reducing the net number of VS by one. Here, the net number of VS equals the number of VS that have arrived *minus* the number of VS that have been nullified by PIs. Unlike in Zhao *et al.* (2018b), we subscribe to the notion that the healing effect of an already occurred PI is not eliminated by the arrival of a VS. The effect lingers on, and k cumulative PIs (whether or not interrupted by a VS in the interim) nullify one VS.

However, we must not forget that the system is also aging: It is not practical to assume that the healing effect will go on forever. Therefore, we divide the system lifetime into two stages. In *Stage 1*, the system retains the capacity to heal. After the net number of VS reaches a given (prefixed) threshold m_1, we say the system has endured sufficient amount of damage so that it has lost its capacity to heal, and as such the system has transitioned into *Stage 2*. The epoch when the system transitions from Stage 1 to Stage 2, is called the *change point* for the system. In Stage 2, even if the PIs keep coming, they do not heal the system. Therefore, in this stage, the net count of VS only keeps on increasing as the VSs arrive. When it reaches another prefixed threshold m_2 ($> m_1$), the system fails, and requires replacement. We denote the sojourn time in Stage 1 by T_1, and time to system failure as T_2, implying that the sojourn time in Stage 2 is $T_2 - T_1$. These notions are illustrated in Figure 3.1.

In Zhao *et al.* (2018b), the shocks that do not cause significant damage and the time-lag from the previous shock exceeds the threshold δ are called "δ-invalid" shocks and a running trail of k *consecutive* δ*-invalid* shocks can trigger a self-healing behavior. But, in this chapter, we have considered that healing occurs when the *cumulative* effect of k PIs nullify one valid shock. Furthermore, the PIs need not be δ-invalid. Thus, although our initial set-up is quite similar to that of Zhao *et al.* (2018b), our motivation and focus are quite different.

A practical example

The model described in this chapter applies to a practical maintenance problem, borrowed from Usynin and Hines (2007). Consider an Electronic Power Supply System made up of k components, numbered $1, 2, \cdots, k$. Sudden changes in temperature or abrupt vibrations are considered valid shocks that inflict cracks in printable circuit board causing equal amount

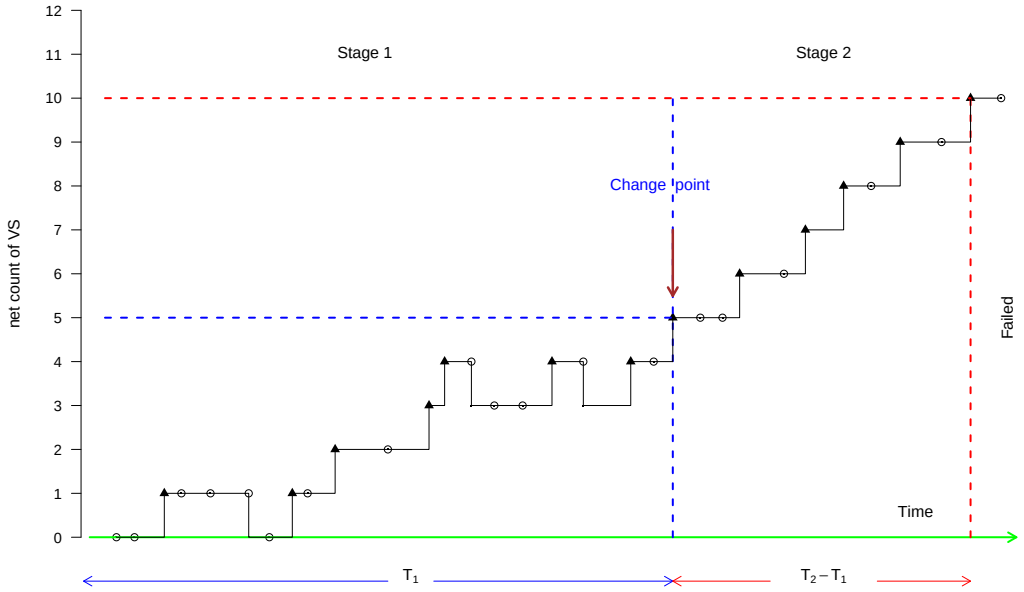


Figure 3.1. The arrival of VS (denoted by ▲) and PI (denoted by ∘) illustrate the net count of VS, and hence the stages. The change point defines transition from Stage 1 to Stage 2. Here, $k = 3$, $m_1 = 5$ and $m_2 = 10$. To understand when a healing occurs, count the PI's: Don't start counting until the first VS arrives. Stop counting PI if the net number of VS drops to 0. Resume counting once the next VS arrives. When the count reaches $k = 3$ a healing occurs. Here, $N_1 = 8, N_2 = 13$.

of "damage" to all k components. The VSs arrive according to a process with independent inter-arrival times distributed as F. On an ad-hoc basis, at random intervals inspections are carried out on these components cyclically; and the damage imposed by one VS is eliminated through repair. Consequently, when k repairs are completed cumulatively, the damage caused by one VS is nullified, and the net count of VS drops by one. If at any point, the net count of VS drops to zero, either no more inspection is made or those inspections do not affect the system. Only when another VS arrives, we resume the healing process. When the net count of VS for any component reaches m_1, healing stops and the system enters Stage 2.

3.3 Theoretical analysis: Distributions of T_1 and T_2

We present two approaches to describe the underlying stochastic process and compute the distributions of Stage 1 duration T_1 and system lifetime T_2.

3.3.1 The counting process

The following steps describe how the sporadic impacts are categorized as VS and PI, and how the net number of shocks is determined.

(A1) Generate a random sample from F yielding the inter-arrival times between successive VS. Similarly, we generate a random sample from G yielding the inter-arrival times between PI. We allow F and G to have *arbitrary* distributions.

(A2) Label VS as type 1 impact and PI as type 0 impact.

(A3) Concatenate the arrival times (cumulative sums of inter-arrival times) of the VS and the PI to form a vector of time points. Similarly, we concatenate type 1 and type 0 impacts into a vector of indicators.

(A4) Create a data frame consisting of two variables: a vector of arrival times of VS and PI, and a vector of indicators of types.

(A5) Sort the data frame with respect to arrival time, and carry along the indicators.

(A6) Start counting the net number of shocks as soon as the first VS arrives. The net count increases by one whenever a VS arrives. As soon as k PI accumulate, the net count decreases by one.

(A7) Stage 1 ends as soon as the net count reaches m_1. This epoch is called T_1, when the system experiences a change point and enters stage 2.

(A8) In Stage 2, there is no more healing. Therefore, the net count keeps on increasing as the VS arrive. Stage 2 ends and the system fails when the total number of shocks reaches m_2. This epoch is called T_2.

(A9) The goal is to find the distributions of Stage 1 duration T_1 and system lifetime T_2. To do so, first we find the distributions of the total number of VSs N_1, N_2 that the system receives in Stage 1 and throughout its lifetime, respectively. Since $N_2 = N_1 + m_2 - m_1$, it suffices to find the distribution of N_1.

(A10) We also find and record D_1, the total number of impacts (VS plus PI) in Stage 1, and D_2, the total number of impacts in Stage 1 and Stage 2 combined.

Next, repeat the above steps for a total of 10^4 iterations. By summarizing the computed random variables, one estimates their distributions.

3.3.2 The adjusted convolution process

The net number of shocks at any given time depends on how many VSs and how many PIs have arrived by that time. When the net count reaches m_1, the change point epoch T_1 is attained. Let the total number of VSs arriving during $(0, T_1]$ be denoted by N_1. In this subsection, we will demonstrate how the distribution of N_1 suffices to reconstruct the entire stochastic process described in Subsection 3.3.1 by obtaining the distributions of T_1 and T_2 starting from that of N_1. Let $S_j = \sum_{i=1}^{j} X_i$ be the arrival time of the j-th VS, and let $U_j = \sum_{i=1}^{j} Y_i$ be the arrival time of the j-th PI. Then, the duration of time the system stays in Stage 1 (T_1), can be found from the adjusted convolution described below in five steps.

(B1) The system receives N_1 VSs in Stage 1. The arrival time of the $(N_1 - 1)$-st VS is S_{N_1-1} and that of the N_1-th VS is S_{N_1}.

(B2) The arrival of which PI caused the last healing in Stage 1? Note that $(N_1 - m_1)$ VSs have been nullified using $(N_1 - m_1)k$ PIs. Letting r denote the number of PI that arrive whenever the system had no accumulated VS, the last healing happened with the arrival of the $[(N_1 - m_1)k + r]$-th PI.

(B3) After the last healing, the accumulated net VS reaches *at most* $(m_1 - 2)$. Thereafter, enough (*at least* two more) VSs came before k more PIs could come, and caused the end of Stage 1. See Figure 3.2. That is, between the last healing and the change point

T_1, the number of additional PIs that came is $h = 0, 1, 2, \ldots, k-1$; and these were not sufficient to cause another healing.

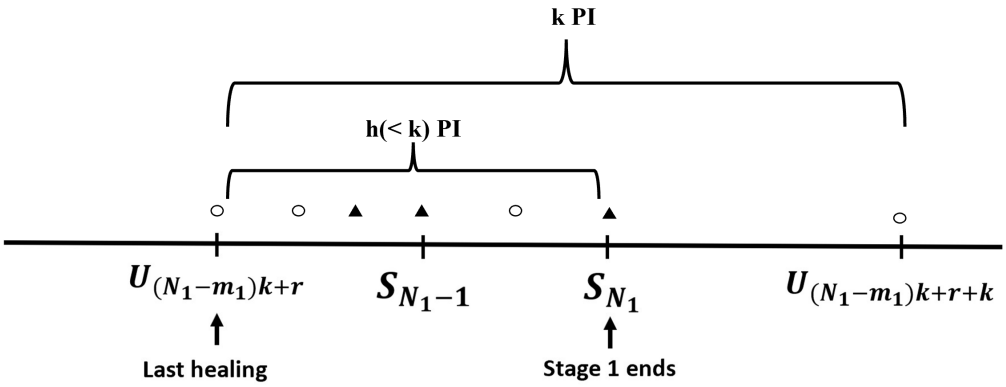


Figure 3.2. A schematic diagram to explain the arrival time of a PI causing the last healing in Stage 1. (A VS is denoted by ▲ and a PI by ∘). No healing occurs in Stage 2.

(B4) The last healing with the arrival of $[(N_1 - m_1)k + r]$-th PI must have happened before the $(N_1 - 1)$-st VS came. Hence, $U_{(N_1-m_1)k+r} < S_{N_1-1}$.

Thereafter, until the N_1-th VS arrives, fewer than k PI have arrived (for otherwise, another healing would have happened). Hence, $S_{N_1} < U_{(N_1-m_1)k+r+k}$.

Combining the two inequalities, we have

$$U_{(N_1-m_1)k+r} < S_{N_1-1} < S_{N_1} < U_{(N_1-m_1)k+r+k} \tag{3.3.1}$$

and the sojourn time in Stage 1 is

$$T_1 = S_{N_1} \tag{3.3.2}$$

(B5) What is the distribution of $T_1 = S_{N_1}$?

If T_1 and N_1 *were* independent, then we could write $T_1 \approx S_j$ with probability $P(N_1 = j)$. Note that S_j has distribution function given by $F*F*\cdots*F$, the j-fold convolution of F,

with $E[S_j] = j\,E[X]$. However, T_1 and N_1 are *not* independent, which motivates us to include an adjustment ϵ_j to the approximation $T_1 \approx S_j + \epsilon_j$ with probability $P(N_1 = j)$. Because N_1 is a stopping time (that determines the end of Stage 1), Wald's First Identity implies that $E[T_1] = E[N_1]\,E[X]$. To ensure the equality in expectation of the above approximation, we must have $E[N_1]\,E[X] \approx j\,E[X] + E[\epsilon_j]$. Hence, we take $\epsilon_j \propto (j - E[N_1])\,E[X]$. In other words, we approximate T_1 as

$$T_1 \approx S_j + \lambda\,(j - E[N_1])\,E[X] \text{ with probability } P(N_1 = j) \tag{3.3.3}$$

for $j \geq m_1$; where the proportionality constant λ depends on F and G. Note that for each choice of λ, the means of T_1 from the point process described in Subsection 3.3.1 and its approximation given in equation (3.3.3) are equal. But upon simulation the mean squared deviations from the mean for the approximation, when compared to that of T_1 from the point process, turns out to be an underestimate when $\lambda = 0$ and an overestimate when $\lambda = 1$. Therefore, to match both the mean and the mean squared deviation from the mean of the point process and the convolution process, we must choose a λ in the interval $(0, 1)$.

Choosing λ: For different combinations of inter-arrival time distributions F and G, we document the numerical values of λ, obtained via a grid search (with increment 0.01) to match the standard deviations of the distributions of T_1 obtained from the point process and the adjusted convolution process. As anticipated, λ is a function of the ratio of standard deviations of F and G.

In view of the above description of the stochastic process as a convolution with adjustment, we state the following two results:

Result 3.1. *The distribution of sojourn time in Stage 1 is approximately a weighted average of j-fold convolutions of F shifted by $\lambda\,(j - E[N_1])\,E[X]$, where $\lambda \in (0, 1)$, with weights given by $P(N_1 = j)$, for $j = m_1, m_1 + 1, \ldots$.*

Next, we obtain the distribution of the system lifetime. Since the system loses its healing property in Stage 2, the system lifetime equals the duration of Stage 1 *plus* $(m_2 - m_1)$

additional inter-arrival times of VS (which is the duration of Stage 2), the system lifetime equals

$$T_2 = T_1 + \sum_{i=N_1+1}^{N_1+m_2-m_1} X_i = S_{N_2}$$

where $N_2 = N_1 + m_2 - m_1$. Using the same justification as for equation (3.3.3), we approximate

$$T_2 \approx S_{j+m_2-m_1} + \lambda\,(j - E[N_1])\,E[X] \text{ with probability } P(N_1 = j) \tag{3.3.4}$$

for $j = m_1, m_1 + 1, \ldots$. In other words, we have a second result.

Result 3.2. *The distribution of time to failure (or system lifetime) T_2 is a weighted average of $(j + m_2 - m_1)$-fold convolution of F shifted by $\lambda\,(j - E[N_1])\,E[X]$, where $\lambda \in (0, 1)$, with weights given by $P(N_1 = j)$, for $j = m_1, m_1 + 1, \ldots$.*

3.3.3 Simulation results

Let us simulate the arrival process of the VS and the PI. We shall consider various different inter-arrival time distributions for X and Y satisfying $E[X] = 1$ and $E[Y] = 2/3$. In each iteration of the point process, we find the net count of VS and PI using description in Subsection 3.3.1; we also find the values of T_1, N_1, T_2 and N_2. By repeating such evaluations 10^4 times, we get a distribution of these random variables. For the adjusted convolution process described in Subsection 3.3.2, we utilize the above probability mass function (PMF) of N_1 together with the conditional probability density function (PDF) of S_j (for $j = m_1, m_1 + 1, \ldots$) constructed based on 100 randomly generated j-th convolution of F. Thereafter, we shift the j-th conditional PDF by $\lambda\,(j - E[N_1])\,E[X]$ and take the weighted average of these adjusted PDF's to find the overall PDF of T_1 and T_2 as described in Results 1 and 2, respectively, of Subsection 3.3.2.

We illustrate some useful details for one particular example: Let the $F \equiv$ Weibull (shape = 2, scale = $2/\sqrt{\pi}$) and $G \equiv$ gamma (shape = 2, scale = 1/3) such that $E[X] = 1$ and $E[Y] = 2/3$. The other cases are similar; hence, omitted. The PMF of N_1, estimated from the point process, is shown in Figure 3.3.

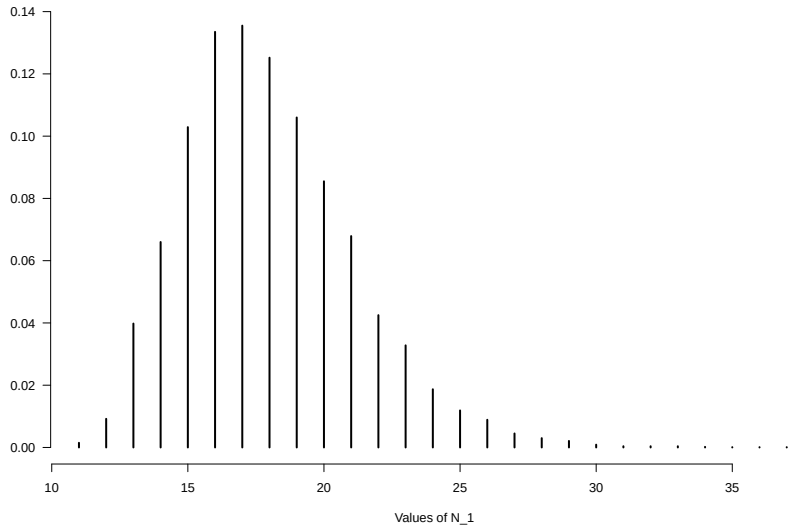


Figure 3.3. Probability distribution of N_1 is unimodal with $E[N_1] \approx 18$, $sd(N_1) = 3.14$, $Q_1 = 16$, $Q_2 = 18$, $Q_3 = 20$, $(N_1)_{0.99} = 27$, $P(N_1 > 30) = 0.0017$.

Figure 3.4 shows that the point process and the adjusted convolution process yield the same PDF for T_1. One measure of agreement between these two PDF's is given by the mean Kullback-Leibler divergence of the adjusted convolution PDF from the point process PDF measuring 0.0008128, which is very small (simulated p-value 0.999). Hence, we conclude that the densities arising from the point process and the adjusted convolution process are the same, supporting Result 1. Similarly, Figure 3.5 shows that the density plots for lifetime according to the two processes are the same, with a mean Kullback-Leibler divergence of 0.0008826 (simulated p-value 0.996), supporting Result 2. In the lower panels of Figures 3.4 and 3.5, we see that the maximum discrepancy between the approximated densities for each time point is at most 0.00015.

