RELIABILITY AND MAINTENANCE FOR INTERRELATED SYSTEMS

by

MICHAEL JOHN ARMSTRONG

BSc, Royal Military College of Canada, 1985
MBA, University of Ottawa, 1991

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

FACULTY OF COMMERCE AND BUSINESS ADMINISTRATION

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June 1996

© Michael J Armstrong, 1996
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Faculty of Commerce and Business Administration
Department of ______________________

The University of British Columbia
Vancouver, Canada

Date 14 June 1996
Abstract

In this paper we consider three topics in reliability and maintenance. Their common feature is the explicit consideration of interdependence, either amongst components within a machine, amongst stages in a logistics chain, or amongst machines sharing maintenance facilities.

Chapter 1 examines measures of the contribution of a component towards a system's reliability; these measures include structural importance, marginal reliability importance, joint reliability importance, and link importance. We review these different measures and examine how they relate to each other. We show that some structural results previously derived for independent components continue to hold for associated components. We then propose extensions to each importance measure to cover components which have two failure modes rather than the conventional one, and show that these extensions largely retain the properties of the original definitions.

Chapter 2 examines a system containing one machine subject to random failure, for which we wish to determine a maintenance policy and a spare ordering policy. We consider the solvability and desirability of jointly optimizing these two traditionally separate policies. We discuss why the general form of the problem is intractable and show how it might be approached using bounds and heuristic policies; more importantly, we examine several special cases which are more tractable. Particular attention is paid to systems which contain only one spare at a time; we show that these systems have some convexity properties which facilitate cost minimization. We subsequently consider systems which allow more than one spare, but make use of other restrictions or assumptions to simplify analysis.

Chapter 3 examines how to coordinate maintenance for machines sharing repair facilities, a task which is commonly called the machine repairman problem. Existing studies of this problem make the restrictive assumption that repairs are performed only after machine failure; our work extends the machine repairman model by allowing
preventive repair. We derive analytical results for a single critical age repair policy, and then use numerical studies to investigate other heuristic policies which can improve the system's cost effectiveness.
# Table of Contents

Abstract .................................................................................................................. ii  
Table of Contents .................................................................................................. iv  
List of Tables ......................................................................................................... vi  
List of Figures ....................................................................................................... vi  
Acknowledgements ................................................................................................. vii  
Foreword .................................................................................................................. vii  

Introduction ............................................................................................................ 1  

Terminology ........................................................................................................... 4  

Importance of Components...................................................................................... 7  
1.01 Chapter Abstract ............................................................................................. 7  
1.02 Introduction ..................................................................................................... 7  
1.03 Notation and Assumptions ............................................................................. 9  
1.04 Measures of Importance and Their Relationship ........................................ 10  
1.05 Numerical Example: The Sign of JRI .......................................................... 24  
1.06 Systems of Dual Failure Mode Components .................................................. 26  
1.07 Importance of Components for DFM Models .............................................. 28  
1.08 Future Research ............................................................................................. 37  
1.09 Concluding Remarks ...................................................................................... 37  

Maintenance and Inventory for a Single Machine .................................................. 39  
2.01 Chapter Abstract ............................................................................................. 39  
2.02 Introduction ..................................................................................................... 39  
2.03 Model Description & Notation ....................................................................... 43  
2.04 Single Spare with Basic Structure ................................................................. 45  
2.05 Early Ordering with Replacement on Receipt ............................................... 63  
2.06 Single Spare With Returns ............................................................................ 64  
2.07 Single Spare Extended Cost Structure........................................................... 70  
2.08 Scheduled versus Unscheduled Orders .......................................................... 71  
2.09 Single Spare, Random Lead Time ................................................................ 75  
2.10 Single Spare, Two Random Lead Times ......................................................... 78  
2.11 Multiple Spares without Stock-outs ............................................................... 83  
2.12 The General Problem ..................................................................................... 87  
2.13 Future Research ............................................................................................. 88  
2.14 Concluding Remarks ...................................................................................... 89  

The Machine Repairman Problem with Preventive Maintenance ......................... 90  
3.01 Chapter Abstract ............................................................................................. 90  
3.02 Introduction ..................................................................................................... 90  
3.03 Basic Model and Notation ............................................................................. 92  
3.04 The Simple Single Age Repair Policy ........................................................... 94
List of Tables

Table 1-1: Calculated JRI with Cov\{x_2,x_5\} = 0.00 ........................................... 25
Table 1-2: Calculated JRI with Cov\{x_2,x_5\} = +0.02 ........................................... 25
Table 1-3: Calculated JRI with Cov\{x_2,x_5\} = -0.02 ........................................... 25