Table 3.1 displays the mean and the standard deviation of the sojourn time T_1 in Stage 1 for different combinations of inter-arrival time distributions of VS and PI. As anticipated (by choosing λ correctly), the mean and the standard deviation from the point process and the adjusted convolution process are almost the same. The very construction of Table 3.1

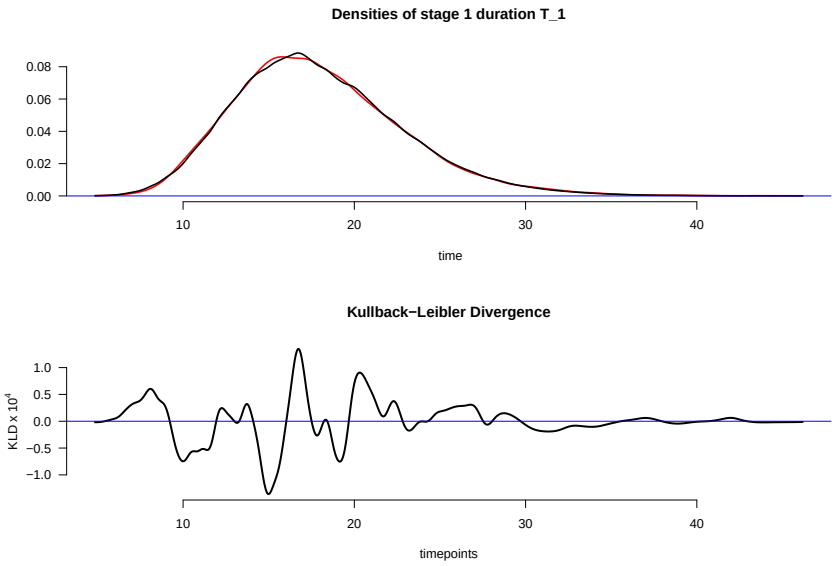


Figure 3.4. Densities of T_1 estimated from a point process (red) and an adjusted convolution process (black), with their absolute difference being at most 1.5×10^{-4}.

demonstrates that we can allow any distribution for the inter-arrival times of VS and PI, as long as we can identify the type of impetus so that we can track the net counts of VS and unused PI. Also, from Table 3.1 we can identify a trend in the values of λ as we scan through the rows and the columns. For a particular choice of F in a row, as we look from left to right across the columns, we see that λ decreases. A closer look at the corresponding standard deviations reveals that λ decreases as the standard deviation of G increases. Similarly, for a fixed choice of G in a particular column, as we go from top to bottom down the rows, we see that λ increases as the standard deviation of F increases. This led us to believe that λ is a function of the ratio of the standard deviations of F and G. When we plotted λ against σ_F/σ_G, we noticed a non-linear relationship. Thereafter, we fitted a linear regression of λ on $\log(\sigma_F/\sigma_G)$ and found an adjusted coefficient of determination of 0.9763. Therefore, we conjecture that $\lambda \approx (3/8) + (\pi/20)\log(\sigma_F/\sigma_G)$. The search for optimal λ can be restricted to a small neighborhood of this approximate value.

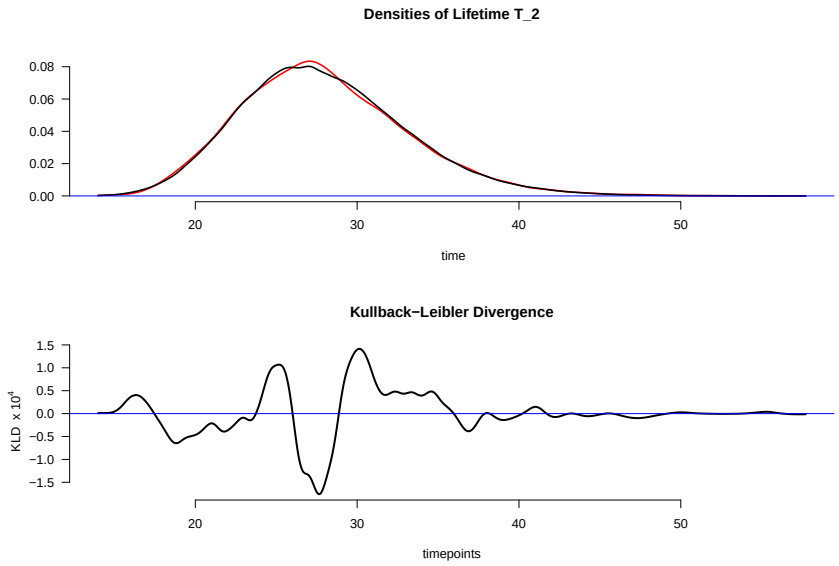


Figure 3.5. Densities of T_2 estimated from a point process (red) and an adjusted convolution process (black), with their absolute difference being at most 1.5×10^{-4}.

Next, we display a similar Table 3.2 for the system lifetime T_2. There is no need to report the same λ already shown in Table 3.1. Again, as anticipated, the mean and the standard deviation from the two processes agree well.

Whereas previous research illustrate only exponential inter-arrival times of VS and PI, we have incorporated *arbitrary* inter-arrival times for VS and PI. Thus, we have extended the application of shock models in reliability maintenance research.

3.4 Preventive maintenance policies

Identifying the distributions of T_1 and T_2 is an important achievement towards making optimum decisions during system maintenance. Taking a cue from Zhao *et al.* (2018b), we consider three maintenance policies. We allow different costs of system maintenance in the two Stages: But unlike Zhao *et al.* (2018b), we consider that the cost of failure replacement is much higher than that of a preventive replacement in Stage 2, which is higher than that in

Table 3.1. For various inter-arrival time distributions satisfying $E[X] = 1$ and $E[Y] = 2/3$, the top entries give mean (standard deviation) of T_1 according to a point process and middle row entries (in *italics*) give the same quantities according to an adjusted convolution process. The third row gives the multiplier λ of the adjustment term.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $sd \approx 0.12$	gamma $(2, \frac{1}{3})$ $sd \approx 0.22$	inv-Gauss $(2/3)$ $sd \approx 0.29$	exponential $(3/2)$ $sd \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $sd \approx 0.27$	17.96 (4.75) *17.95 (4.72)* $\lambda = 0.50$	17.97 (4.96) *17.97 (4.96)* $\lambda = 0.40$	17.98 (5.12) *17.97 (5.10)* $\lambda = 0.35$	17.93 (5.41) *17.93 (5.41)* $\lambda = 0.29$
gamma $(2, \frac{1}{2})$ $sd \approx 0.50$	17.92 (6.06) *17.92 (6.06)* $\lambda = 0.59$	17.91 (6.26) *17.93 (6.25)* $\lambda = 0.50$	17.93 (6.35) *17.94 (6.34)* $\lambda = 0.46$	17.85 (6.58) *17.86 (6.56)* $\lambda = 0.38$
inv-Gauss (1) $sd \approx 1.00$	17.63 (8.02) *17.74 (8.01)* $\lambda = 0.68$	17.61 (8.10) *17.71 (8.07)* $\lambda = 0.61$	17.67 (8.25) *17.75 (8.21)* $\lambda = 0.58$	17.61 (8.33) *17.72 (8.31)* $\lambda = 0.52$
exponential (1) $sd \approx 1.00$	17.85(8.48) *17.83 (8.28)* $\lambda = 0.69$	17.87 (8.39) *17.87 (8.39)* $\lambda = 0.63$	17.92 (8.58) *17.91 (8.58)* $\lambda = 0.60$	17.82 (8.55) *17.79 (8.57)* $\lambda = 0.54$

Stage 1. Let c_0 be the cost of intitial installation of the system, c_{p_1} be the cost of replacement in Stage 1, c_{p_2} be that of Stage 2, and c_f be the cost of failure replacement, satisfying $c_{p_1} \leq c_{p_2} \leq c_f$. This is because we believe that in Stage 1, the system is young, therefore early replacement would incur a relatively smaller cost; furthermore, if we replace in early Stage 2, even then we are not utilizing the system lifetime enough and on the other hand, the system has already aged, which means maintenance/repair at this stage would cost more. Therefore, we find it logical to consider that replacement cost in Stage 2 is higher than that in Stage 1. Furthermore, there is an initial cost c_0 of setting up a new system. For illustration, we choose $c_0 = 100, c_{p_1} = 10, c_{p_2} = 15, c_f = 200$.

One could also consider differential revenues earned per unit time when the system operates. However, for the sake of simplicity, we assume that revenue is earned at a constant

Table 3.2. For various inter-arrival time distributions satisfying $E[X] = 1$ and $E[Y] = 2/3$, the top entries give mean (standard deviation) of system lifetime T_2 according to a point process, and the bottom entries (in *italics*) show the same quantities according to an adjusted convolution process.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$	gamma $(2, \frac{1}{3})$	inv-Gauss (2/3)	exponential (3/2)
Weibull $(2, \frac{2}{\sqrt{\pi}})$	27.94 (5.03) *27.97 (5.06)*	27.96 (5.24) *27.98 (5.23)*	27.96 (5.39) *27.99 (5.37)*	27.92 (5.65) *27.86 (6.95)*
gamma $(2, \frac{1}{2})$	27.91 (6.47) *27.91 (6.46)*	27.92 (6.67) *27.93 (6.63)*	27.93 (6.73) *27.94 (6.71)*	27.86 (6.95) *27.85 (6.94)*
inv-Gauss (1)	27.57 (8.63) *27.67 (8.64)*	27.58 (8.73) *27.66 (8.65)*	27.60 (8.85) *27.69 (8.83)*	27.54 (8.94) *27.63 (8.93)*
exponential (1)	27.86 (8.83) *27.82 (8.90)*	27.90 (8.97) *27.85 (9.02)*	27.93 (9.12) *27.88 (9.2)*	27.82 (9.07) *27.78 (9.16)*

rate throughout the lifetime of the system. Thus, we focus on minimizing the maintenance cost per unit time.

3.4.1 Maintenance policy 1

Suppose that the monitoring equipment can detect the arrival of an impetus, but it cannot distinguish between a VS and a PI, nor can it identify whether the system has transitioned from Stage 1 to Stage 2. The system will be replaced when it has failed or has experienced a specified number of impetus N (the sum of VS and PI).

Within one cycle (between two successive replacements of the system), the total cost of replacement under Policy 1 is a random variable taking three possible values:

(1) c_{p_1}, if N impetus arrive while the system is still in Stage 1, with an associated probability of $P(D_1 > N)$.

(2) c_{p_2}, if the system has already moved to Stage 2 and the N impetus have arrived before the system fails, with an associated probability of $P(D_1 \leq N < D_2)$.

(3) c_f, if the system has already failed before the arrival of N impetus, with an associated probability of $P(D_2 \leq N)$.

Hence, the expected cost (C) and the expected cycle time (CT) under Policy 1 are given by

$$\begin{aligned} E[C \mid \text{Policy 1}] = c_{p_1} P(D_1 > N) + c_{p_2} P(D_1 \leq N < D_2) \\ + c_f P(D_2 \leq N) \end{aligned} \tag{3.4.1}$$

and, writing W_j as the arrival time of the j-th impact (either VS or PI), we have

$$\begin{aligned} E[CT \mid \text{Policy 1}] &= E[\min\{W_N, W_{D_2}\}] \\ &= E[W_N \mid D_1 > N]\ P(D_1 > N) \\ &+ E[W_N \mid D_1 \leq N < D_2]\ P(D_1 \leq N < D_2) \\ &+ E[W_{D_2} \mid D_2 \leq N]\ P(D_2 \leq N) \end{aligned} \tag{3.4.2}$$

Therefore, the expected cost per unit time is the ratio

$$E[C \mid \text{Policy 1}]/E[CT \mid \text{Policy 1}] \tag{3.4.3}$$

which we must minimize by choosing N. Under Policy 1 and the assumed cost structure, Figure 3.6 shows that the expected cost per unit time is minimized when we choose $N = 53$. Moreover, note that for any other choice of N in the vicinity of the optimal value 53, say between 48 and 55, the expected cost per unit time increases only slightly (no more than 3%). Such a robustness result allows us to rely on the optimal value even when the inter-arrival time distributions deviate slightly from the stated ones. Table 3.3 documents the optimal choices of N for other combinations of inter-arrival times, together with the optimal results of the other policies.

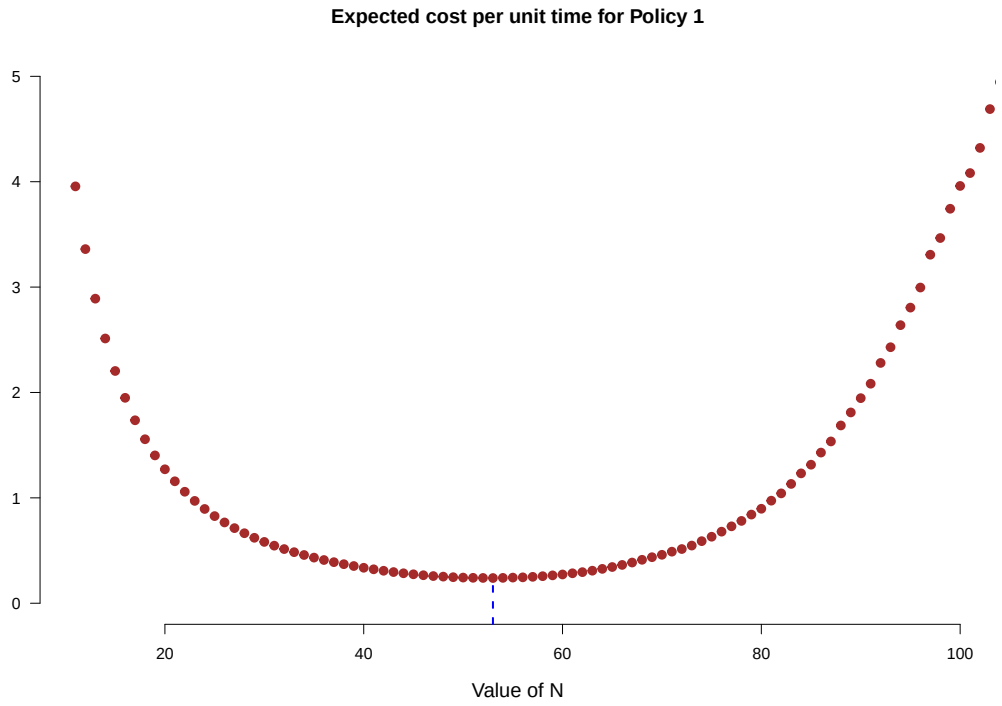


Figure 3.6. Under Policy 1 and cost parameters $c_0 = 100$, $c_{p_1} = 10$, $c_{p_2} = 15$, $c_f = 200$, the optimal number of impetus for preventive replacement is $N = 53$.

3.4.2 Maintenance policy 2

Suppose that the monitoring equipment can identify the stages of the system. If the system is in Stage 1, we do not replace it at all. After the system enters Stage 2, if the system is still functioning for an additional t units of time, we replace it immediately at epoch $T_1 + t$; otherwise, we replace the system immediately if it fails before the additional time t. Note that once the optimal t is determined, such a policy is logistically more convenient than waiting to replace the system at a random time when the $(N_2 - 1)$-st VS arrives. Our objective is to determine an optimum additional time t in Stage 2 when the system should be replaced. To do so, we minimize the expected cost per unit time, where the expected cost (C) is given by

$$E[C \mid \text{Policy 2}] = c_{p_2} P(T_2 > T_1 + t) + c_f P(T_2 \leq T_1 + t) \tag{3.4.4}$$

and the expected length of the cycle time (CT) is

$$E[CT \mid \text{Policy 2}] = E[min(T_2, T_1 + t)] = E[T_1] + E[min(T_2 - T_1, t)] \tag{3.4.5}$$

We wish to minimize the expected cost per unit time

$$E[C \mid \text{Policy 2}]/E[CT \mid \text{Policy 2}] \tag{3.4.6}$$

by choosing t. Under Policy 2 and the assumed cost parameters, Figure 3.7 shows that the expected cost per unit time is minimized when we choose $t = 6.55$. In fact, we identified this optimal t value via a grid search between the first and the 99-th percentiles of system lifetime with an increment of 0.05. This choice suffices because any other choice of t in the interval $[6, 7]$ increases the cost per unit time only marginally. Table 3.3 documents the optimal choices of t for other combinations of inter-arrival times.

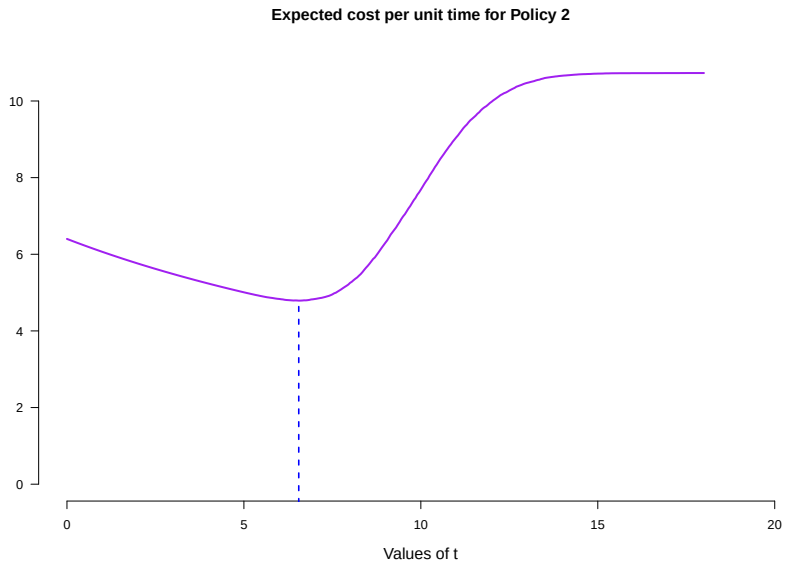


Figure 3.7. Using cost parameters $c_0 = 100$, $c_{p_1} = 10$, $c_{p_2} = 15$, $c_f = 200$, the optimal duration in Stage 2 after which preventive replacement must be scheduled is $t = 6.55$.

3.4.3 Maintenance policy 3

Suppose that the system state can be identified and it is possible to detect system failure immediately when it occurs. However, the system is not under continuous monitoring. Instead, at a cost of c_I per inspection, scheduled inspections will be conducted at inter-inspection intervals $t_1 > t_2 > t_3 > \ldots$, until the system is found to be in Stage 2, when the system will be preemptively replaced immediately. Finally, if the system is found to be still in Stage 1 after the completion of a (predetermined) number of inspections u, at time $v_u = \sum_{i=1}^{u} t_i$, then it will be replaced at the $(u+1)$-st inspection at time $v_{u+1} = v_u + t_{u+1}$, whether or not the system enters Stage 2.

If inspections start too early or if they occur in quick successions, we will need too many inspections costing too much and/or not utilize the system long enough at the preventive replacement time v_{u+1}. On the other hand, if the inspections begin late or if they are further apart, then we run a high risk of a costly system failure before preventive replacement. To balance the two extremes, we should not consider the same u number of inspections before preventive replacement irrespective of how the inspection times are scheduled. To obtain an optimum u, we have considered the following criteria: First, we allow the inspection times to be spread out in a geometric series starting with t_1 (to be determined) and thereafter taking $t_{i+1} = \alpha^i \; t_1$ for $i \geq 1$, where we allow a fixed $\alpha \in (0, 1)$, since more frequent inspections are preferred as the system ages. Then, we select u such that

$$v_u = \sum_{i=1}^{u} t_i = t_1 \frac{1-\alpha^u}{1-\alpha} \leq (T_1)_{0.95}, \tag{3.4.7}$$

the 95-th percentile of the Stage 1 sojourn time T_1. Thus, for different choices of the starting time of the first inspection t_1, we get different optimal choices of the required number of inspections $(u+1)$. Algorithm 1 computes the cost and the cycle time under Policy 3.