Table 2-1: Example Solutions .................................................................................. 60
Table 2-2: Results Summary .................................................................................... 61

Table 3-1: Two machines, low failure variance (24 cases) ........................................ 110
Table 3-2: Three machines, low failure variance (12 cases) ........................................ 110
Table 3-3: Two machines, high failure variance (24 cases) ....................................... 111
Table 3-4: Three machines, high failure variance (12 cases) ...................................... 111
Table 3-6: Mean Improvement by Number of Machines ........................................... 115
Table 3-7: Mean Improvement by Other Factor ....................................................... 115

List of Figures

Figure 1-1: Structure of three component system .................................................... 10
Figure 1-2: Minimum Path Representation of an Arbitrary System ......................... 20
Figure 1-3: Components Re-Ordered ...................................................................... 20
Figure 1-4: New Modules Formed .......................................................................... 21
Figure 1-5: Structure of Five Component System ................................................... 24

Figure 2-1: Timing of Events .................................................................................. 46
Figure 2-2: Possible locations of KKT solution points ............................................. 53
Figure 2-3: Timing of Events with Random Lead Times .......................................... 76
Figure 2-4: Relations Amongst Single Spare Models .............................................. 82

Figure 3-1: A CQN with four machines and two repair shops ............................... 94
Figure 3-2: Single Age Policy .............................................................................. 100
Figure 3-3: Double Age Policy ............................................................................ 102
Figure 3-4: WUA Policy ..................................................................................... 102
Figure 3-5: Average Age Policy .......................................................................... 104
Figure 3-6: Constant Cost Policy ........................................................................ 104
Figure 3-7: Typical Optimal Policies .................................................................... 111

Figure B-1: Interior KKT Point ............................................................................ 130
Figure B-2: Diagonal Boundary KKT Points ......................................................... 134
Figure B-3: Interior KKT Point Again ................................................................... 136
Figure B-4: Interior and Diagonal KKT Points ...................................................... 138
Figure B-5: Lower Boundary and Diagonal KKT Points ....................................... 138
Acknowledgments

I wish to thank my thesis supervisor Derek Atkins for the encouragement and patience he showed me during my stay here at UBC; working with him helped to make the long and often frustrating task of writing a thesis a much more agreeable experience. Thanks are also due to Hong Chen, Bertrand Clarke, Moren Levesque, Ernie Love, Tom McCormick, Maurice Queyranne, Marty Puterman, and Daning Sun for their collective help over the past four years.

During this research the author received funding from the Department of National Defence and from Natural Science and Engineering Research Council grant OGP 0004743.

Foreword

Some of the material contained in this thesis has already been accepted for journal publication. The following describes the publication status at the time of this writing:

(a) a paper [Ar95] containing an earlier version of the work in section 1.04 on joint reliability importance was published in IEEE Transactions on Reliability in 1995; and,

(b) a paper [AA96] containing an earlier version of the work in section 2.04 was published in IIE Transactions in 1996.

Note to the Reader:

The paper [AA96] was coauthored by myself and Michael Armstrong. Michael was listed as the first author, and he was in fact the principal author who derived most of the results and wrote most of the text. My contribution was in the role of advisor, in that I assisted him with some of the more difficult parts of some proofs and helped him to edit his work into an appropriate style for our field of research. I therefore think it appropriate that this work be included in his thesis.

Derek Atkins

Professor of Management Science
Introduction

In this thesis we consider several problems in the field of reliability and maintenance. Much of the classic work in this area analyzes a single component or machine in isolation; this simplifies reliability analysis and maintenance optimization, but obviously does not reflect the reality of many applications in which a machine is part of and interacts with a larger system. A feature common to all of our work herein is an explicit consideration of system interdependence, whether it be amongst different components within a system, amongst successive stages in a logistics chain, or amongst several machines sharing common maintenance facilities.

Chapter 1 considers how to evaluate the contribution of individual components to the functioning of the overall system. There currently exist several measures of the importance of a component within a system in determining the system's reliability; these measures include structural importance, marginal reliability importance, joint reliability importance, and link importance. We review these different measures and examine how they relate to each other. We show that structural results previously obtained for joint reliability importance of independent components continue to hold for associated components. We then propose extensions to each importance measure to cover reliability models in which the components have two failure modes rather than the conventional one, and show that these extensions have properties similar to those in the single failure mode model. The chapter concludes with some numerical examples which illustrate our results.

Our main contributions in this first chapter are as follows:
(a) we show that structural results previously obtained for joint reliability importance continue to hold when component failures are not independent of each other;
(b) we extend the various measures of importance to dual failure mode systems; and,
(c) we show how the various measures of importance relate to each other.

Chapter 2 examines the interaction between the maintenance and inventory functions of a system. We consider a system containing one machine subject to random
failure, for which we wish to determine an age replacement maintenance policy and a spare ordering inventory policy. Our goal is to evaluate the solvability and desirability of jointly optimizing these two traditionally separate policies. We discuss why the general form of the problem is largely intractable and then proceed to examine several special cases which are somewhat easier to analyze. Special attention is paid to systems which are only allowed to contain a single spare; we show that for these systems our problem has some convexity properties which make it amenable to cost minimization. We subsequently consider systems which allow more than one spare, but make use of other restrictions or assumptions to somewhat simplify the analysis.

For the general form of the problem considered, we are only able to suggest some heuristic solution procedures; exact optimality appears to be intractable, at least using our style of analysis. When we assume that we are to use a single age maintenance policy and can either borrow a spare or instantly expedite one, then we are able to derive the cost function and specify some of its properties. When we restrict our inventory policy to having a single spare in stock or on order at any point in time, then we are able to specify the optimal policy form, derive the cost function, and show how to determine some of the optimal values of the decision variables for this function. When we further restrict the lead time to be deterministic, we are able to determine the full set of optimal decision variables. We find that the objective function is pseudo-convex in each decision variable and any economically viable Karush-Kuhn-Tucker point is a global minimum. A numerical study shows that joint optimization of maintenance and inventory decisions can be worthwhile in comparison to traditional separate optimization. The cost difference between joint optimal and separate optimal solutions can be arbitrarily large in extreme cases, while for the numerical data we consider, joint optimization yields an average improvement of 3% over the best form of separate optimization with one tenth of our cases having a difference greater than 10%. We also note the effects of allowing spares to be returned unused by an
improving machine, of requiring that spares be used upon receipt, and of allowing the lead
time to depend upon the order type (i.e. scheduled or unscheduled).

Our main contributions in this second chapter are as follows:
(a) we enhance the existing single-spare optimization literature by introducing the
notions of breakage costs, return of spares, service constraints, and double age policies;
(b) we combine and generalize previous work in single-spare optimization with
regards to such analytical properties as pseudo-convexity of the cost function; and,
(c) we illustrate the importance and difficulties of joint optimization of maintenance
and inventory.

Chapter 3 examines the maintenance of machines sharing common repair facilities,
a situation often referred to as the machine repairman problem. Although this has been a
subject of research for quite some time, it seems that all existing studies have made the
reasonable but restrictive assumption that repairs are made only after machine failure, i.e.
preventive maintenance is ignored. Our work extends the machine repairman model by
allowing for preventive repair under the assumption that the machines have an increasing
hazard function. We derive analytical results for the case where the form of the repair
policy is restricted to a single critical age, and then investigate other heuristic policies
which can improve the system's cost effectiveness. Numerical studies are then used to
compare the performance of our assorted policies for small systems. Our work shows that
for the type of system we consider, single age policies are quite tractable to analyze and
optimize; but their performance is sub-optimal and can result in meaningful inefficiencies.
We can improve on this situation using heuristic policies; we find that these tend to be
most worthwhile in systems with more machines, low shortage costs, low failure variance,
and a high repair rate. Our conclusion from this chapter is that although we are unable to
describe the truly optimal policy structure, the heuristics we examine seem to perform well
enough to make consideration of more complicated policies unnecessary.
Our main contributions in this third chapter are as follows:

(a) we introduce the use of preventive repair into the study of the machine repairman problem;
(b) we derive some basic properties of the traditional single critical age repair policy as used in this context, and,
(c) we introduce and evaluate several heuristics that improve upon the performance of single age policies.

Terminology

In discussing our work we will make extensive use of terminology from the reliability literature and to a lesser extent from the inventory literature. In this section we briefly introduce many of these terms; for more detailed explanations, refer to [BP75], [SS90], and [Co87] for reliability terminology and [GRZ93] for inventory terminology. For surveys of the existing maintenance literature, see [CP91], [VF89], and [SA85]; for the inventory literature [GRZ93] provides a broad survey.