Algorithm 1: Finding cost and cycle time

Result: cost, cycle time

Initialize:

a = 1; cost = 0; cycle time = 0 ;

while $a < u$ **do**

  if $T_1 \leq \sum_{i=1}^{a} t_i < T_2$ **then**

    cost = $c_0 + c_{p_2} + (a \times c_I)$;

    cycle time = $\sum_{i=1}^{a} t_i$

    break

  end

  if $\sum_{i=1}^{a} t_i \geq T_2$ **then**

    cost = $c_0 + c_f + (a \times c_I)$;

    cycle time = T_2

    break

  else

    a = a+1;

  end

end

if $a = u$ **then**

  if $\sum_{i=1}^{a} t_i < T_1$ **then**

    if $\sum_{i=1}^{a+1} t_i < T_1$ **then**

      cost = $c_0 + c_{p_1} + ((a+1) \times c_I)$;

      cycle time = $\sum_{i=1}^{a+1} t_i$

    end

    if $T_1 \leq \sum_{i=1}^{a+1} t_i < T_2$ **then**

      cost = $c_0 + c_{p_2} + ((a+1) \times c_I)$;

      cycle time = $\sum_{i=1}^{a+1} t_i$

    else

      cost =$c_0 + c_f + ((a+1) \times c_I)$;

      cycle time = T_2

    end

  end

  if $T_1 \leq \sum_{i=1}^{a} t_i < T_2$ **then**

    cost = $c_0 + c_{p_2} + (a \times c_I)$;

    cycle time = $\sum_{i=1}^{a} t_i$

  else

    cost = $c_0 + c_f + (a \times c_I)$;

    cycle time = T_2

  end

end

The expected cost (C_u) under Policy 3 is

$$\begin{aligned} E[C_u] = {} & c_{p1} P(v_{u+1} < T_1) \\ & + c_{p2}\{P(v_u < T_1 \le v_{u+1} < T_2) + P(T_1 \le v_u < T_2)\} \\ & + c_f\{P(v_u < T_1 < T_2 \le v_{u+1}) + P(v_{u-1} < T_1 < T_2 \le v_u)\} \end{aligned} \tag{3.4.8}$$

Similarly, the expected cycle time (CT_u) is

$$\begin{aligned} E[CT_u] = {} & v_{u+1}\{P(v_{u+1} < T_1) + P(v_u < T_1 \le v_{u+1} < T_2)\} \\ & + v_u P(T_1 \le v_u < T_2)\} \\ & + T_2\{P(v_u < T_1 < T_2 \le v_{u+1}) + P(v_{u-1} < T_1 < T_2 \le v_u)\} \end{aligned} \tag{3.4.9}$$

We must choose t_1 (and the associated u) to minimize the expected cost per unit time $E[C_u]/E[CT_u]$ under Policy 3. In our numerical example, together with $c_I = 5$ and $\alpha = 0.95$, we conducted a grid search for t_1 over the interval $[2.5, 15.5]$ at increments of 0.05. Figure 3.8 indicates that the optimal time of first inspection is $t_1 = 9.3$, and the system must be replaced, even though it has not failed, after the fourth inspection.

3.5 Summary

Having shown the details of the optimal choices for the special case of $F \equiv$ Weibull (shape $= 2$, scale $= 2/\sqrt{\pi}$) and $G \equiv$ gamma (shape $= 2$, scale $= 1/3$), for the sake of brevity, in Table 3.3 we simply document the optimal choices of N, t and t_1 (u) for Policy 1, Policy 2 and Policy 3, respectively, for the different combinations of F and G considered in the previous illustrations, and the above mentioned choices of the cost parameters. From the table we see that for Policy 1, the optimal N lies between 48 to 55; for Policy 2, the optimal t is somewhere between 4 and 7, and for Policy 3, the optimum t_1 is between 8 to 10, with the corresponding number of inspections u roughly 3 or 4.

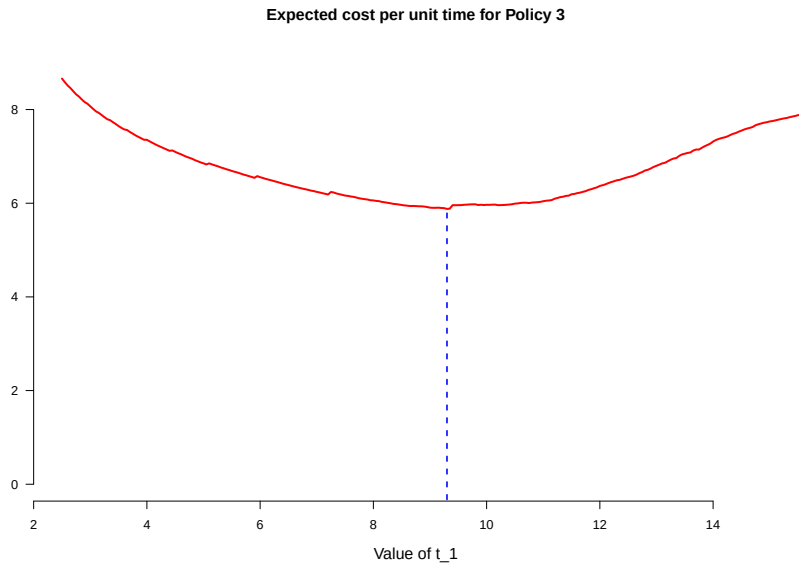


Figure 3.8. Using $c_0 = 100, c_{p_1} = 10, c_{p_2} = 15, c_f = 200$, $c_I = 5$, and $\alpha = 0.95$, the optimal time for the first inspection is $t_1 = 9.3$ and the associated $u = 3$.

Table 3.3. For various inter-arrival time distributions satisfying $E[X] = 1$ and $E[Y] = 2/3$ and cost parameters $c_0 = 100, c_{p_1} = 10, c_{p_2} = 15, c_f = 200$, $c_I = 5$, and inter-inspection duration factor $\alpha = 0.95$, the first row gives the optimal value of N for Policy 1, the second row gives the optimal t for Policy 2 and the third row gives the optimal t_1 and the associated u (in parenthesis) for Policy 3, for every choice of (F, G).

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $sd \approx 0.12$	gamma $(2, \frac{1}{3})$ $sd \approx 0.22$	inv-Gauss $(2/3)$ $sd \approx 0.29$	Exponential $(3/2)$ $sd \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $sd \approx 0.27$	$N = 54$ $t = 6.40$ $t_1 = 9.20$ (3)	$N = 53$ $t = 6.55$ $t_1 = 9.30$ (3)	$N = 52$ $t = 6.70$ $t_1 = 9.30$ (3)	$N = 50$ $t = 6.55$ $t_1 = 9.75$ (3)
gamma $(2, \frac{1}{2})$ $sd \approx 0.50$	$N = 52$ $t = 6.05$ $t_1 = 8.85$ (3)	$N = 51$ $t = 5.95$ $t_1 = 9.60$ (3)	$N = 51$ $t = 5.90$ $t_1 = 9.60$ (3)	$N = 49$ $t = 5.85$ $t_1 = 9.35$ (3)
inv-Gauss (1) $sd \approx 1.00$	$N = 49$ $t = 5.00$ $t_1 = 8.10$ (4)	$N = 50$ $t = 5.00$ $t_1 = 8.50$ (4)	$N = 48$ $t = 4.90$ $t_1 = 8.60$ (4)	$N = 47$ $t = 4.85$ $t_1 = 8.75$ (4)
exponential (1) $sd \approx 1.00$	$N = 51$ $t = 4.95$ $t_1 = 8.05$ (4)	$N = 48$ $t = 4.95$ $t_1 = 8.55$ (4)	$N = 50$ $t = 4.65$ $t_1 = 8.65$ (4)	$N = 48$ $t = 4.60$ $t_1 = 8.75$ (4)

4. COST MINIMIZATION UNDER SPORADIC SHOCKS AND HEALING IMPETUS WHEN THE HEALING STAGE IS SUBDIVIDED

Content in the following chapter was previously published by Society of Statistics, Computer and Applications (SSCA) in the proceedings of their 23rd Annual Conference. 2021, May: https://ssca.org.in/media/2_Spl_Proceedings_2021_Jyotirmoy_Sarkar_280521_Final.pdf.

Debolina Chatterjee and Jyotirmoy Sarkar are co-authors of the published work.

4.1 Introduction

In machine maintenance and reliability engineering, it is often necessary to study the impacts of external shocks. In addition to degradation due to aging, system lifetime is affected by the accumulated damage due to shocks. Because a system failure causes a severe loss, it is preferable to replace a system before it fails, but only after utilizing its potential life to the extent possible. Therefore, we seek optimal replacement policies before the system fails.

In Chapter 3, two types of impacts — valid shocks (VS) and positive interventions (PI) — are considered, with their inter-arrival times having *arbitrary* distributions, and the system lifetime is split into two stages — Stage 1 where it can heal, and Stage 2 where it cannot heal. However, healing occurs when the *cumulative* effect of k PIs (not necessarily consecutive) nullify one VS. Furthermore, the PIs need not be δ-invalid. This continues until the system reaches a "change point" beyond which it can no longer heal.

The main focus of the current work is to extend the two-stage model by splitting Stage 1 further into two parts. Initially, when the system is young, healing can happen faster; but later, when the system has aged and has experienced several shocks, healing is slower. In the earlier part of Stage 1, called Stage 1A, k_A PIs nullify the damaging effect of one VS. In the later part of Stage 1, called Stage 1B, k_B ($> k_A$) PIs can heal one VS. The system is in Stage 1A until the net VS reaches a threshold m_A; thereafter, it enters Stage 1B. Next, the

system reaches the change point and enters Stage 2 when the net VS reaches m_1. Therefore, $m_1 - m_A = m_B$ is the net number of VS allowed in Stage 1B.

Previous research considered either healing or degradation, or they have assumed that the shocks/impacts have inter-arrival times exponentially distributed. Although some works mention non-exponential inter-arrival times, they illustrate only exponential examples. As in Chapter 3, here we illustrate with several non-exponential inter-arrival time distribution. As long as we can count the number of VS and PI, we can determine the distributions of duration of Stage 1 and system lifetime.

Section 4.2 describes the evolution of the system under shocks and healing; Section 4.3 illustrates two approaches to calculate the distributions of Stage 1 duration and lifetime; Section 4.4 compares Stage 1 duration and system lifetime between divided versus undivided Stage 1; Section 4.5 obtains optimal decisions for two maintenance policies; and finally Section 4.6 summarizes the main findings of this research.

4.2 The system set up

External impacts to the system are of two types: Valid Shocks (VS) that cause damage to the system and Positive Interventions (PI) that do not have any damaging effect; on the contrary, the accural of a certain (predetermined) number of PIs nullify the effect of one VS. This behaviour is what we call "healing", which means the net number of VS (VS arrived minus VS nullified by PIs) reduces by one. For simplicity, we assume each VS causes an equal amount of damage. Hence, leaving for future the study of magnitude of damages, we focus on counting the net number of VS to the system.

The lifetime of the system is divided into two stages depending on the net VS it receives. In Stage 1, the system has healing ability as described above and the system remains in this stage until the net VS reaches a certain predetermined threshold m_1. Thereafter, the system moves to Stage 2 where it can no longer heal; that is, new PIs no longer reduce net VS. The system fails when the net VS reaches another higher threshold m_2. Furthermore, Stage 1 is subdivided into two parts: Stage 1A requires fewer and Stage 1B requires larger number of PI's to nullify one VS.

The inter-arrival times of VS, denoted by $\{X_i; i \in \mathbb{N}\}$ are independently and identically distributed (IID) with an *arbitrary* cumulative distribution function (CDF) F. Likewise, $\{Y_j; j \in \mathbb{N}\}$, the inter-arrival times of PIs are IID with another *arbitrary* CDF G. The arrival processes of PIs and VS are stochastically independent. Let the duration of system in Stage 1 be denoted as T_1 and the system lifetime be denoted as T_2. The total number of VS in Stage 1 be N_1 and that until failure be N_2. Note that $m_2 - m_1 = N_2 - N_1$ since in Stage 2 there is no healing. Let r denote the number of PIs rendered unused towards healing in Stage 1. Furthermore, let D_1 and D_2 denote the total number of impacts (VS+PI) in Stage 1 and until failure, respectively.

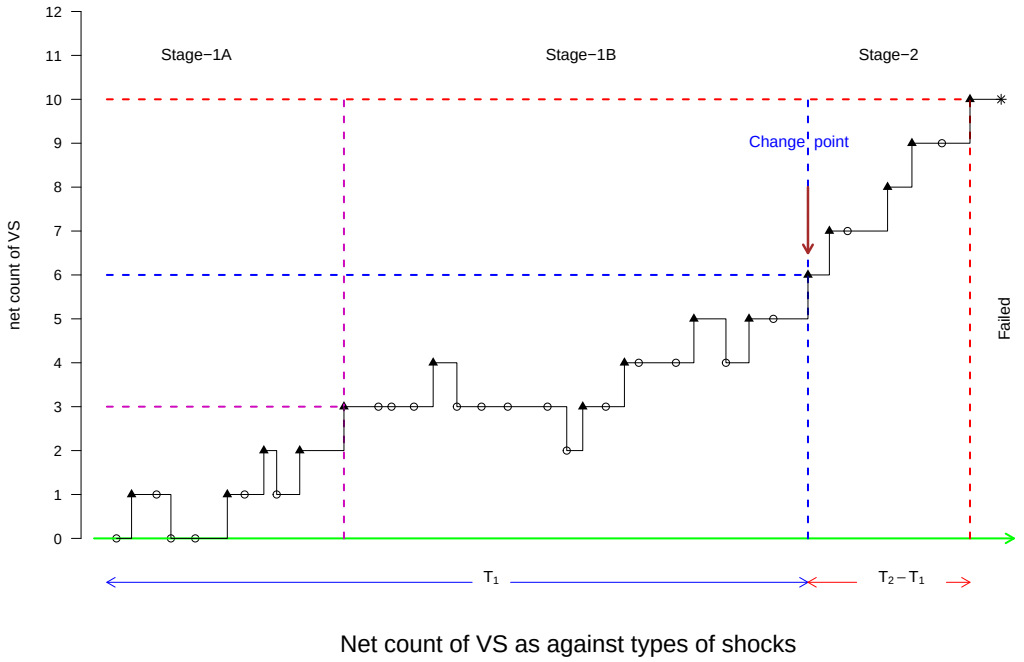


Figure 4.1. The arrival processes of VS (denoted by ▲) and PI (denoted by ○) illustrate the net count of VS, and hence the stages. Here, $k_A = 2, k_B = 4, m_A = 3, m_1 = 6$, and $m_2 = 10$. Do not start counting until the first VS arrives. Stop counting PI if the net number of VS, drops to 0. Resume counting once the next VS arrives. The *change point* T_1 defines the transition from Stage 1 (Stages 1A and 1B combined) to Stage 2.

4.3 Theoretical analysis: Distributions of T_1 and T_2

We describe the underlying stochastic process in terms of two approaches: a counting process approach and a convolution process approach.

4.3.1 The counting process

Given the constant integers k_A, k_B, m_A, m_1 (hence, $m_B = m_1 - m_A$) and m_2, we describe a simulation of the system status as follows.

Generate a sequence of inter-arrival times of VS $\{X_i; i \in \mathbb{N}\}$ IID with CDF F, and another sequence of inter-arrival times of PIs $\{Y_j; j \in \mathbb{N}\}$ IID with CDF G. Take the cumulative sums of the two sequences, to obtain the arrival times. Sort these arrival times of the impacts (VS and PI) and associate with each arrival time an indicator **1** to denote VS and **0** to denote PI. Start counting as soon as the first VS arrives. Ignore all PIs (0) before this moment.

Stage 1A: Count the VS. Arrival of k_A PIs nullify one VS. Compute the net VS as the VS arrived minus VS nullified. If the net VS ever drops to 0, stop counting; resume counting when again another VS arrives. When net VS reaches m_A, the system enters Stage 1B. We keep record of the total number of VS that arrived in Stage 1A, namely N_A.

Stage 1B: In this later part of Stage 1, arrival of a VS increases its count by one, but now to nullify one VS we need k_B PIs ($k_B > k_A$). Again, we stop counting if the net count ever drops down to 0; and resume counting when a new VS arrives. When the net VS reaches m_1, the system enters Stage 2. Let N_B denote the total number of VS that arrive Stage 1B, and let r denote the number of PIs that have arrived in Stage 1.

Stage 2: In this stage, the system does not heal. The VS keeps accumulating one by one without being nullified since the PIs have no effect. The system fails when the net VS reaches a threshold m_2.

Thus, in one iteration of the simulation, we obtain as outputs the following quantities: $N_1 = N_A + N_B$, N_2, T_1, T_2, r, and $D_1 = N_1 + r$, which is the total number of impacts in Stage 1. Next, we repeat the above steps for a total of 10^4 iterations.

We will approximate the probability mass functions (PMF) of N_1 and N_2 from the relative frequencies observed in the simulation. Also, based on the simulation, we will directly

approximate the probability density function (PDF) of T_1 and T_2. Alternatively, we will reconstruct these PDFs using the PMFs of N_1 and N_2, respectively, through a convolution process as explained below.

4.3.2 The adjusted convolution process

The underlying stochastic process is described below:

In Stage 1A, $(N_A - m_A)$ VS have been nullified by the arrival of $(N_A - m_A) * k_A$ PIs. Similarly in Stage 1B, $(N_B - m_B)$ VS have been nullified by the arrival of $(N_B - m_B) * k_B$ PI and there may have arrived h more PIs, where $0 \leq h \leq k_B - 1$, which are insufficient to nullify another VS. Therefore, in total, the arrival of $Q = (N_A - m_A) * k_A + (N_B - m_B) * k_B$ PIs has contributed to nullifying $(N_1 - m_1)$ PIs.

Let us denote $S_j = \sum_{i=1}^{j} X_i$ as arrival time of the j-th VS, and $U_j = \sum_{i=1}^{j} Y_i$ as arrival time of the j-th PI. We describe how Stage 1 duration T_1 depends on the number of VS N_1.

(1) The system receives N_1 VS in Stage 1. The arrival time of the $(N_1 - 1)$-st VS is S_{N_1-1} and that of the N_1-th VS is S_{N_1}.

(2) Let U_Q be the arrival time of a PI which causes the $(N_1 - m_1)$-th nullification of a VS, and let U_{Q+h} be the arrival time of the $(Q + h)$-th PI, which do not nullify any VS (where $h = 1, 2, \ldots, k_2 - 1$), since the count of unused PI has not reached k_B yet.