A conventional or binary-state machine (or component, module, system, etc.) has two possible states: working (or good) and failed (or bad). The reliability of a machine is the probability that it is working. A system of $n$ components is often represented by a probabilistic network or graph, in which each edge of the graph corresponds to a component. The probability of being able to trace a path of working components (also called a path-set) through the graph amongst a set of $k$ nodes is the $k$-terminal reliability of the system; in particular, if $k = n$ then we talk of the all terminal reliability, while if we are only interested in paths between two particular nodes (the source and sink) so that $k = 2$ we speak of the $s$-$t$ reliability of the system. A path which contains no extra edges is referred to as a minimum path. A cut in the graph (i.e. a set of failed components, also called a cut-set) that prevents the required paths from being traced results in system failure; if this cut contains no extra edges, then it is a minimum cut.
A module is any convenient self-contained group of components (or sub-graph) that can be represented by (condensed into) a single edge of the graph. Components are said to be in parallel if any one working is sufficient to allow a system (or module) to work, while they are in series if all are required to be working for the system to be working. The usage of the term parallel-series system seems to vary from paper to paper. In this thesis we use it to mean a system which can be sequentially decomposed into modules, sub-modules, and so on, so that as each module is decomposed the result is a set of sub-modules all either in parallel or in series with each other. This definition is more general than that of [BP75] but less than that of [Co87]. A component is irrelevant if its functioning or failure never has any effect on system operation; otherwise it is relevant. A binary-state system is coherent if all of its components are relevant and system reliability is a non-decreasing function of component reliability. The reliability function expresses the reliability of a system in terms of the reliability of its components, assuming that the behaviour of the components is statistically independent; it is a \( n \)-th order polynomial. The structure function expresses the state of the system in terms of the states of the components, regardless of any statistical dependence.

The lifetime of a stochastically failing component can be described by a (cumulative) distribution function and a corresponding density function (which we assume exists). The survival function is equal to one minus the distribution function. The hazard rate or failure rate of a component is the ratio of the density function to the survival function, and can conveniently be thought of as the instantaneous conditional probability of component failure as it ages. If the failure rate is (strictly) monotonically increasing, we say that the component has a (strictly) increasing failure rate or is (strictly) IFR. Similarly, if the failure rate is monotonically decreasing the component is said to be DFR, while if the failure rate is constant then the component is CFR. Corrective maintenance is performed on a component after it has already failed, while preventive maintenance occurs prior to failure. If a repair fully restores a component to its original condition (i.e.
makes it good-as-new) then we describe that as *perfect repair*. Conversely, *minimal repair* only returns a component to its condition just before failure (i.e. makes it bad-as-old).

Inventory systems can be classified according to how often they review stock levels and make ordering decisions. A system is said to use *continuous review* if stock levels are reviewed constantly so that orders may be placed at any time; a system which performs the review and ordering functions only at fixed intervals (i.e. periodically) is said to use *periodic review*. An *(s-1,s)* inventory policy is a continuous review policy which places an order for one new unit whenever it receives a demand for one, so as to always have a total of *s* units in stock or on order; this is also called a *one-for-one ordering policy*. 
Chapter 1
Importance of Components

1.01 Chapter Abstract
Each component in a system plays a role in determining the reliability of that system; some components are more important in this regard than are others. The existing measures of this importance include structural importance, marginal reliability importance, joint reliability importance, and link importance. In this chapter we discuss these measures and examine how they relate to each other. We show that structural results previously derived for joint reliability importance of independent components continue to hold for associated components. We then propose extensions to each importance measure to cover reliability models in which the components have two failure modes instead of the conventional one, and show that these extensions have properties similar to those of the original versions. Throughout our discussion we present some numerical examples which illustrate our results.

1.02 Introduction
Consider a system, such as the radar system in an aircraft, which is composed of stochastically failing components. There are many questions that one could ask about such a system during its design, evaluation, and operation. How reliable is it? Why does it fail? If we make a particular change to one of the components, how does that affect the system as a whole? Given limited resources (e.g. a fixed engineering budget), how can we best improve the system's reliability (i.e. which components should be upgraded first)?

In answering such questions, the concept of the importance of components naturally arises. For example, if in a given system the failure of one component is more likely to lead to the system's failure than is the failure of another component, then we would intuitively consider the first component to be more important in some sense than the second. Similarly, more complicated questions of cost effectiveness and design priorities hinge on having some measure of each component's contribution or worth to the
system's reliable performance. For example, an aircraft's radar system has many components, each of which could potentially be made more reliable through redesign efforts, etc. The decision as to which of these components should receive priority for engineering improvements clearly would be made easier by having some method of ranking the components with respect to their contribution to system reliability.

To address such questions we have at our disposal several notions of importance that have been previously defined by other authors. Two of these, structural importance and reliability importance, have been in use for some time and are well described by Barlow & Proschan in [BP75]. More recently, Page & Perry [PP94] defined a measure they called link importance, and Hong & Lie [HL93] introduced the concept of joint reliability importance.

There are several closely related topics which are not addressed herein. We do not consider methods of comparing the reliability of entire networks, as was done by [AB84], nor do we address the question of determining the optimal configuration of components within a network, as in [BHP63] and [PM94]. For the reliability measures we do discuss, we make only passing reference to the computational issues involved in actually calculating their values, though we acknowledge that this is in itself an important issue; see [Co87] and [DR86] for discussions on this topic.

We begin our paper by reviewing the original definitions of these measures and examining how they relate to each other. We then show that structural results previously derived for joint reliability importance of independent components continue to hold for associated components. We propose extensions to all of these measures to cover components which have two possible failure modes rather than the conventional one, and show that the extended measures largely retain their original properties. We conclude with some numerical examples.
1.03 Notation and Assumptions

We consider a system of \( n \) components. The state of each component \( i = 1, \ldots, n \) is represented by indicator variable \( x_i \); for components with one failure mode, \( x_i \in \{1,0\} \), where 1 indicates that the component is working, and 0 indicates that it is failed. We use \( x \) to represent the vector of all component indicator variables, so \( x = (x_1, \ldots, x_n) \). The probability that component \( i \) is working is \( p_i = \Pr[x_i = 1] \), while the probability that component \( i \) is failed is \( q_i = 1 - p_i \). The vector of all component reliabilities is \( p = (p_1, \ldots, p_n) \). \( S(x) \) is the structure function for the system: \( S(x) = 1 \) if the system is working, 0 otherwise. \( R(p) \) is the reliability function for the system: \( R(p) = \Pr[S(x) = 1] = E[S(x)] \), so it gives the probability that the system is working. Sometimes we wish to evaluate one of these functions conditional on one or more component states being known; thus we write for example, \( S(1_i, x) \) to indicate the structure function for a system in which \( x_i = 1 \) (i.e. component \( i \) always works). Similarly we use \( R(0_i, x) = \Pr[S(x) = 1 | x_i = 0] \) (i.e. it is a conditional probability), or we can condition on the states of two variables as in \( R(0_i, 1_j, x) = \Pr[S(x) = 1 | x_i = 0, x_j = 1] \). We will introduce other notation later as required for our discussion.

We assume that we have a coherent system which can be modeled as a probabilistic network. For convenience only two-terminal (or s-t) reliability is discussed, but our results can also be shown to apply to all-terminal and \( K \)-terminal reliability using similar proofs with appropriately revised notation. The discussion is phrased in terms of a system and its components, though one could equally well speak of a graph and its edges or a set and its elements. We initially assume that our components have the conventional single failure mode, though later in the chapter this assumption will be relaxed.

In the course of our discussion we will refer to a small example system for which we calculate each of the various importance measures discussed. Figure 1-1 shows the structure of the system. For convenience in calculation it is assumed that all three
components shown are independent and identical, with a probability of working \( p = 0.8 \) and probability of being failed \( q = 0.2 \).

**Figure 1-1: Structure of three component system**

(a) Component version

(b) Graph version

The structure function for this system is

\[ S(x) = x_1 x_2 + x_1 x_3 - x_1 x_2 x_3 \]

and the reliability function is

\[ R(p) = p_1 p_2 + p_1 p_3 - p_1 p_2 p_3 = 2p^2 - p^3 \]

where the latter simplification is due to all components being identical. Note that by symmetry, components 2 and 3 must have the same importance value and ordering for all measures. The reliability of the system for the given component reliabilities is

\[ R(p) = 2(0.8)^2 - (0.8)^3 = 0.768 \]

In other words, at any given time there is about a 77% chance that the system is functional.

### 1.04 Measures of Importance and Their Relationship

*Structural importance* (SI) as defined in chapter 1 of [BP75] is one of the simplest measures of the importance of a component, as it ignores any consideration of the individual reliability of components and instead concentrates on the topological structure of the system.

**Definition of SI from [BP75]:**

\[ SI_i = 2^{-(n-1)} \sum [ S(1_j,x) - S(0_j,x) ] \text{ where the summation is over all } x \text{ such that } x_i = 1. \]
A system of \( n \) components has \( 2^n \) different state condition combinations, of which half have \( x_i = 1 \); hence the \( 2^{-(n-1)} \) term in the definition.