(3) Before the $(N_1 - 1)$-st VS arrives, the Q-th PI has already arrived and $Q + h$-th PI must have happened. Hence, $U_{Q+h} < S_{N_1-1}$. Thereafter, until the N_1-th VS arrives, fewer than k_B PI have arrived in Stage 1B. Hence, $S_{N_1} < U_{Q+h+k_2}$. Therefore, the arrival times satisfy the inequality

$$U_{Q+h} < S_{N_1-1} < S_{N_1} < U_{Q+h+k_2} \tag{4.3.1}$$

and the sojourn time in Stage 1 is

$$T_1 = S_{N_1} \tag{4.3.2}$$

Note that S_j has a CDF given by $F * F * \cdots * F$, the j-fold convolution of F. Furthermore, since N_1 is a random stopping time (that determines the end of Stage 1), by Wald's first identity, we have $E[T_1] = E[N_1] \times E[X]$. However, N_1 and T_1 are not independent. Therefore, using a second-order approximation (by matching the mean and the mean squared deviation from the mean), we model

$$T_1 = S_j + \lambda\,(j - E[N_1])\,E[X] \text{ with probability } P(N_1 = j) \tag{4.3.3}$$

for $j = m_1, m_1 + 1, \ldots$; where $\lambda \in [0, 1]$ depends on F and G. That is, the distribution of T_1 is modeled as a weighted average of adjusted j-fold convolutions of F, where the adjustment equals a suitable fraction of the departure of N_1 from its expectation *times* the expected inter-arrival time between shocks, with weights given by the probability masses $P(N_1 = j)$ for $j = m_1, m_1 + 1, \ldots$.

The above explanations justify the following results:

Result 4.1. *The distribution of Stage 1 duration is a weighted average of j-fold convolutions of F shifted by $\lambda\,(j - E[N_1])\,E[X]$, where $\lambda \in (0, 1)$ is described below, with weights given by $P(N_1 = j)$, the probability that N_1 VS arrive in Stage 1, for $j = m_1, m_1 + 1, \ldots$.*

Description of λ: For several combinations of inter-arrival time distributions F and G, the fraction λ is numerically obtained via a grid search (with increment 0.01) to match the standard deviations of the distribution of T_1 obtained from the point process and the convolution process.

The lifetime of the system is equal to the duration of Stage 1 *plus* $(m_2 - m_1)$ additional inter-arrival times of VS (which is the duration of Stage 2), since the system can no longer heal in Stage 2. Hence, the system lifetime is

$$T_2 = T_1 + \sum_{i=N_1+1}^{N_1+m_2-m_1} X_i = S_{N_2}$$

where $N_2 = N_1 + m_2 - m_1$. Using Result 4.1, we can describe

$$T_2 = S_{j+m_2-m_1} + \lambda\,(j - E[N_1])\,E[X] \text{ with probability } P(N_1 = j) \tag{4.3.4}$$

for $j = m_1, m_1 + 1, \ldots$.

Result 4.2. *The distribution of time to failure T_2 is a weighted average of $(j + m_2 - m_1)$-fold convolution of F shifted by $\lambda(j - E[N_1])E[X]$, where $\lambda \in (0, 1)$, with weights given by $P(N_1 = j)$, for $j = m_1, m_1 + 1, \ldots$.*

It is noteworthy that the above results are exactly the same as in Chapter 3. The intuition behind it is that both the results involve the PMF of N_1, the total number of VS in Stages 1A and 1B combined. Exactly how Stage 1 is subdivided (based on the requirements of healing) is irrelevant to describe the Stage 1 duration and the lifetime.

4.3.3 Simulation results

We shall consider different inter-arrival time distributions for X and Y satisfying $E[X] = 1$ and $E[Y] = 2/3$. The distribution of sojourn time in Stage 1 is found directly by repeating the point process 10^4 times. In Chapter 3, we had considered $k = 3$ and $m_1 = 10$ for illustration. For comparability, here we choose $k_A = 2, k_B = 4$ to keep the overall average number of PIs required to nullify one VS roughly the same, and we choose $m_A = 5, m_B = 5$ so that $m_1 = 10$.

We emphasize that while similar results hold for all combinations of inter-arrival times, to save space, we will show detailed results for one particular combination of inter-arrival times: $F \equiv$ Weibull (shape $= 2$, scale $= 2/\sqrt{\pi}$) and $G \equiv$ gamma (shape $= 2$, scale $= 1/3$), such that $E[X] = 1$ and $E[Y] = 2/3$. Figure 4.3 shows the simulated PMF of N_1.

Figure 4.3 shows the PDF of T_1 obtained both directly from the point process and from the adjusted convolution. One measure of agreement between these two PDF's is given by the mean Kullback-Leibler divergence of the adjusted convolution PDF from the point process PDF measuring 0.001466, which is very small (simulated p-value 0.997). Hence, we conclude that the densities arising from the point process and the adjusted convolution process are the same, supporting Result 4.1.

Similarly, Figure 4.4 shows that the density plots for lifetime according to the two processes are the same, with a mean Kullback-Leibler divergence of 0.00102 (simulated p-value 0.999),

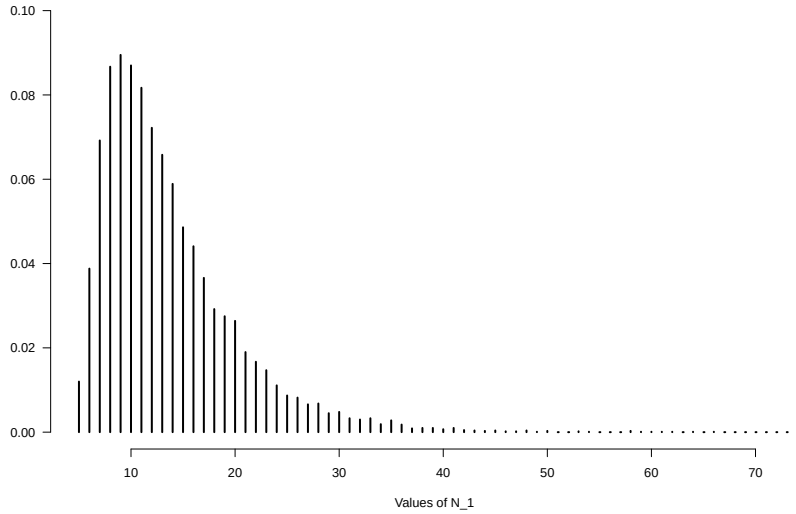


Figure 4.2. Probability distribution of N_1 is unimodal with mode 17, $E[N_1] \approx 21$, $sd(N_1) = 6.52$, $Q_1 = 17$, $Q_2 = 20$, $Q_3 = 24$, 99-th percentile $(N_1)_{0.99} = 36$, $P(N_1 > 40) = 0.0057$.

supporting Result 4.2. In Figures 4.3 and 4.4, we see that the maximum discrepancy of the approximated densities for each time point is at most 0.0003.

Let us now consider all combinations of F and G simultaneously. In the Table 4.1, we show the mean and the standard deviations of T_1 obtained from the two processes for various choices of F and G. Similarly, in Table 4.2, we show the mean and the standard deviations of T_2. We show the corresponding λ's for each combination of F and G in Table 4.3.

4.4 Comparison with undivided Stage 1

In Table 4.3, we compare the means of the Stage 1 duration, showing the percentage change, between the divided Stage 1 studied here and the undivided Stage 1 studied in Chapter 3. We also report the λ's obtained in the current research and compare them to the λ's reported in 3.

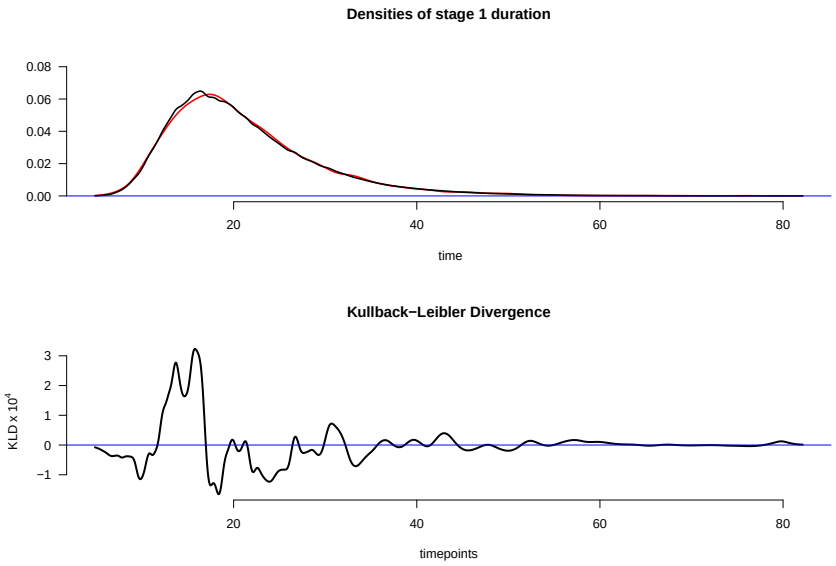


Figure 4.3. Densities of T_1 estimated from a point process (red) and an adjusted convolution process (black), with their difference being within 3.5×10^{-4} of 0.

From Table 4.3, we identify a trend in the values of λ as we scan the rows and columns. For a particular choice of F in a row, as we look from left to right across the columns, we see that λ decreases. A closer look at the corresponding standard deviations reveals that λ decreases as the standard deviation of G increases. Similarly, for a fixed choice of G in a particular column, as we go from top to bottom down the rows, we see that λ increases as the standard deviation of F increases. This led us to believe that λ is a function of the ratio of the standard deviations of F and G. When we plotted λ against σ_F/σ_G, we noticed a non-linear relationship. Thereafter, we fitted a linear regression of λ on $\log(\sigma_F/\sigma_G)$ with slope $= 0.11833$, intercept $= 0.20644$ and adjusted coefficient of determination of 0.832.

Let us look at the change in λ before and after the subdivision of Stage 1. When the sd of F is ≈ 0.25 or ≈ 0.50, the λ's for the divided Stage 1 is about one-half to three-fifths of the λ's from the undivided Stage 1; but when the standard deviation of F is ≈ 1, the λ's for the divided Stage 1 is about two-thirds of the λ's from the undivided Stage 1. Thus, there is a relation between the standard deviation of the VS and the λ's.

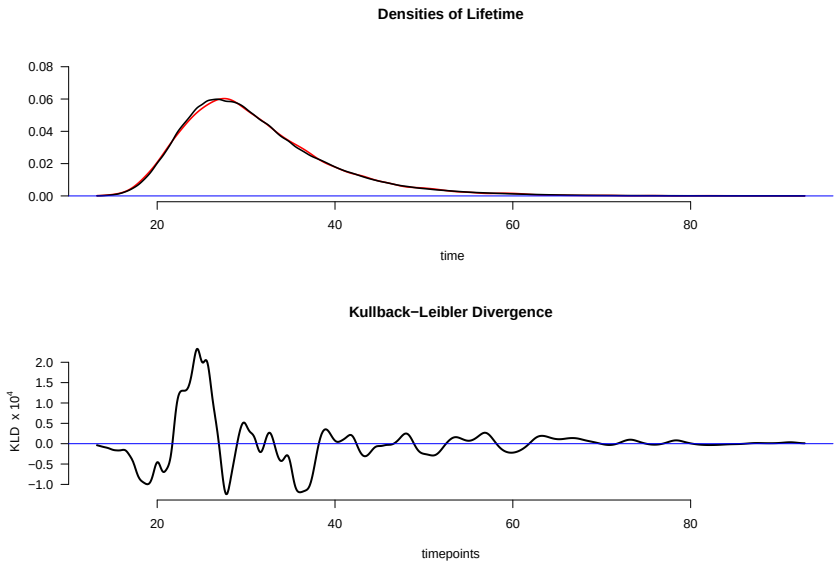


Figure 4.4. Densities of T_2 estimated from a point process (red) and an adjusted convolution process (black), with their difference being within 2.5×10^{-4} of 0.

In Table 4.4, we compare the means of the lifetimes, showing the percentage change, between the divided Stage 1 studied here and the undivided Stage 1 studied in Chapter 3.

4.5 Preventive maintenance policies

System failure being disruptive to the production process and too expensive to recover from, oftentimes a maintenance engineer must intervene to replace a functioning unit. Clearly there is a tension between utilizing the remaining lifetime of the system and the prevention of failure. We consider here two types of preventive maintenance policies. Let c_{p_A} be the cost of replacement in Stage 1A, c_{p_B} in Stage 1B, c_{p_2} in Stage 2, and c_f after failure. Furthermore, we assume that the costs of replacement is the same throughout Stage 1, because the healing rate ought not affect the cost of replacement. We consider $c_{p_A} = c_{p_B} \leq c_{p_2} << c_f$ with justification as follows: In Stage 1, the system is young, and so an early replacement will incur a loss; if we replace in early part of Stage 2, we are not utilizing the system lifetime sufficiently, but the

Table 4.1. For various inter-arrival time distributions satisfying $E[X] = 1$ and $E[Y] = 2/3$ the top entries give mean (standard deviation) of Stage 1 duration T_1 according to a point process, and the bottom entries (in *italics*) show the same quantities according to an adjusted convolution process.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $sd \approx 0.12$	Gamma $(2, \frac{1}{3})$ $sd \approx 0.22$	Inv-Gauss $(2/3)$ $sd \approx 0.29$	Exponential $(3/2)$ $sd \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $sd \approx 0.27$	21.51 (8.12) *21.50 (8.15)*	21.17 (8.35) *21.15 (8.38)*	21.01 (8.50) *21.01 (8.53)*	20.34 (8.39) *20.31 (8.39)*
Gamma $(2, \frac{1}{2})$ $sd \approx 0.50$	20.77 (9.13) *20.76 (9.12)*	20.42 (9.07) *20.44 (9.07)*	20.43 (9.21) *20.43 (9.20)*	19.82 (9.05) *19.81 (9.01)*
Inv-Gauss (1) $sd \approx 1.00$	19.53 (10.16) *19.61 (10.16)*	19.26 (10.01) *(19.36 (10.01)*	19.18 (10.11) *19.27 (10.07)*	18.88 (9.97) *18.98 (9.97)*
Exponential (1) $sd \approx 1.00$	19.59 (10.36) *19.59 (10.37)*	19.41 (10.36) *19.37 (10.39)*	19.46 (10.59) *19.43 (10.58)*	18.91 (10.04) *18.89 (10.06)*

system has already aged, and so maintenance/repair at this stage will cost more. Therefore, we find it logical to consider that replacement cost in Stage 2 is higher than that in Stage 1. Finally, a system failure is highly expensive. Furthermore, there is an initial cost c_0 of setting up a new system. For illustration, we choose $c_0 = 100, c_{p_A} = 10, c_{p_B} = 10, c_{p_2} = 15, c_f = 200$. In Figures 4.5 and 4.6, we illustrate decision making when $F \equiv$ Weibull $(2, 2/\sqrt{\pi})$ and $G \equiv$ gamma $(2, 1/3)$, such that $E[X] = 1$ and $E[Y] = 2/3$.

4.5.1 Maintenance policy 1

Suppose that a monitoring equipment can detect the arrival of an impetus, but cannot distinguish between a VS and a PI, nor can it identify whether the system is in Stage 1 or Stage 2. The system will be replaced when it has failed or has experienced a specified number of impetus N (the sum of VS and PI).

Within one cycle (between two successive replacements of the system), the total cost of replacement under Policy 1 is a random variable taking three possible values:

Table 4.2. For various inter-arrival time distributions satisfying $E[X] = 1$ and $E[Y] = 2/3$ the top entries give mean (standard deviation) of system lifetime T_2 according to a point process, and the bottom entries (in *italics*) show the same quantities according to an adjusted convolution process.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $sd \approx 0.12$	Gamma $(2, \frac{1}{3})$ $sd \approx 0.22$	Inv-Gauss $(2/3)$ $sd \approx 0.29$	Exponential $(3/2)$ $sd \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $sd \approx 0.27$	31.48 (8.27) *31.50 (8.31)*	31.14 (8.51) *31.15 (8.55)*	31.01 (8.65) *31.03 (8.69)*	30.32 (8.58) *30.33 (8.55)*
Gamma $(2, \frac{1}{2})$ $sd \approx 0.50$	30.77 (9.44) *30.76 (9.39)*	30.44 (9.34) *30.45 (9.33)*	30.42 (9.48) *30.43 (9.47)*	29.82 (9.34) *29.80 (9.28)*
Inv-Gauss (1) $sd \approx 1.00$	29.45 (10.69) *29.55 (10.67)*	29.23 (10.53) *29.32 (10.51)*	29.17 (10.66) *29.23 (10.57)*	28.83 (10.54) *28.91 (10.49)*
Exponential (1) $sd \approx 1.00$	29.62 (10.80) *29.57 (10.89)*	29.42 (10.82) *29.34 (10.88)*	29.51 (11.07) *29.43 (11.08)*	28.95 (10.53) *28.90 (10.58)*

(1) c_{p_1}, if N impetus arrive while the system is still in Stage 1, with an associated probability of $P(D_1 > N)$.

(2) c_{p_2}, if the system has already moved from Stage 1 to Stage 2 and the N impetus have arrived before system failure, with an associated probability of $P(D_1 \leq N < D_2)$.

(3) c_f, if the system has already failed before the arrival of N impetus, with an associated probability of $P(D_2 \leq N)$.

Hence, the expected cost (C) and the expected cycle time (CT), under Policy 1, are

$$E[C \mid \text{Policy 1}] = c_{p_1} P(N_1 > N) + c_{p_2} P(N_1 \leq N < N_2) + c_f P(N_2 \leq N) \tag{4.5.1}$$

Table 4.3. For various inter-arrival time distributions satisfying $E[X] = 1$ and $E[Y] = 2/3$, the top entries give the mean duration of T_1 for divided Stage 1 (undivided Stage 1), the middle row gives the % increase in T_1 after subdivision of Stage 1, and the third row gives the multiplier λ of adjusted convolution for the divided Stage 1 (undivided Stage 1).

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $sd \approx 0.12$	Gamma $(2, \frac{1}{3})$ $sd \approx 0.22$	Inv-Gauss $(2/3)$ $sd \approx 0.29$	Exponential $(3/2)$ $sd \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $sd \approx 0.27$	21.51 (17.96) $\approx 19.76\%$ $\lambda = 0.25(0.50)$	21.17 (17.97) $\approx 17.8\%$ $\lambda = 0.21(0.40)$	21.01 (17.98) $\approx 16.85\%$ $\lambda = 0.18(0.35)$	20.34 (17.93) $\approx 13.44\%$ $\lambda = 0.16(0.29)$
Gamma $(2, \frac{1}{2})$ $sd \approx 0.50$	20.77 (17.92) $\approx 15.90\%$ $\lambda = 0.31(0.59)$	20.42 (17.91) $\approx 14.00\%$ $\lambda = 0.27(0.50)$	20.43 (17.93) $\approx 13.94\%$ $\lambda = 0.25(0.46)$	19.82 (17.85) $\approx 11.03\%$ $\lambda = 0.22(0.38)$
Inv-Gauss (1) $sd \approx 1.00$	19.53 (17.63) $\approx 10.78\%$ $\lambda = 0.43(0.68)$	19.26 (17.61) $\approx 9.37\%$ $\lambda = 0.40(0.61)$	19.18 (17.67) $\approx 8.55\%$ $\lambda = 0.37(0.58)$	18.88 (17.61) $\approx 7.21\%$ $\lambda = 0.36(0.52)$
Exponential (1) $sd \approx 1.00$	19.59 (17.85) $\approx 9.75\%$ $\lambda = 0.45(0.69)$	19.41 (17.87) $\approx 8.62\%$ $\lambda = 0.43(0.63)$	19.46 (17.92) $\approx 8.59\%$ $\lambda = 0.39(0.60)$	18.91 (17.82) $\approx 6.12\%$ $\lambda = 0.37(0.54)$

and writing W_j as the arrival time of the j-th impact (either VS or PI), we have

$$\begin{aligned} E[CT \mid \text{Policy 1}] &= E[\min\{W_N, W_{N_2}\}] \\ &= E[W_N \mid N_1 > N]\ P(N_1 > N) \\ &\quad + E[W_N \mid N_1 \le N < N_2]\ P(N_1 \le N < N_2) \\ &\quad + E[W_{N_2} \mid N_2 \le N]\ P(N_2 \le N) \end{aligned} \tag{4.5.2}$$

Therefore, the expected cost per unit time is the ratio

$$E[C \mid \text{Policy 1}]/E[CT \mid \text{Policy 1}] \tag{4.5.3}$$

which we must minimize by choosing N. For the example considered, Figure 4.5 shows that the expected cost per unit time is minimized when we choose $N = 56$. Moreover, note that

Table 4.4. For various inter-arrival time distributions satisfying $E[X] = 1$ and $E[Y] = 2/3$, the top row gives the mean duration of T_2 for divided Stage 1 (undivided Stage 1), the bottom row gives approximate % increase in mean T_2 after subdividing Stage 1.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $sd \approx 0.12$	Gamma $(2, \frac{1}{3})$ $sd \approx 0.22$	Inv-Gauss $(2/3)$ $sd \approx 0.29$	Exponential $(3/2)$ $sd \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $sd \approx 0.27$	31.48 (27.94) $\approx 12.67\%$	31.14 (27.96) $\approx 11.37\%$	31.01 (27.96) $\approx 10.91\%$	30.02 (27.92) $\approx 7.52\%$
Gamma $(2, \frac{1}{2})$ $sd \approx 0.50$	30.77 (27.91) $\approx 10.25\%$	30.44 (27.92) $\approx 9.03\%$	30.42 (27.93) $\approx 8.92\%$	29.82 (27.86) $\approx 7.03\%$
Inv-Gauss (1) $sd \approx 1.00$	29.45 (27.57) $\approx 6.82\%$	29.23 (27.58) $\approx 5.98\%$	29.17 (27.60) $\approx 5.69\%$	29.83 (27.54) $\approx 8.32\%$
Exponential (1) $sd \approx 1.00$	29.62 (27.86) $\approx 6.32\%$	29.42 (27.90) $\approx 5.45\%$	29.51 (27.93) $\approx 5.66\%$	28.95 (27.82) $\approx 4.06\%$

for any other choice of N in the vicinity of the optimal value 56, say between 50 and 60, the expected cost per unit time increases only slightly (no more than 3%). Such a robustness result allows us to rely on the optimal value even when the inter-arrival time distributions deviate slightly from the stated ones. Table 4.5 documents the optimal choices for other combinations of inter-arrival times.

4.5.2 Maintenance policy 2

Suppose that the monitoring equipment can identify the stages of the system. If the system is in Stage 1, we do not replace it at all. After the system enters Stage 2, if the system is still functioning for an additional t units of time, we replace it immediately at epoch $T_1 + t$; otherwise, we replace the system immediately on failure during $[T_1, T_1 + t)$. Our objective is to determine an optimum additional time t in Stage 2. To do so, we minimize the expected cost per unit time, where the expected cost (C) is given by

$$E[C \mid \text{Policy 2}] = c_{p_2} P(T_2 > T_1 + t) + c_f P(T_2 \leq T_1 + t) \tag{4.5.4}$$

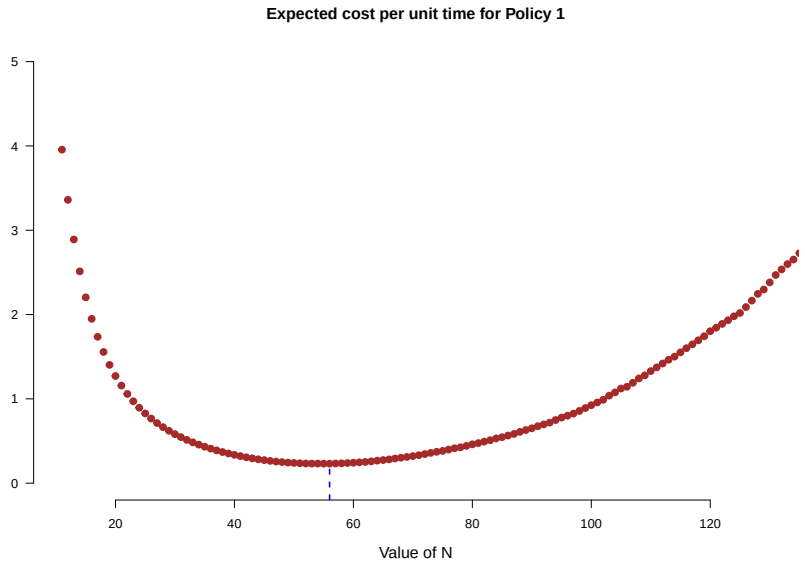


Figure 4.5. Under Policy 1 and cost parameters $c_0 = 100$, $c_{p_A} = 10$,$c_{p_B} = 10$, $c_{p_2} = 15$, $c_f = 200$, if $F \equiv$ Weibull $(2, 2/\sqrt{\pi})$ and $G \equiv$ gamma $(2, 1/3)$, then to minimize the maintenance cost per unit time, the optimal number of impetus for preventive replacement is $N = 56$.

and the expected length of the cycle time (CT) is

$$E[CT \mid \text{Policy 2}] = E[min(T_2, T_1 + t)] = E[T_1] + E[min(T_2 - T_1, t)] \tag{4.5.5}$$

We wish to minimize the expected cost per unit time

$$E[C \mid \text{Policy 2}]/E[CT \mid \text{Policy 2}] \tag{4.5.6}$$

by choosing t. Under Policy 2, the assumed cost parameters, and $F \equiv$ Weibull $(2, 2/\sqrt{\pi})$ and $G \equiv$ gamma $(2, 1/3)$, Figure 4.6 shows that the expected cost per unit time is minimized at $t = 6.6$. In fact, we identified this optimal t value via a grid search between the first and the 99-th percentiles of system lifetime with an increment of 0.05. This choice suffices because any other choice of t in the interval $[6, 7]$ increases the cost per unit time only marginally. Table 4.5 documents the optimal choices of t for other combinations of inter-arrival times.

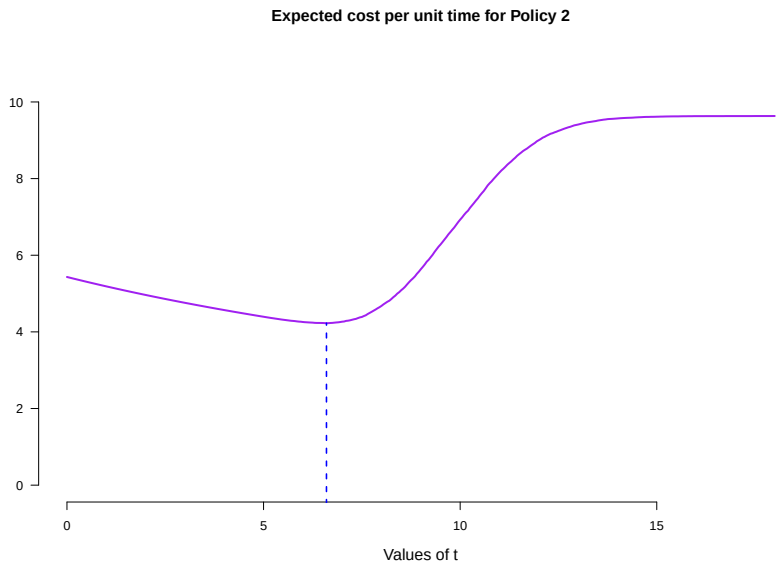


Figure 4.6. Using cost parameters $c_0 = 100$, $c_{p_A} = 10$,$c_{p_B} = 10$, $c_{p_2} = 15$, $c_f = 200$, $F \equiv$ Weibull $(2,\ 2/\sqrt{\pi})$ and $G \equiv$ gamma $(2,\ 1/3)$, to minimize the maintenance cost per unit time, the optimal duration in Stage 2 after which preventive replacement must be scheduled is $t = 6.6$.

The following tables give the summary of the optimal choices of N and t for Policy 1 and Policy 2, respectively, for different choices of F and G.

We see that for policy 1, the total number of impacts N is only 0-3 impacts more than the optimal values of N when Stage 1 was not divided. This is because, for our choice of (k_A, k_B) and (m_A, m_B), the average impact throughout the entire undivided Stage 1 is comparable to that in the undivided Stage 1 case. Hence, there is only a negligible amount of change in N due to subdivision. Similarly, for policy 2, there is no significant change in t, because it depends only on the arrival rate of VS in Stage 2 and not at all on the subdivision of Stage 1 to accommodate varying rates of healing.

4.6 Summary

In this chapter, we subdivided Stage 1, where healing is permissible, into two parts: initially the system heals at a faster rate requiring a few PIs to nullify one VS; but once

Table 4.5. For various inter-arrival time distributions F and G satisfying $E[X] = 1$ and $E[Y] = 2/3$, to minimize the maintenance cost per unit time, the optimal N for Policy 1 is shown in the first row, and the optimal t for Policy 2 is shown in the second row.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $sd \approx 0.12$	Gamma $(2, \frac{1}{3})$ $sd \approx 0.22$	Inv-Gauss $(2/3)$ $sd \approx 0.29$	Exponential $(3/2)$ $sd \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $sd \approx 0.27$	$N = 55$ $t = 6.45$	$N = 56$ $t = 6.60$	$N = 53$ $t = 6.55$	$N = 52$ $t = 6.60$
Gamma $(2, \frac{1}{2})$ $sd \approx 0.50$	$N = 54$ $t = 5.70$	$N = 51$ $t = 5.75$	$N = 52$ $t = 5.80$	$N = 50$ $t = 5.85$
Inv-Gauss (1) $sd \approx 1.00$	$N = 51$ $t = 4.85$	$N = 50$ $t = 4.65$	$N = 49$ $t = 5.05$	$N = 49$ $t = 4.90$
Exponential (1) $sd \approx 1.00$	$N = 50$ $t = 4.75$	$N = 50$ $t = 4.65$	$N = 50$ $t = 4.65$	$N = 49$ $t = 4.80$

enough net VS have accumulated, more PIs are needed to nullify one VS. The theoretical investigation of λ remains an open problem. Moreover, in this research we found that subdivision of Stage 1 leads to an increase in the Stage 1 duration, and hence the system lifetime. In the next chapter, we will consider varying magnitudes of VS and PI and allow natural system degradation.

5. AN OPTIMAL REPLACEMENT POLICY UNDER VARIABLE SHOCKS AND DIFFERENT PATTERNS OF SELF-HEALING

5.1 Introduction

Industrial systems are often challenged by external impetus that affect their normal functioning. An impetus that inflicts a damaging effect is called a "shock". In the last few years, there have been extensive studies on different types of external shocks and their effects on a system. On the other hand, when an impetus produces a positive effect on the system by improving its current state, it is called a "healing effect". When the system heals by default, without requiring intervention, it is called "self-healing". Such natural and continuous self-healing is commonplace in many industrial, ecological, and biological systems and may continue either indefinitely or for a specific duration.

Self-healing exists in software debugging systems, where bugs, malware invasion, license expiration, etc. are considered shocks, while automatic system cleansing is considered self-healing. The software industry spends millions of dollars to monitor and maintain such systems to prevent them from failing, especially when such a failure is catastrophic. These maintenance policies seek optimal rules to replace the system before risking it's failure.

In recent decades, many shock models with healing effects have been studied that also permit sporadic shocks of variable magnitudes and continuous internal degradation. For example, in Shen *et al.* (2018) shocks arrive according to a Poisson process with changing intensities. Depending on the degree of accumulated damage, the system performance can be divided into several states. In some states, self-healing reduces the accumulated damage; however, self-healing can stop when the system reaches a specific state. There may also be an internal degradation process. For instance, in some systems or components such as micro-electro-mechanical systems (MEMS) and servo motors as described in Wang *et al.* (2020), natural degradation affects the consequences caused by shocks and vice versa. That paper allows a natural-degradation-state (NDS) function to classify shocks into safety, damage, and fatal zones according to their thresholds, and derives closed-form reliability function and

failure time distribution function. Dong *et al.* (2020b) introduced a "damage recovery factor" to quantify self-healing and its effect on the reliability function and the mean failure time. They allow random shocks to accelerate internal degradation rate and discuss a preventive replacement policy. Similarly, Kong and Yang (2020) formulated a reliability model under multiple competing failure processes. They considered the magnitudes of the shock and their duration simultaneously to study their impacts on the degradation processes, describing both recovery level and recovery time. Ranjkesh *et al.* (2019), evaluate system reliability using dependencies between inter-arrival times having phase-type distributions and shock magnitudes.

A survey of various maintenance policies for industrial systems is provided in Wang (2002), including a broad spectrum of replacement policies such as age-dependent, periodic, failure limit and sequential preventive maintenance policies. These policies essentially use one of the following optimization criteria: maximize availability, minimize expected cost per unit time, minimize downtime, and minimize limiting failure rate. They also consider various repair policies, such as perfect or imperfect repair, and various monitoring strategies such as monitoring at discrete time points or continuously. We will next seek an optimal maintenance policy that minimizes the average cost per unit time.

Preventive maintenance (PM) and corrective maintenance (CM) are the two classical types of maintenance policies undertaken to maximize profit or minimize loss due to failure. In Chien *et al.* (2012), each period of operation inflicts a random amount of damage to the system and those damages accumulate to trigger a PM or a CM action. The long-run expected cost rate is minimized to determine the optimal policy. Qiu *et al.* (2020) develops a novel reliability model characterizing self-healing effect on system reliability. They carry out an imperfect repair following each minor failure, and replace the system based on its lifetime and the number of minor failures. The optimal replacement time is determined using a stochastic dynamic programming formulation that minimizes the expected total cost of system failure and imperfect repairs. In Dong *et al.* (2021) external shocks and their damaging effects are considered for multi-component systems that are subject to dependent and competing failure processes. Generalized shock models are presented under several shock categories. A block replacement policy is introduced, and the Nelder-Mead downhill simplex

method is employed to determine the optimal replacement interval based on the derived system reliability.

Some applications of systems exposed to external shocks, internal degradation, and experiencing self-healing are the following. Dong *et al.* (2020b) considers micro-electro-mechanical systems (MEMS), where electrostatic, piezoelectric, optical, mechanical vibration or magnetic stimuli are classified as external shocks, whereas self-healing is induced in the electrode by a polymer binder. Dong *et al.* (2021) discusses a similar system with internal degradation. Kong and Yang (2020) discusses an example of insulators in power-transmission systems, where temperature, humidity, ultraviolet radiation, and corona discharge are external damaging shocks; and hydrophobicity degradation and corona discharge cause internal degradation of the system. They consider the system recovery level and time to characterize a self-healing mechanism.

We build this current chapter on our previous Chapters 3 and 4, where the system was exposed to randomly arriving external shocks of the same magnitude. Here, unlike previously, we let external shocks inflict damages of varying magnitudes. We also permit the system to begin to heal instantaneously and continue to heal at a fixed rate while also continuously degrading due to aging. Under this more general model of variable damage, our objective is to determine an optimal replacement time that minimizes the cost per unit time.

The remainder of this chapter is organized as follows. Section 5.2 describes the evolution of the system as a continuous-time stochastic process that renews itself after a preventive or corrective replacement. Subsection 5.2.1 explains the method of computing the expected cost per unit time. Subsection 5.2.2 reports the optimal replacement times obtained in simulation studies. Furthermore, once we have established the optimal replacement policy, we also consider different variations of our assumed system in Section 5.3: First, the damaging shocks can heal only for a finite duration τ (Subsection 5.3.1); second, a fixed proportion of shocks are non-healable (that is, $\tau = 0$ for these shocks) (Subsection 5.3.2); and third, there are two different types of shocks—healable for a finite duration τ and non-healable (Subsection 5.3.3). Details of the simulation studies for the three sub-cases are mentioned in Section 5.3. Section 5.4 summarizes our research findings.

5.2 Stochastic evolution of systems

The system described in this research is either a single- or a multi-unit system. External shocks arrive with random inter-arrival times inflicting damages of random magnitudes. Immediately after a shock arrives, the system begins to heal, which reduces the accumulated damage by automated replenishment which is known as the self-healing behavior. We assume that healing occurs continuously and indefinitely at a constant rate according to an exponential function of time; therefore, cumulative damage decreases exponentially. The system fails when cumulative damage crosses a certain threshold, which decreases with time as a result of aging.

The set-up and assumptions:

(A1) Let $X_1, X_2, \ldots, X_n$ denote the inter-arrival times of shocks which are IID with arbitrary CDF F.

(A2) Let $Y_1, Y_2, \ldots, Y_n$ denote the corresponding magnitudes of damage caused by the external shocks, which are IID with arbitrary CDF G.

(A3) Damages from the shocks accumulate over time.

(A4) The system self-heals from the damages at a constant rate. Hence, at any given time, either a shock arrives, causing the cumulative damage to shoot up, or the system continuously heals from the effects of all previous shocks, causing the cumulative damage to decrease continuously.

(A5) The system fails when the accumulated damage exceeds a certain boundary threshold, which decreases over time at a faster rate as the system ages, making it more vulnerable to failure. Here, for illustration, we assume that the boundary is a quadratically decreasing function of time. Thus, the system fails in one of two ways:

(i) a new shock arrives so that the cumulative damage exceeds the boundary;

(ii) the accumulated damage, though decreasing, crosses the boundary while the system is healing because aging causes the boundary to come down faster.

Figure 5.1, which depicts the accumulated damage as a function of time, illustrates these two types of failure using two sample paths. The black sample path crosses the boundary threshold when a shock of sufficient magnitude arrives. The red sample path exceeds the boundary while the system is healing, but the boundary is reducing faster.