Intuitively, \( S_I \) is the proportion of possible states of the rest of the system in which the functioning of component \( i \) makes the difference between system failure and system functioning. Because \( S_I \) ignores component reliabilities, its value is easier to calculate than other measures, and it can be used even if these reliabilities are unknown or subject to change e.g. during the system design process. On the other hand, it can not distinguish between components which occupy similar structural positions but have drastically different reliabilities. In the example system of Figure 1-1, \( S_{I_1} = 1/4 \) \([0-0] + (1-0) + (1-0) + (1-0)\) = 0.75; a similar computation gives \( S_{I_2} = 0.25 \), so that \( S_{I_1} > S_{I_2} \) and \( S_{I_2} = S_{I_3} \).

The other measure of importance described in [BP75] is the reliability importance of a component (also known as marginal reliability importance or MRI).

**Definition of MRI from [BP75]:**

\[
MRI_i = E[ S(1_i,x) - S(0_i,x) ]
\]

\[
= R(1_i,p) - R(0_i,p) = \frac{\partial}{\partial p_i} R(p) \quad \text{for independent components.}
\]

Note that the definition uses conditional expectations; i.e. \( E[S(1_i,x)] \) is equivalent to \( E[S(x) \mid x_i = 1] \).

MRI is essentially the rate at which the system reliability changes with respect to changes in the reliability of a given component. An improvement in the reliability of the component with the highest MRI value causes the greatest increase in reliability of the overall system. This information can be used to determine which components should be improved first in order to make the largest improvement in the system. In certain cases, the structure of the system will indicate which components have the highest MRI; examples from [BP75] include pure parallel and pure series systems, while more complicated examples include the \( k\text{-out-of-}r\text{-of-}n \) linear systems described in [PS91]. For our example system, \( MRI_1 = [(0.8_1 + 0.8 - 0.6_4) - (0 + 0 - 0)] = 0.96 \) and similarly \( MRI_2 = 0.16 \), so \( MRI_1 > MRI_2 \) and \( MRI_2 = MRI_3 \).
MRI gives a more informative measure of the importance of a component than does SI, as MRI considers both the topology of the system and the reliability of the components in its determination. As noted in [BP75], $\text{SI}_i = \text{MRI}_i$ if the components are independent with reliabilities all equal to $1/2$. This means that under these conditions both measures give the same relative comparisons for any two components; i.e. when all components are independent with $p_i = p = 0.5$ for all $i$ we will have:

(a) $\text{SI}_i > \text{SI}_j \iff \text{MRI}_i > \text{MRI}_j$;

(b) $\text{SI}_i = \text{SI}_j \iff \text{MRI}_i = \text{MRI}_j$.

If the components are not independent or if the reliabilities are not identically one half, then these two measures need not agree; we could have $\text{SI}_i > \text{SI}_j$ but $\text{MRI}_i < \text{MRI}_j$. In general MRI provides more information than does SI, but of course it requires more information to calculate as well.

The link importance (LI) is a measure defined fairly recently by Page & Perry [PP94]. Whereas SI and MRI both provide numerical values for each component and hence determine a complete ordering (possibly including ties) of the components' importance, LI does not compute a value but provides only a partial ordering. There are actually two definitions in [PP94] for LI; we repeat them below, referring to the first one as LI proper and to the second one as pathset-cutset importance (PCI).

**Definition of LI from [PP94]:** $\text{LI}_i \geq \text{LI}_j$ (read, "component $i$ is at least as important in terms of LI as component $j$") if both of the following are true for any assignment of probabilities such that the reliabilities of all the components are independent and equal:

(a) $R(0, p) \geq R(0, p)$;

(b) $R(1, p) \leq R(1, p)$.

If both $\text{LI}_i \geq \text{LI}_j$ and $\text{LI}_i \leq \text{LI}_j$, then the two components are said to be equivalent in LI and so we write $\text{LI}_i \approx \text{LI}_j$.

The intuitive notion behind LI is that one component is more important than another if that first component makes the greater difference to system reliability both by its
functioning and by its failure. Note that the definition effectively assumes that all components are identical, and gives an ordering only if both inequalities hold true for any common reliability; MRI on the other hand is calculated using a specific set of reliabilities, which need not be identical. For the example system, \( L_1 \geq L_2 \) and \( L_2 \approx L_3 \).