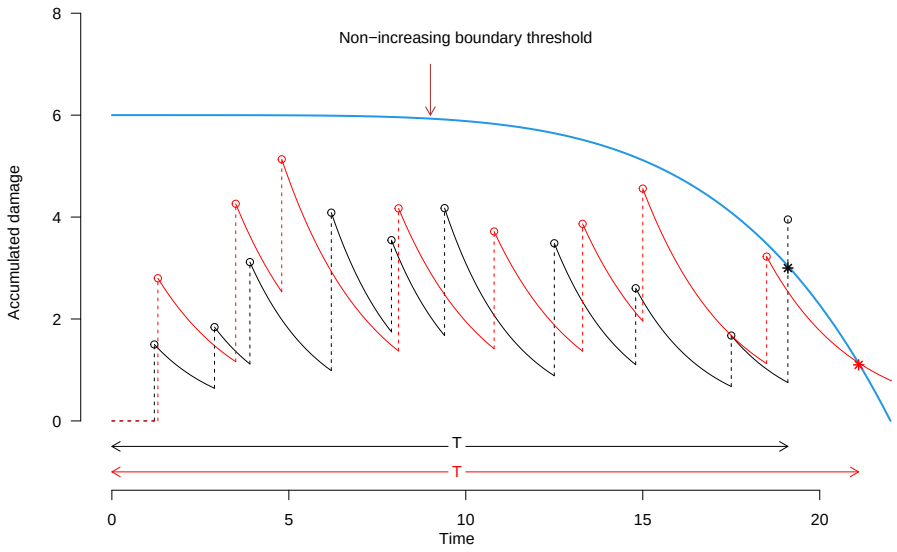


Figure 5.1. Depicting cumulative damage (black and red curves) as shocks arrive randomly. The blue curve represents the quadratically decreasing boundary threshold; dotted vertical segments denote random amount of damage inflicted by each shock, and the continuous curves represent exponential decay of cumulative damage due to constant healing. When the cumulative damage exceeds the boundary threshold, the system fails.

5.2.1 Methodology

Let us explain how to calculate the cumulative damage to the system at any given time: Let $S_j = \sum_{i=1}^{j} X_i$ be the arrival time of the j-th shock. Let $N(t)$ denote the number of

shocks that have arrived till time t (observed at increments of Δ). Then the cumulative damage D, at time t is computed as

$$D(t) = \sum_{i=1}^{N(t)} Y_i \mathrm{e}^{-\kappa(t-S_i)} \tag{5.2.1}$$

where Y_i is the damage inflicted by the i-th shock, and κ is the fixed healing rate.

For monitoring purpose, the system is observed at regular epochs at increments of Δ over a window of time $[0, \mathcal{T}]$. Borrowing the discretization approach from Chapter 2, we choose Δ sufficiently small so that for all practical purposes, we observe the system almost continuously. At each observation epoch, we measure the difference between the cumulative damage and the boundary threshold given by

$$B(t) = a + bt - ct^2 \tag{5.2.2}$$

where $a, c \in \mathbb{R}^+$ and $b \in \mathbb{R}$. Note that, the above choice of the boundary threshold is illustrative only, one can consider any other non-increasing function of time. For $k = 1, 2, \ldots,$ if the cumulative damage $D(t)$ does not cross the threshold $B(t)$ at the $(k-1)$-st observation epoch, but is found to have crossed it at the k-th inspection epoch, then the system has failed and we replace it at time $T = k\Delta$.

However, replacement upon failure is not desirable due to the high cost of replacing a failed system and the loss of revenue until a new system is installed. Instead, we must determine a preventive maintenance policy in which we replace the system before failure; however, we must not replace the system too early and forfeit its remaining lifetime. Therefore, we propose the following maintenance policy: Whenever the cumulative damage $D(t)$ enters within d units of the threshold $B(t)$, where $d > 0$ has yet to be determined, an alarm is activated and we replace the system after an additional time t^* which depends on the tolerable risk probability (say, between 10% and 20%) that the system might fail before that epoch. Thus, the choice parameter d is related to the additional duration t^* after the alarm sets off, when we replace the unfailed system. For any choice of $d \in [D_1, D_2]$, at increments of 0.1, we apply

Algorithm 5.1 to compute lifetime, replacement time, residual life-time and the number of shocks.

Algorithm 5.1

(S1) Generate n shocks with inter-arrival time $X_1, X_2, \ldots, X_n$ IID with CDF F and magnitudes. $Y_1, Y_2, \ldots, Y_n$ IID with CDF G, where n is sufficiently large; say, $n \approx 2\mathcal{T}/E[X]$.

(S2) Calculate cumulative damage $D(t)$ at each epoch $t = j\Delta$ (for $j = 1, 2, \cdots$) using equation (5.2.1). Suppose that $T = k\Delta$, is the first time $D(t)$ exceeds the boundary. Then we must replace the system at epoch T, called *failure time*.

(S3) We record the number N of shocks that the system endures until failure.

(S4) Let $T' = l\Delta$, for some $l < k$, be the first time that cumulative damage $D(t)$ is within d units of the boundary. Had we replaced the system at T', the lifetime of the system lost due to premature replacement would have been $T - T'$, and would be called *residual lifetime*.

(S5) Repeat steps (S1) to (S4) 10^4 times. For each repetition, obtain T, T', N, and compute $r = T - T'$.

Note that our objective is to utilize the system to its fullest. Therefore, we should not replace the system too soon. How long could we allow the system to function before replacing it so that the chance of a system failure within this additional duration would be 10%, 15% or 20% (equivalently, survival probabilities would be 90%, 85% or 80%), respectively? Let us denote these survival percentiles after the alarm sets off by $t^*_{90}(d)$, $t^*_{85}(d)$, and $t^*_{80}(d)$, respectively (collectively denoted by $t^*_\gamma(d)$ for $\gamma = 0.90, 0.85, 0.80$). We compute these survival percentiles using the 1000 values of the residual lifetime $r = T - T'$. This we do for every choice of $d \in [D_1, D_2]$, at increments of 0.1. Our main objective is to find an optimal d such that the expected cost per unit of time is minimized when we are willing to risk a small (10%, 15% or 20%) chance of system failure within the next $t^*_\gamma(d)$ units of time after the alarm sets off.

To compute the expected cost per unit time, let c_0 be the initial cost of installing the system, c_I be the per unit time cost of inspection and maintenance (although the inspection

cost is incurred at increments of Δ, we redistribute the cost over the entire interval), c_{op} be the per unit time cost of operating the system, c_{rev} be the per unit time revenue earned by the system while operating (it is a negative cost), and c_f be the additional cost of failure replacement. We define as cycle time the duration from the time a system is installed for operation until it is replaced at epoch $\min\{T, T' + t^*_\gamma(d)\}$. Then, the expected cost (EC) within a cycle time is:

$$EC[t^*_\gamma(d)] = c_0 + (c_I + c_{op} - c_{rev}) \times E[\min\{T, T' + t^*_\gamma(d)\}] + c_f I_F \tag{5.2.3}$$

where $I_F = 1$ if the system experiences failure and $I_F = 0$ if the system is replaced before failure. We wish to minimize $EC[t^*_\gamma(d)]/E[\min\{T, T' + t^*_\gamma(d)\}]$ with respect to d.

5.2.2 Simulations

Let us demonstrate how to find the optimal d and associated expected $t^*_\gamma(d)$, in our proposed policy we choose $\mathcal{T} = 100, \Delta = 0.05$, $\kappa = 0.01$ or 0.02, $E[X] = 1$ and $E[Y] = 10$, respectively. We take $n = 200$ so that in every iteration, the cumulative damage (almost) surely crosses the boundary (equation (5.2.2)) with $a = 500$, $b = 0$ and $c = 0, 1/60, 1/50, 1/40, 1/30$. We calculate the cumulative damage at time t using equation (5.2.1). Let the various costs be $c_0 = 5000, c_I = 50, c_{op} = 100, c_{rev} = 200, c_f = 1000$. We search for optimal $d \in [D_1 = 8, D_2 = 16]$.

For inter-arrival time between shocks, we choose $X \sim$ Weibull (shape $= 2$, scale $= 2/\sqrt{\pi}$) so that $E[X] = 1$. For the magnitude of the shocks, we choose $Y \sim$ Weibull (shape $= 10$, scale $= 50/\Gamma(1/5)$) so that $E[Y] = 10$. For each choice of $d \in [8, 16]$ in increments 0.1, values of t^*_γ are obtained for $\gamma = 0.90, 0.85, 0.80$. In Figure 5.2, as an illustration, we display the survival plot for a particular choice $d = 12$ showing that the additional time after the alarm sets off that the system can be operated with $90\%, 85\%$ and 80% survival probabilities, respectively, are $0.50, 0.80, 1.04$.

Next, using equation (5.2.3), the expected cost per unit time is calculated for every choice of d and the associated t^*_γ, for survival probabilities $\gamma = 90\%, 85\%, 80\%$. Figure 5.3(a) demonstrates that the larger d is (that is, the farther the accumulated damage from the

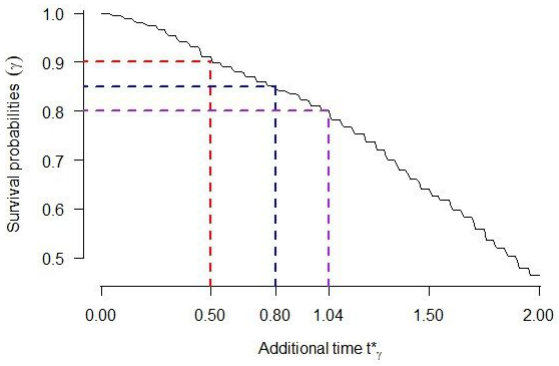


Figure 5.2. When $B(t) = 500 - t^2/50$, $\kappa = 0.01$ and $d = 12$, the additional time t^*_γ that the system should be allowed to operate after the alarm sets off, for $\gamma = 0.90, 0.85, 0.80$.

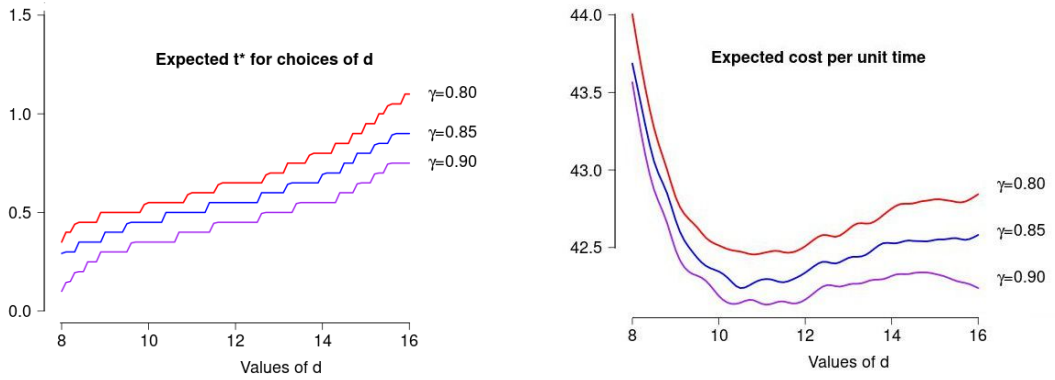
boundary), the larger is the t^*_γ, for each $\gamma = 0.90, 0.85, 0.80$. Figure 5.3(b) shows the expected cost per unit time as a function of d when $X \sim$ Weibull (shape = 2, scale = $2/\sqrt{\pi}$), $Y \sim$ Weibull (shape = 2, scale =1/2) and $B(t) = 500 - t^2/50$. If expected cost per unit time has multiple minima, we choose the smallest one, since we wish to utilize the system as much as possible without compromising the cost per unit time.

Table 5.1 displays the simulation results showing the optimal d for different choices of κ and boundary thresholds with different quadratic coefficients, but keeping the inter-arrival distribution Weibull and the magnitude of shocks Weibull.

Here are some lessons learned from Table 5.1:

1. If the healing rate κ increases, the system heals faster so that it takes longer for the accumulated damage to come within d units of the boundary, thus increasing the replacement time. Using the same logic, a higher κ also causes the optimal d to be smaller.

2. Furthermore, for a fixed κ and for a particular choice of boundary threshold, if the survival probabilities are chosen to be smaller, then t^* increases or at least remains the same (because, by definition, $t^*_{80} \geq t^*_{85} \geq t^*_{90}$), and the optimal d also increases, because



(a) (b)

Figure 5.3. (a) The percentiles $t^*_{80} \geq t^*_{85} \geq t^*_{90}$ are increasing functions of d. (b) The expected cost per unit time is minimized at $d = 10.3$ for $\gamma = 0.90$; at $d = 10.4$ for $\gamma = 0.85$ and $\gamma = 0.80$. If there are multiple minima, choose the smallest one.

of monotonic relation with t^* exhibited in Figure 5.3(a). On the contrary, if we demand a higher survival rate, then the optimal d decreases.

3. As the boundary threshold decreases at a faster rate, the optimal d increases for each of the survival rates 90%, 85% and 80%. This is because when the boundary decreases faster, we should let the alarm go off earlier to avoid potential failure.

For boundary $B(t) = 500 - t^2/50$, healing rate $\kappa = 0.02$, and Y following Weibull distribution, Table 5.2 shows the optimal d and the associated t^*_γ for various distributions of inter-arrival time of shocks.

When X has Weibull or gamma distribution, the additional time $t^*(\gamma)$ after the alarm sets off, are comparable, and the optimal d is robust around 10. However, when X is inverse-Gaussian, the t^*_γ values are much lower, because inverse-Gaussian distribution has a heavier right tail than Weibull and gamma. Likewise, because the exponential distribution has an even thicker right tail, the corresponding t^*_γ's are even smaller. Among all inter-arrival time distributions considered here, exponential is the most heavy-tailed; hence its survival function is the highest, and the optimal d is the highest.

Table 5.1. For $X \sim$ Weibull(2,2/π), $Y \sim$ Weibull(2,1/2), and for various choices of κ and $B(t)$, the optimal d and [the associated t^*_γ] are displayed for $\gamma = 0.90, 0.85, 0.80$.

$B(t) = 500$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	9.1 [0.625]	9.1 [0.803]	9.6 [0.963]
0.02	8.2 [1.105]	8.2 [1.137]	8.8 [1.423]
$B(t) = 500 - t^2/60$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	9.6 [0.448]	9.6 [0.583]	9.6 [0.709]
0.02	9.3 [0.610]	9.3 [0.782]	9.3 [0.953]
$B(t) = 500 - t^2/50$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	10.3 [0.456]	10.4 [0.575]	10.4 [0.679]
0.02	9.3 [0.601]	10.1 [0.774]	10.1 [0.934]
$B(t) = 500 - t^2/40$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	10.6 [0.401]	10.6 [0.534]	10.6 [0.638]
0.02	9.5 [0.530]	10.2 [0.698]	10.2 [0.843]
$B(t) = 500 - t^2/30$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	10.8 [0.376]	10.8 [0.495]	10.8 [0.606]
0.02	9.7 [0.526]	10.7 [0.681]	10.7 [0.813]

In the next section, we discuss some variations on the healing pattern.

5.3 Variations in the healing effect and the shock types

In this section, we discuss some variations on the stochastic modelling of the system evolution described earlier.

5.3.1 Case 1: Healing stops after a finite duration

Unlike in the previous section where healing continues indefinitely so that the damage eventually heals 100%, in this subsection, healing continues only up to a finite duration τ, and

Table 5.2. For $Y \sim$ Weibull(2,1/2), $\kappa = 0.02$, and $B(t) = 500 - t^2/50$, and various inter-arrival time distributions, the optimal d and [the associated t^*_γ] are displayed for $\gamma = 0.90, 0.85, 0.80$.

$X \sim$ *gamma* (*shape* $= 3$, *scale* $= 1/3$)		
$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
8.7 [0.542]	8.7 [0.706]	8.4 [0.865]
$X \sim$ *Weibull* (*shape* $= 2$, *scale* $= 2/\sqrt{\pi}$)		
$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
9.3 [0.601]	10.1 [0.698]	10.1 [0.934]
$X \sim$ *inverse-Gaussian* (*mean* $= 1$)		
$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
9.9 [0.334]	9.9 [0.432]	10.5 [0.529]
$X \sim$ *exponential* (*rate* $= 1$)		
$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
11.4 [0.143]	10.9 [0.229]	10.7 [0.315]

thereafter stops, so that only a certain percentage of the inflicted damage heals. Conversely, if we specify what proportion of the inflicted damage will heal, we can find the corresponding τ. Thus, the shocks are not totally healable, and a residual damage is left behind. Figure 5.4 illustrates the accumulated damage until system failure.

For illustration, we make the following choices: For exponential healing with rate $\kappa = 0.01$, we choose $\tau = 50$ to attain a 40% healing of the inflicted damage, and $\tau = 25$ for a 22% healing. When the healing rate increases to $\kappa = 0.02$, a choice of $\tau = 25$ attains a 40% healing, and $\tau = 50$ attains a 64% healing. Hence, at a given time t, the cumulative damage D to the system at time t is calculated as

$$D(t) = \sum_{i=1}^{N(t)} Y_i \mathrm{e}^{-\kappa[(t-S_i)\wedge\tau]} \tag{5.3.1}$$

where the notation "$\wedge$" stands for the minimum. We see that the equation (5.3.1) matches equation (5.2.1) when we let $\tau = \infty$. As in the previous section, we calculate the lifetimes, replacement times and associated t^*_γ' s for survival probabilities $\gamma = 90\%, 85\%, 80\%$ after 1000 repetitions of the stochastic process. We implement the same preventive maintenance

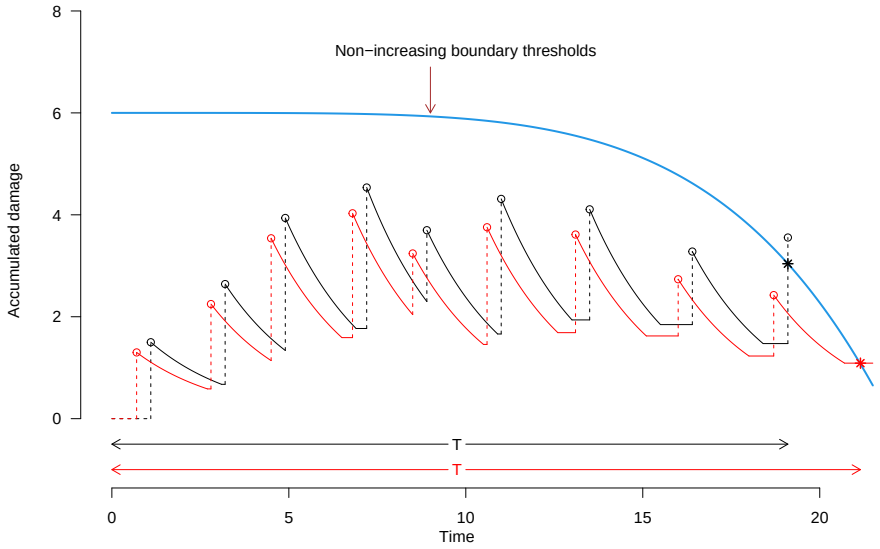


Figure 5.4. Depicting cumulative damage (black and red) as shocks arrive randomly. The blue curve represents the quadratically decreasing boundary threshold; dotted vertical segments denote random amount of damage inflicted by each shock, and the continuous curves represent exponential decay of cumulative damage upto a finite duration $\tau = 1$ due to constant healing. When the cumulative damage exceeds the boundary threshold, the system fails.

policy as in Subsection 5.2.2 to document in Table 5.3 and Table 5.4 the optimal d and the associated t^*_γ under the modified healing rule for $\tau = 50$ and 25 respectively.