What we refer to as PCI is defined only in terms of the network topology and ignores component reliabilities; in this way it is similar to SI. The idea behind PCI is to examine the size of the path sets and cut sets that two components are part of; the component which belongs to the smaller sets will tend to be more important to system operation than the one which belongs to larger sets because a component's state is relatively less influential in a larger set.

**Definition of PCI from [PP94]:** PCI, \( \geq \) PCI, (read, "component \( i \) is at least as important in terms of PCI as component \( j \)") if all four of the following are true:

(a) \( \{ \text{the number of components in a minimal size path set when } x_i = 1 \} \leq \{ \text{the number of components in a minimal size path set when } x_j = 1 \}. \) If equality holds, then also require that \( \{ \text{the number of minimal size path sets when } x_i = 1 \} \geq \{ \text{the number of minimal size path sets when } x_j = 1 \}; \)

(b) \( \{ \text{the number of components in a minimal size cut set when } x_i = 1 \} \geq \{ \text{the number of components in a minimal size cut set when } x_j = 1 \}. \) If equality holds, then also require that \( \{ \text{the number of minimal size cut sets when } x_i = 1 \} \leq \{ \text{the number of minimal size cut sets when } x_j = 1 \}; \)

(c) \( \{ \text{the number of components in a minimal size path set when } x_i = 0 \} \geq \{ \text{the number of components in a minimal size path set when } x_j = 0 \}. \) If equality holds, then also require that \( \{ \text{the number of minimal size path sets when } x_i = 0 \} \leq \{ \text{the number of minimal size path sets when } x_j = 0 \}; \)

(d) \( \{ \text{the number of components in a minimal size cut set when } x_i = 0 \} \leq \{ \text{the number of components in a minimal size cut set when } x_j = 0 \}. \) If equality holds, then also require
that \( \{ \text{the number of minimal size cut sets when } x_j = 0 \} \geq \{ \text{the number of minimal size cut sets when } x_j = 0 \} \).

If both \( PCI_i \geq PCI_j \) and \( PCI_i \leq PCI_j \), then \( PCI_i \approx PCI_j \). A minimal size path set is simply the path set which has the fewest components in it. In our example, \( PCI_1 > PCI_2 \) and \( PCI_2 \approx PCI_3 \).

The previous definition for LI was intended for systems of independent components; we can remove this requirement and revise the definition in terms of expectations as follows.

**Definition 1-1:** LI \(_i \geq LI_j \) (read, "component \( i \) is at least as important in terms of LI as component \( j \) ") if both of the following are true for any assignment of probabilities such that the marginal probabilities for each component's functioning are all equal:

(a) \( E[S(0_j, x)] \geq E[S(0_i, x)] \);
(b) \( E[S(1_j, x)] \leq E[S(1_i, x)] \).

If both \( LI_i \geq LI_j \) and \( LI_i \leq LI_j \), then \( LI_i \approx LI_j \).

Both LI and PCI are limited by being partial orderings, in that there are components that cannot be categorized by them; e.g. we could have two components such that neither \( LI_i \geq LI_j \) nor \( LI_i \leq LI_j \) nor \( LI_i \approx LI_j \). As illustrated by the examples in [PP94], LI is a stricter or more demanding test of relative importance than is PCI; that is, the use of LI results in fewer equivalencies or "ties" than PCI, but is more likely to be unable to express an ordering at all. It is therefore difficult to say which would be most useful in general; [PP94] does offer the following observation.

**Observation from [PP94]:** LI implies the following about PCI:

(a) \( LI_i \geq LI_j \implies PCI_i \geq PCI_j \);
(b) \( LI_i \approx LI_j \implies PCI_i \approx PCI_j \).

Page & Perry do not discuss how LI relates to SI or MRI, so we do that here. By comparing the respective definitions, we can see that the orderings determined by LI will be duplicated by SI, and also by MRI if the components all have identical reliabilities.
Proposition 1-1: LI implies the following about SI and MRI:

(a) $LI_i \geq LI_j \Rightarrow SI_i \geq SI_j$;
(b) $LI_i \approx LI_j \Rightarrow SI_i = SI_j$;
(c) $LI_i \geq LI_j \Rightarrow MRI_i \geq MRI_j$ when all components are identical;
(d) $LI_i \approx LI_j \Rightarrow MRI_i = MRI_j$ when all components are identical.