Note that the overall optimal replacement time t^*_γ when τ is a finite number is lower than that in Subsection 5.2.2 where $\tau = \infty$. This is anticipated because when the shocks do not heal indefinitely, their residual damages bring the collective damage closer to the boundary threshold much earlier. In general, we also see that the t^*_γ's are significantly lower and the optimal d's are larger in Tables 5.3 and 5.4 as compared to Table 5.1, implying that the alarm goes off when the distance from the boundary is larger and we wait a shorter duration after the alarm goes off to replace the system. Further, comparing Tables 5.3 and 5.4, we see that for the latter one, the optimal d's are larger and the expected t^*_γ's are lower beacuse when $\tau = 25$, the healing continues for a shorter duration than when $\tau = 50$.

Table 5.3. For $\tau = 50$, $X \sim$ Weibull(2,2/π), $Y \sim$ Weibull(2,1/2), and for various choices of κ and $B(t)$, the optimal d and [the associated t^*_γ] are displayed for $\gamma = 0.90, 0.85, 0.80$.

$B(t) = 500 - t^2/60$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	9.9 [0.447]	9.9 [0.575]	11.1 [0.690]
0.02	9.3 [0.578]	9.8 [0.742]	9.8 [0.880]
$B(t) = 500 - t^2/50$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	10.5 [0.435]	11 [0.564]	11.3 [0.683]
0.02	9.6 [0.552]	9.6 [0.714]	10.4 [0.863]
$B(t) = 500 - t^2/40$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	10.6 [0.392]	10.6 [0.527]	11.4 [0.634]
0.02	9.7 [0.505]	10.1 [0.642]	10.2 [0.774]

5.3.2 Case 2: Some shocks are not healable

In this subsection, we consider the situation when not all shocks are healable. A fixed proportion p of shocks never heal; that is, their damage is permanent. Equivalently, for such shocks $\tau = 0$. We incorporate the effect of such shocks, not by an increase in accumulated damage, but by a sudden drop in the threshold boundary. Figure 5.6 illustrates the cumulative damage until it exceeds the boundary threshold.

(1) Classify a shock as nonhealable with probability p.

(2) Let $N(t)$ denote the number of shocks that have arrived by time t (observed at increments of Δ). Let H_i be an indicator function that takes value 1 if the i-th shock is healable, and 0 otherwise. Then the boundary curve drops by the corresponding magnitude of the nonhealable shock, making the modified boundary

$$B(t) = a + bt - ct^2 - \sum_{i=1}^{N(t)} (1 - H_i) Y_i \tag{5.3.2}$$

Table 5.4. For $\tau = 25$, $X \sim$ Weibull(2,2/π), $Y \sim$ Weibull(2,1/2), and for various choices of κ and $B(t)$, the optimal d and [the associated t^*_γ] are displayed for $\gamma = 0.90, 0.85, 0.80$.

$B(t) = 500 - t^2/60$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	10.8 [0.338]	11.3 [0.449]	11.3 [0.568]
0.02	10.5 [0.426]	10.5 [0.562]	10.5 [0.682]
$B(t) = 500 - t^2/50$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	11.3 [0.349]	11.3 [0.456]	11.3 [0.561]
0.02	10.7 [0.430]	10.7 [0.560]	10.7 [0.669]
$B(t) = 500 - t^2/40$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	11.3 [0.339]	11.5 [0.442]	11.7 [0.541]
0.02	10.8 [0.394]	11.1 [0.517]	11.1 [0.626]

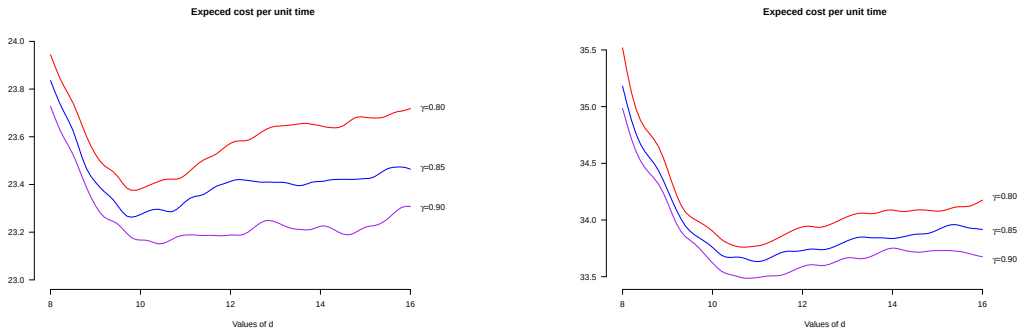
(3) We record the cumulative damage inflicted by healable shocks only. Therefore, the cumulative damage D to the system at time t, is calculated as in Subsection 5.2.1

$$D(t) = \sum_{i=1}^{N(t)} H_i \; Y_i e^{-\kappa[(t-S_i)\wedge\tau]} \tag{5.3.3}$$

As in Subsection 5.3.1, here also on average, compared to Subsection 5.2.2, the overall waiting time until replacement after the alarm goes off is shorter.

5.3.3 Case 3: The arrival times of healable and non-healable shocks have different distributions, so do their magnitudes

Suppose that the shocks affecting the system are of two types based on their healing capabilities. The first type of shock is self-healable for a finite duration τ (or up to a certain percentage of the damage heals and the rest is permanent). We assume such healable shocks arrive with inter-arrival times $X_1, X_2, \ldots, X_n$ which are IID with arbitrary CDF F. Moreover, the magnitudes of each of such shocks are denoted by $Y_1, Y_2, \ldots, Y_n$ which are IID with arbitrary CDF G. The second type of shocks are nonhealable (or $\tau = 0$) and their impact is



(a) (b)

Figure 5.5. (a) With $\kappa = 0.02$ and $\tau = 50$, the expected cost per unit time is minimized at $d = 9.3$ for $\gamma = 0.90$ and $\gamma = 0.85$ and at $d = 9.9$ for $\gamma = 0.80$. (b) With $\kappa = 0.02$ and $\tau = 25$, the expected cost per unit time is minimized at $d = 10.7$ for $\gamma = 0.90, 0.85, 0.80$. If there are multiple minima, choose the smallest one.

characterized by drops in the non-increasing boundary threshold causing the system degrade more severely than under natural aging. Let $Z_1, Z_2, \ldots, Z_m$ denote the inter-arrival times of the nonhealable shocks, which are IID with arbitrary CDF H. Let $U_1, U_2, \ldots, U_m$ denote the magnitudes of such shocks, which are IID with arbitrary CDF K. The system fails in one of three ways:

(i) a new healable shock arrives so that the cumulative damage exceeds the boundary;

(ii) the accumulated damage curve, although decreasing because of healing, crosses the boundary which decreases faster due to aging;

(iii) a new nonhealable shock arrives so that the boundary suddenly drops below the (otherwise) gently decreasing cumulative damage curve.

Figure 5.7 illustrates the type (i) and (ii) failures using black and red sample paths respectively which depict accumulated damage as a function of time. The black sample path crosses the boundary threshold when a healable shock of sufficient magnitude arrives. The red sample path exceeds the boundary while the system is healing, but the boundary comes down faster due to aging. Type (iii) failure is self-explanatory (and not shown in the illustration).

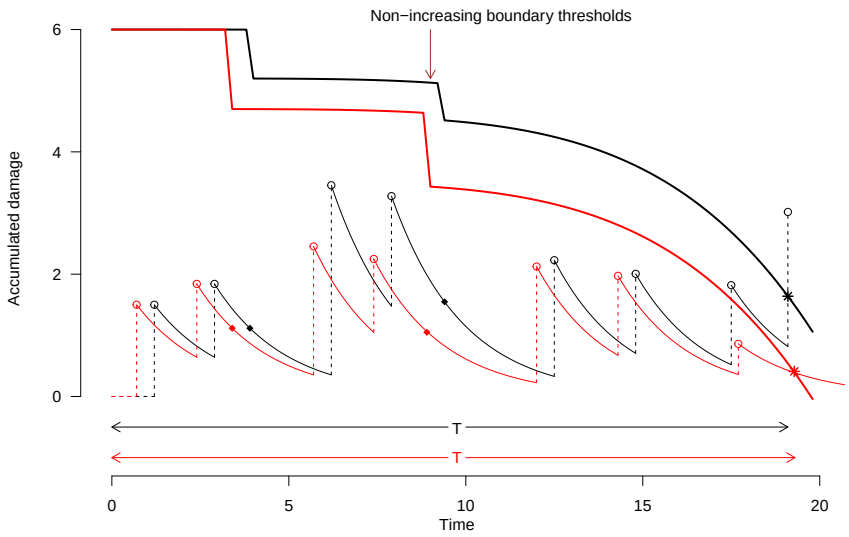


Figure 5.6. Depicting cumulative damage and the corresponding boundary curves (black and red curves) as shocks arrive randomly. The boundaries drop due to arrival of nonhealable shocks denoted by diamond shaped dots on the stochastic paths.

Figure 5.8 shows the expected cost per unit time as a function of d when $B(t) = 500 - t^2/50$. Table 5.6 displays the simulation results showing the optimal d for different choices of κ and boundary thresholds with different quadratic coefficients, but keeping the inter-arrival time distribution of healable shocks and their magnitudes such that their means are 1 and 10 units respectively; and the inter-arrival times of non-healbale shocks and their damage contributions such that their means are 5 and 3 units respectively. The given choice is considered to ensure that non-healable shocks are not more frequent than healable shocks.

5.4 Summary

In this chapter, we incorporated random magnitude of shocks and allowed the system to degrade over time due to aging while it heals at a constant rate. The main objective of our research has been to design a preventive maintenance policy. In each of the different scenarios

Table 5.5. For $p = 0.2$ proportion of all shocks nonhealable, for $X \sim$ Weibull(2,2/π), $Y \sim$ Weibull(2,1/2), and for various choices of κ and $B(t)$, the optimal d and [the associated t^*_γ] are displayed for $\gamma = 0.90, 0.85, 0.80$.

$B(t) = 500 - t^2/60$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	10.0 [0.263]	10.2 [0.367]	10.2 [0.476]
0.02	9.0 [0.462]	9.0 [0.613]	9.0 [0.754]
$B(t) = 500 - t^2/50$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	10.6 [0.263]	10.6 [0.378]	10.9 [0.485]
0.02	9.0 [0.425]	9.0 [0.598]	9.0 [0.745]
$B(t) = 500 - t^2/40$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	11.2 [0.263]	11.2 [0.374]	11.4 [0.480]
0.02	10.3 [0.398]	10.4 [0.545]	10.4 [0.676]

that we considered, we found an optimum d such that when the cumulative damage comes within d units of the boundary threshold an alarm sets off, and we replace the system after an additional duration dependent on a tolerable risk of failure. The optimization criterion is to minimize the cost per unit time. We allow different distributions for inter-arrival times and magnitude of shocks, different rates of healing, and different rates of degradation due to aging for the originally proposed system. Thus, we are not limited to using only certain types of inter-arrival time distributions. We also see that as the healing patterns change or the system degrades much faster, we wait for a relatively shorter duration of time before replacing the system in order to not risk failure.

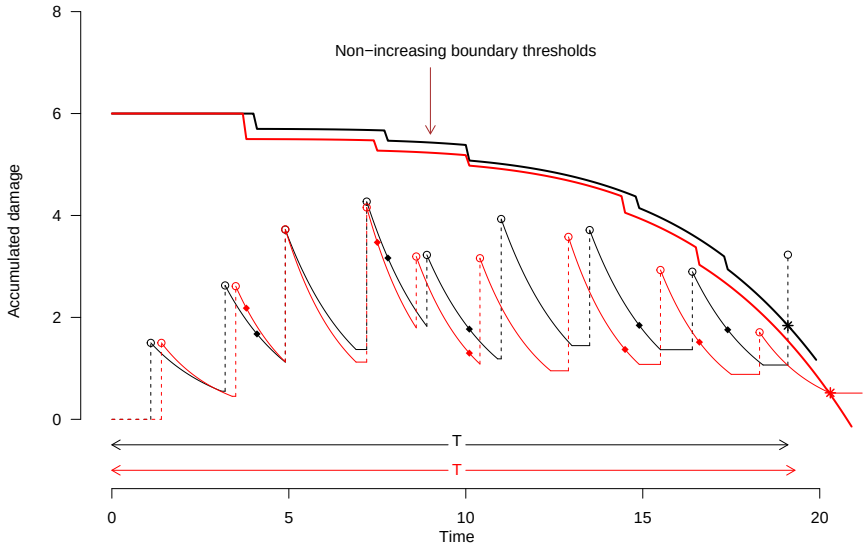


Figure 5.7. Depicting cumulative damage as shocks arrive randomly. The black and red stepwise decreasing curves represent the boundary threshold corresponding to the black and red sample paths respectively. Diamond shaped dots represent the arrival times of nonhealable shocks. For illustration we consider $\tau = 2$ and that nonhealable shocks arrive twice as faster as healable shocks.

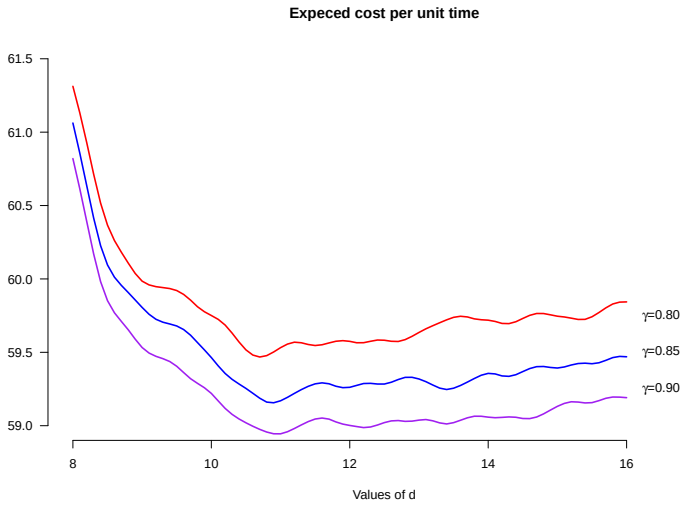


Figure 5.8. The expected cost per unit time is minimized at $d = 11$ for $\gamma = 0.90$; at $d = 11.3$ for $\gamma = 0.85$ and $d = 11.5$ for $\gamma = 0.80$. If there are multiple minima, choose the smallest one.

Table 5.6. For $X \sim$ Weibull(2,2/π), $Y \sim$ Weibull(2,1/2), $Z \sim$ Weibull(2,10/$\sqrt{\pi}$), $U \sim$ gamma (3,1), and for various choices of κ and $B(t)$, the optimal d and [the associated t^*_γ] are displayed for $\gamma = 0.90, 0.85, 0.80$.

$B(t) = 500 - t^2/60$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	11.0 [0.305]	11.0 [0.401]	11.3 [0.495]
0.02	10.8 [0.415]	11.0 [0.543]	11.1 [0.647]
$B(t) = 500 - t^2/50$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	11.3 [0.290]	11.3 [0.383]	11.3 [0.467]
0.02	11.1 [0.382]	11.1 [0.512]	11.1 [0.631]
$B(t) = 500 - t^2/40$			
κ	$\gamma = 0.90$	$\gamma = 0.85$	$\gamma = 0.80$
0.01	11.5 [0.278]	11.5 [0.365]	11.5 [0.448]
0.02	11.1 [0.382]	11.5 [0.502]	12.4 [0.605]

6. SUMMARY

In this chapter, we present the chapter-wise conclusions and the possible directions of future research.

6.1 Conclusions

Let us revisit the conclusions of each chapter to knit together the thematic message of this dissertation.

In Chapter 2, we study a one-unit repairable system supported by identical spare units on cold standby and repair facilities for which we remove the restrictive assumption of exponential life- or repair time distributions. The exponential distribution assumption is very common in the literature due to its lack of memory property, which ensures that the successive differences between life- or repair times are independent exponential variables (with different rates), and hence closed form expressions for the limiting average availability can be obtained. We allow arbitrary lifetime and arbitrary repair time distributions for any number of spare units and repair facilities by devising a discretization approach in which we inspect the system only at discrete time points; intervening only when during inspection we detect failure of unit(s) or revival of a failed system due to completion of at least one repair. In particular, we do not intervene at all even if a repair has been completed as long as the operating unit has not failed. Thus, this approach essentially discretizes the time variable and simplifies continuous monitoring to periodic monitoring (at inspection times only); thus, making it logistically preferable. We provide a simple computational approach by using the discretization which allows us to incorporate *any arbitrary* life- and repair time distributions as well as increase the number of repair facilities and/or the number of spare units. This broadens the horizon of research related to repair time distributions.

In Chapter 3, we shift focus to complex systems that are exposed to external impacts which can be broadly classified into two types: the damaging impacts are called valid shocks (VS) (we assume that each VS causes equal damage); and the ones that have a positive impact to the system are referred to as positive interventions (PI), accumulation of k of which triggers a healing effect, wherein the effect of one VS is nullified. The VS and PIs arrive

according to independent stochastic processes, and we essentially focus on the net count of shocks at any given time. In the literature of shock models, either shocks/impacts are assumed to have exponential inter-arrival times, or even if non-exponential inter-arrival times are mentioned, the illustrations are presented with only exponential examples. We generalize the inter-arrival time distribution of VSs and PIs to be *arbitrary*. Whenever we can count the number of VS and PIs, the distribution of Stage 1 duration and the lifetime of the system can be calculated. Furthermore, we consider that the system can lose its healing capacity as it ages, therefore, it is divided into Stage 1 (where it can heal) and Stage 2 (where PIs do not have any effect and the system cannot heal). We find that the distributions of Stage 1 duration T_1 and the system lifetime T can be described by a weighted convolution process with an adjustment involving a multiplier λ obtained numerically. Indeed, it is a function of the logarithm of the ratio of the standard deviations of inter-arrival time distributions of VS and PIs. The use of various distributions with choice of parameters yielding the same mean shows that we can allow any arbitrary inter-arrival distributions of shocks. Furthermore, we study three replacement policies, each of which optimizes the average cost per unit time to operate the system under different scenarios.

In Chapter 4 we basically maintain the same setup as in Chapter 3, the new addition being subdividing Stage 1 into two parts: initially the system heals at a faster rate requiring a few PIs to nullify one VS; but once enough net VS have accumulated, more PIs are needed to nullify one VS. We show that given a predetermined net number of shocks that the system can withstand in various stages, we can work out the distributions of Stage 1 duration T_1 and lifetime T_2 using a point process or an adjusted convolution process, as long as we can count the net number of VS. Moreover, we discover that subdivision of Stage 1 leads to an increase in Stage 1 duration, and hence system lifetime.

In Chapter 5, unlike in the previous chapters, we let external shocks inflict damage of varying magnitudes. Thus, we do not limit ourselves to considering only counts of shocks, but instead incorporate their magnitudes. We also allow the system to begin to heal instantaneously and continue to heal exponentially at a fixed rate κ (> 0) while also continually degrading due to aging. We have designed a time-dependent maintenance policy which focuses on risk assessment of the system at a given time as soon as the system comes

dangerously close (that is, within d units) to the boundary threshold, at which instant we are warned of a high probability of failure in the near future, and thus we determine an optimal d and associated replacement time by minimizing the cost per unit time. We study changes in healing behavior such as healing happening only for a fixed duration τ; changes in types of shocks, wherein with a certain probability p, some shocks are healable and the others are non-healable which leave some permanent damage to the system by suddenly degrading the system by a random amount; and also a combination of both types of shocks. Here too we allow arbitrary inter-arrival time distribution of all types of shocks. We make four important discoveries from this chapter:

- As the boundary degrades faster, the optimal d increases and the associated expected $t^*_\gamma(d)$ (the additional t^* units of time that the system can be allowed to function once it has reached within d units of the boundary with a probability of survival γ) decreases, which means due to higher risk, we allow the system to not come too close to the boundary threshold and also allow it to run for a shorter duration of time once the risk is detected; also when the healing rate becomes faster, we can allow the optimal d to be smaller and thus allow the system to function for a little bit longer.

- When shocks do not heal indefinitely, but rather for a fixed duration τ, their residual damage brings the collective damage closer to the boundary threshold much earlier. In general, we also see that the $t^*_\gamma(d)$'s are significantly lower and the optimal d's are larger as compared to the former setup implying that the alarm goes off when the distance from the boundary is larger and we wait a shorter duration after the alarm goes off to replace the system. Furthermore, if τ is shorter, the optimal d' s are even larger and the expected $t^*_\gamma(d)$'s are even smaller because healing continues for a shorter duration, increasing the risk of failure.

- When a fixed proportion p of shocks never heal; that is, their damage is permanent, we see that on average, as compared to Subsection 5.2.2, the overall waiting time until replacement after the alarm goes off is shorter.

- When there are two types of shock where the first type of shock is self-healable for a finite duration τ (or when a certain percentage of the damage heals and the rest is permanent), and the second type of shock is non-healable, we have similar interpretations: As healing patterns change or the system degrades much faster, we wait a relatively shorter duration of time before replacing the system to reduce the risk of failure.

This research provides a comprehensive view of different types of shocks and degradation rates. Although the simulations and illustrations consider some standard parametric distributions, the approach can easily be replicated for any type of distribution where parameters can be approximated from the data.

6.2 Directions of future research

6.2.1 Thoughts on research in reliability theory

Let us discuss some directions of future research in reliability theory and applied probabilistic modelling that may follow from each chapter of this dissertation.

The following are the possible directions from Chapter 2.

- One plausible extension is to increase the number of spare units and/or the number of repair facilities and study the changes in system availability. However, the inclusion of additional spares or repair facilities will lead to an increase in the number of states, and thus might lead to computational complexity.

- One may also consider implementing the proposed discretization method to study other systems, such as a k-out-of-$N : G$ system.

- Studies involving cost per unit time for repair and maintenance of spare units may be undertaken. Design policies to determine the optimal number of repair facilities to be established and the optimum number of spare units to be kept in hand so that the overall availability of the system is not compromised and at the same time the cost is within control.

- Moreover, one may possibly think about a more realistic repair situation since a perfect repair policy is often practically infeasible, and hence imperfect repair can be considered. The lifetime of an imperfectly repaired unit is stochastically shorter than that of a perfectly repaired unit, and the next imperfect repair time is stochastically larger than the previous repair time. Therefore, an intermediate repair policy may be significantly better than a perfect repair policy if the gain from early completion of repair and reduction of cost exceeds the loss due to shortening of the lifetime and expansion of the next repair time. Thus, a system may be subjected to several imperfect repairs before replacement.

From Chapters 3 and 4, the following potential future directions can be taken.

- Data-driven estimation techniques may be used to incorporate various semiparametric and non-paramteric forms of interarrival distributions of the VSs and PIs.
- Other replacement policies, such as a block replacement policy may be considered.
- Similar to the subdivision of Stage 1 in Chapter 4, further subdivisions are possible for various healing rates based on the magnitude of shocks.

Finally, Chapter 5 opens the following possibilities for future research.

- Different self-healing functions like the one described below can be incorporated.

$$h(t) = \begin{cases} \frac{\alpha}{(\alpha+\beta t)^{1+\gamma}} \text{ for } t \geq 0, \\ 0, \text{ otherwise} \end{cases} \tag{6.2.1}$$

  where $h(t)$ is the time-dependent healing function and $\alpha, \beta \in \mathbb{R}$, $\gamma \geq 0$ (Cui *et al.*, 2018).

- Different degradation processes, such as the gamma process, the Wiener process, etc. can be considered for the system.
- The inter-arrival time distribution of shocks may be from some other type of distribution such as phase-type distribution, non-homogeneous Poisson process, Hawkes process, Affine process etc.

- In addition to those described in the current research, various competing failure processes can be considered with healing. For example, in an electronic circuit board, one failure mode can be random voltage spikes, which cause damage by overloading the system, eventually causing failure. Another failure mode may be wearout, which usually occurs only after the system has been running for several cycles. The objective will be to determine the overall reliability of the components after N cycles and to find optimal policies to replace the system before risking failure.

6.2.2 Thoughts on statistical and computational issues

This research involved probabilistic modelling of various maintained systems. Moreover, the resultant simulations required heavy computational approaches. We have identified some potential statistical research as well as implementation challenges that could arise from such problems, which are described below.

- *The choice of thresholds in Chapter 3:* In the counting process of the number of arriving impacts, the thresholds for various states of the process are assumed to be predetermined. Some notable works in this direction are Chien *et al.* (2012); Zhao *et al.* (2018b); Cui *et al.* (2018); Dong *et al.* (2020b) among others. One way to determine the threshold values is by calculating the reliability of the system at any given time. Subsequently, we can determine the choice of thresholds m_1 and m_2 that achieve some specified measure of reliability. These may be empirically calculated using the inter-arrival times of valid shocks and positive interventions. As described in Dong and Cui (2019), identifying such thresholds remains an open problem. To bypass this issue, several time-based thresholds, commonly known as duration thresholds, are often considered by studying the reliability of the system at a given observation epoch.

  Alternatively, prior information on similar systems can be used to determine the optimal threshold based on the cost involved, losses incurred, average system downtime and remaining system lifetime (Rafiee *et al.*, 2015). We discovered in Chapter 4 that subdividing Stage 1 and further incorporating varied thresholds can increase the overall

lifetime of the system. Determining the thresholds mathematically is a challenging task which has potential for future research in Bayesian reliability.

- *The exact value of λ in Chapter 3:* We have established that λ is a function of the logarithm of the ratio of the standard deviations of F and G. However, we have not tried to find a closed-form expression of λ, as it was not the main focus of our research. The value of λ may be estimated using non-parametric estimation techniques such as kernel density estimation.

- *The choice of d in Chapter 5:* Because of randomness in the stochastic paths, the system fails when either a shock arrives or while the system is healing. Even though at times the cumulative damage comes very close to the boundary, the system may not fail. Moreover, even if at a given inspection epoch the cumulative damage is far from the boundary, two things can happen: (1) before the next inspection epoch, a shock arrives, making the cumulative damage high enough to cross the boundary, or (2) the cumulative damage crosses the boundary while the system was healing. Due to this randomness, there can be multiple minima when the expected cost per unit time is plotted against the various choices of d (as in Figure 5.4(b)). The cycle time in equation (5.2.3) is minimum of the failure time T and the replacement time $T' + t^*_\gamma(d)$. With some non-negative probability, sometimes the system might fail before the cumulative damage comes within d of the boundary. Therefore, finding a global minimum remains unsolved.

  Finding a suitable confidence interval of d based on the boundary function and the cumulative damage remains another open statistical problem. To find the confidence interval of d, we may use a data-driven approach such as the bootstrap method. This will give a range of values of d as a function of the inter-arrival times of shocks, their magnitudes, and the boundary function. Even though closed-form expressions of T and T' can be obtained for exponential inter-arrival time distribution of the incoming chocks, it will become increasingly challenging for other inter-arrival time distributions.

- *Computational aspect:* The approaches taken throughout the various chapters require heavy computation. For example, in Chapter 5, the simulation of the entire stochastic path for various choices of d over a range and the monitoring of the system at observation epochs in small increments of Δ, are computationally demanding. Parallel computing is the most efficient technique to perform the simulations simultaneously for different choices of the (hyper)parameters to search for the optimal solutions.

To summarize, the main focus of this dissertation has been to propose probabilistic modelling approach to systems supported by various spare and/or repair conditions or exposed to external stress factors such as shocks that have a deteriorating impact or impetus that can sometimes induce healing from the damages. We have established policies to maximize reliability measures such as availability and system lifetime. The results are illustrated by taking various choices of parameters and hyperparameters to define the conditions under which the systems function. The methodologies undertaken in this research enable us to study complex systems and various system environments in the future and will be beneficial for studying the reliability of not only industrial systems but also systems found in economic, environmental, biological, and actuarial sciences.

REFERENCES

Barlow R. E. and Proschan F. (1996) *Mathematical Theory of Reliability*, SIAM.

Bhuyan P. and Dewanji A. (2017) Estimation of reliability with cumulative stress and strength degradation, *Statistics* **51**(4), 766–781.

Biswas A. and Sarkar J. (2000) Availability of a system maintained through several imperfect repairs before a replacement or a perfect repair, *Statistics & Probability Letters* **50**(2), 105–114.

Biswas A., Sarkar J. and Sarkar S. (2003) Availability of a periodically inspected system, maintained under an imperfect-repair policy, *IEEE Transactions on Reliability* **52**(3), 311–318.

Chien Y.-H. and Sheu S.-H. (2006) Extended optimal age-replacement policy with minimal repair of a system subject to shocks, *European Journal of Operational Research* **174**(1), 169–181.

Chien Y.-H., Sheu S.-H. and Zhang Z. G. (2012) Optimal maintenance policy for a system subject to damage in a discrete time process, *Reliability Engineering & System Safety* **103**, 1–10.

Chien Y.-H., Sheu S.-H., Zhang Z. G. and Love E. (2006) An extended optimal replacement model of systems subject to shocks, *European Journal of Operational Research* **175**(1), 399–412.

Chung K. L. (2001) *A Course in Probability Theory*, Academic press.

Cui L., Chen Z. and Gao H. (2018) Reliability for systems with self-healing effect under shock models, *Quality Technology & Quantitative Management* **15**(5), 551–567.

Cui L., Huang J. and Li Y. (2016) Degradation models with wiener diffusion processes under calibrations, *IEEE Transactions on Reliability* **65**(2), 613–623.

Cui L. and Xie M. (2005) Availability of a periodically inspected system with random repair or replacement times, *Journal of Statistical Planning and Inference* **131**(1), 89–100.

de Smidt-Destombes K. S., van der Heijden M. C. and van Harten A. (2004) On the availability of a k-out-of-n system given limited spares and repair capacity under a condition based maintenance strategy, *Reliability Engineering & System Safety* **83**(3), 287–300.

de Smidt-Destombes K. S., van der Heijden M. C. and van Harten A. (2007) Availability of k-out-of-n systems under block replacement sharing limited spares and repair capacity, *International Journal of Production Economics* **107**(2), 404–421.

Dong Q. and Cui L. (2019) A study on stochastic degradation process models under different types of failure thresholds, *Reliability Engineering & System Safety* **181**, 202–212.

Dong W., Liu S., Bae S. J. and Cao Y. (2021) Reliability modelling for multi-component systems subject to stochastic deterioration and generalized cumulative shock damages, *Reliability Engineering & System Safety* **205**, 107260.

Dong W., Liu S., Cao Y. and Bae S. J. (2020a) Time-based replacement policies for a fault tolerant system subject to degradation and two types of shocks, *Quality and Reliability Engineering International* **36**(7), 2338–2350.

Dong W., Liu S., Cao Y., Javed S. A. and Du Y. (2020b) Reliability modeling and optimal random preventive maintenance policy for parallel systems with damage self-healing, *Computers & Industrial Engineering* **142**, 106359.

Eryilmaz S. and Kan C. (2019) Reliability and optimal replacement policy for an extreme shock model with a change point, *Reliability Engineering & System Safety* **190**, 106513.

Gao H., Cui L. and Dong Q. (2020) Reliability modeling for a two-phase degradation system with a change point based on a wiener process, *Reliability Engineering & System Safety* **193**, 106601.

Gong M., Eryilmaz S. and Xie M. (2020) Reliability assessment of system under a generalized cumulative shock model, *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* **234**(1), 129–137.

Gorjian N., Ma L., Mittinty M., Yarlagadda P. and Sun Y. (2010) A review on degradation models in reliability analysis, *Engineering Asset Lifecycle Management* pp. 369–384.

Hoyland A. and Rausand M. (2009) *System Reliability Theory: Models and Statistical Methods*, John Wiley & Sons.

Huang T., Zhao Y., Coit D. W. and Tang L.-C. (2021) Reliability assessment and lifetime prediction of degradation processes considering recoverable shock damages, *IISE Transactions* **53**(5), 614–628.

Jia X., Xing L. and Song X. (2020) Aggregated markov-based reliability analysis of multi-state systems under combined dynamic environments, *Quality and Reliability Engineering International* **36**(3), 846–860.

Keedy E. and Feng Q. (2013) Reliability analysis and customized preventive maintenance policies for stents with stochastic dependent competing risk processes, *IEEE Transactions on Reliability* **62**(4), 887–897.

Kong D., Balakrishnan N. and Cui L. (2017) Two-phase degradation process model with abrupt jump at change point governed by wiener process, *IEEE Transactions on Reliability* **66**(4), 1345–1360.

Kong X. and Yang J. (2020) Reliability analysis of composite insulators subject to multiple dependent competing failure processes with shock duration and shock damage self-recovery, *Reliability Engineering & System Safety* **204**, 107166.

Lafont U., van Zeijl H. and van der Zwaag S. (2012) Increasing the reliability of solid state lighting systems via self-healing approaches: A review, *Microelectronics Reliability* **52**(1), 71–89.

Levitin G., Xing L. and Dai Y. (2015) Optimal loading of system with random repair time, *European Journal of Operational Research* **247**(1), 137–143.

Levitin G., Xing L. and Dai Y. (2017) Optimal loading of series parallel systems with arbitrary element time-to-failure and time-to-repair distributions, *Reliability Engineering & System Safety* **164**, 34–44.

Levitin G., Xing L. and Huang H. Z. (2019) Dynamic availability and performance deficiency of common bus systems with imperfectly repairable components, *Reliability Engineering & System Safety* **189**, 58–66.

Marshall A. W. and Olkin I. (2007) *Life Distributions*, vol. 13, Springer.

Nakagawa T. (2006) *Maintenance Theory of Reliability*, Springer Science & Business Media.

Nakagawa T. (2007) *Shock and Damage Models in Reliability Theory*, Springer Science & Business Media.

Qiu Q., Cui L. and Wu B. (2020) Dynamic mission abort policy for systems operating in a controllable environment with self-healing mechanism, *Reliability Engineering & System Safety* **203**, 107069.

Rafiee K., Feng Q. and Coit D. W. (2015) Condition-based maintenance for repairable deteriorating systems subject to a generalized mixed shock model, *IEEE Transactions on Reliability* **64**(4), 1164–1174.

Ranjkesh S. H., Hamadani A. Z. and Mahmoodi S. (2019) A new cumulative shock model with damage and inter-arrival time dependency, *Reliability Engineering & System Safety* **192**, 106047.

Robins J. M., van der Vaart A. and Ventura V. (2000) Asymptotic distribution of p values in composite null models, *Journal of the American Statistical Association* **95**(452), 1143–1156.

Ross S. M. (2014) *Introduction to Probability Models*, Academic press.

Ross S. M., Kelly J. J., Sullivan R. J., Perry W. J., Mercer D., Davis R. M., Washburn T. D., Sager E. V., Boyce J. B. and Bristow V. L. (1996) *Stochastic Processes*, vol. 2, Wiley New York.

Sarkar J. and Chaudhuri G. (1999) Availability of a system with gamma life and exponential repair time under a perfect repair policy, *Statistics & Probability Letters* **43**(2), 189–196.

Sarkar J. and Li F. (2006) Limiting average availability of a system supported by several spares and several repair facilities, *Statistics & Probability Letters* **76**(18), 1965–1974.

Sarkar J. and Sarkar S. (2000) Availability of a periodically inspected system under perfect repair, *Journal of Statistical Planning and Inference* **91**(1), 77–90.

Sarkar J. and Sarkar S. (2001) Availability of a periodically inspected system supported by a spare unit, under perfect repair or perfect upgrade, *Statistics & Probability Letters* **53**(2), 207–217.

Sen P. K. and Bhattacharjee M. C. (1984) *Nonparametric Estimators of Availability Under Provisions of Spare and Repair*, University of North Carolina at Chapel Hill. Institute of Statistics.

Shen J., Cui L. and Yi H. (2018) System performance of damage self-healing systems under random shocks by using discrete state method, *Computers & Industrial Engineering* **125**, 124–134.

Sheu S.-H. and Chien Y.-H. (2004) Optimal age-replacement policy of a system subject to shocks with random lead-time, *European Journal of Operational Research* **159**(1), 132–144.

Tekin M. and Eryilmaz S. (2019) An application of phase-type distributions in a reliability shock model, *Journal of Universal Mathematics* **2**(1), 16–21.

Usynin A. and Hines J. W. (2007) Uncertainty management in shock models applied to prognostic problems., in *AAAI Fall Symposium: Artificial Intelligence for Prognostics*, p. 137.

Wald A. (1944) On cumulative sums of random variables, *The Annals of Mathematical Statistics* **15**(3), 283–296.

Wang H. (2002) A survey of maintenance policies of deteriorating systems, *European Journal of Operational Research* **139**(3), 469–489.

Wang J., Bai G. and Zhang L. (2020) Modeling the interdependency between natural degradation process and random shocks, *Computers & Industrial Engineering* **145**, 106551.

Wang N., Li M., Xiao B. and Ma L. (2019) Availability analysis of a general time distribution system with the consideration of maintenance and spares, *Reliability Engineering & System Safety* **192**, 106197.

Wang Q., He Z., Lin S. and Li Z. (2017) Failure modeling and maintenance decision for gis equipment subject to degradation and shocks, *IEEE Transactions on Power Delivery* **32**(2), 1079–1088.

Wu W., Tang Y., Yu M. and Jiang Y. (2014) Reliability analysis of a k-out-of-n: G repairable system with single vacation, *Applied Mathematical Modelling* **38**(24), 6075–6097.

Wu W., Tang Y., Yu M., Jiang Y. and Liu H. (2018) Reliability analysis of ak-out-of-n: G system with general repair times and replaceable repair equipment, *Quality Technology & Quantitative Management* **15**(2), 274–300.

Zhao X., Cai K., Wang X. and Song Y. (2018a) Optimal replacement policies for a shock model with a change point, *Computers & Industrial Engineering* **118**, 383–393.

Zhao X., Guo X. and Wang X. (2018b) Reliability and maintenance policies for a two-stage shock model with self-healing mechanism, *Reliability Engineering & System Safety* **172**, 185–194.

Zhao X., Lv Z., He Z. and Wang W. (2019) Reliability and opportunistic maintenance for a series system with multi-stage accelerated damage in shock environments, *Computers & Industrial Engineering* **137**, 106029.