Proof:

Consider part (c) first. $LI_i \geq LI_j$ by definition means that both $E[S(0_j, x)] \geq E[S(0_i, x)]$ and $E[S(1_j, x)] < E[S(1_i, x)]$. From this it follows that

$$E[S(1_i, x)] + E[S(0_i, x)] \geq E[S(1_j, x)] + E[S(0_j, x)]$$

$$E[S(1_i, x)] - E[S(0_i, x)] \geq E[S(1_j, x)] - E[S(0_j, x)]$$

$MRI_i \geq MRI_j$.

For part (d), $LI_i \approx LI_j$ means that both: $LI_i \geq LI_j$, hence $MRI_i \geq MRI_j$; and also $LI_i \leq LI_j$, hence $MRI_i \leq MRI_j$. For both MRI inequalities to be true, we must have $MRI_i = MRI_j$.

For parts (a) & (b), set $p = 1/2$ and $q = 1/2$ to have $MRI = SI$; then (a) follows from (c) and (b) follows from (d).

QED

We conjecture that PCI has a similar relationship with SI but not with MRI, i.e. that PCI implies the following about SI:

(a) $PCI_i \geq PCI_j \Rightarrow SI_i \geq SI_j$;
(b) $PCI_i \approx PCI_j \Rightarrow SI_i = SI_j$.

We have been unable to prove this conjecture, however.

Aside from the observations above, the various importance measures need not agree on the ordering they provide. In terms of being able to discriminate amongst the importance of different components, we can rank these measures roughly in order of decreasing informativeness but increasing ease of calculation as follows: MRI, LI, PCI, SI. In actual application, we might expect to see SI, PCI or LI to be used during the initial
design of a new system, when component reliabilities are as yet unknown (these measures do not use component reliability in their computation). MRI would be most appropriate for detailed analysis of an existing system, when presumably information about component reliabilities and covariances is available.

The last measure of importance we discuss is joint reliability importance (JRI) which was introduced by Hong & Lie in [HL93]. JRI is quite different from previous measures: whereas SI, LI, PCI, and MRI all evaluate the effect of one component on the system reliability, JRI evaluates the effect of one component on another component's importance. This makes JRI potentially more useful for making the kind of trade-off decisions that commonly arise during system design, since it gives some indication of how components interact in determining the system's reliability. Intuitively, \( JRI_{ij} \) is the rate of increase in the importance of component \( i \) as we increase the reliability of component \( j \) (and vice versa).

**Definition from [HL93]:** For independent components:

\[
JRI_{ij} = R(1_i,1_j,p) + R(0_i,0_j,p) - R(0_i,1_j,p) - R(1_i,0_j,p) = \frac{\partial^2}{\partial p_i \partial p_j} R(p)
\]

This definition of JRI and the structural results derived in [HL93] are restricted to the case where component failures are assumed to be independent of each other. Unfortunately, in many real world systems the component failures are not independent. For example, all of the components in an aircraft radar system would presumably be subjected to many of the same stresses: the vibration from the engines, the shock of landing, the irregularities in the power supplied, etc. It is not unreasonable to therefore expect that there may be positive correlations amongst the components (i.e. if one circuit has failed due to a power surge, another is likely to have failed too). Less commonly, the independence assumption may be violated by the presence of negative correlations. The following definition extends the concept of JRI to more general systems where the components need not be independent.
**Definition 1-2:**

\[ \text{JRI}_{ij} = \mathbb{E}[ S(1_i, 1_j, x) + S(0_i, 0_j, x) - S(0_i, 1_j, x) - S(1_i, 0_j, x) ] \]

When components are independent this simplifies to the original definition, since then \( \mathbb{E}[S(x)] = R(p) \), and so for example \( \mathbb{E}[ S(1_i, 1_j, x) ] = R(1_i, 1_j, x) \). We can interpret this definition as follows. If we could guarantee that two components in a system would always work, we would expect about twice the improvement in system reliability that we would get from a guarantee for only one of the components. If this is actually the case, then JRI is zero. If there is more improvement ("synergy") then JRI is positive. If there is less improvement ("diminishing returns") then JRI is negative. For the example system, \( \text{JRI}_{12} = [1 + 0 - 0 - 0.8] = 0.2 \), which confirms that the reliability of component 2 becomes more critical when component 1 is improved (and vice versa); conversely, \( \text{JRI}_{23} = [0.8 + 0 - 0.8 - 0.8] = -0.8 \), which shows that component 2 becomes less critical when component 3 is improved (and vice versa).

The next lemma shows that JRI can be represented in terms of MRI in a modified system in which one component is in effect guaranteed either to be working or to be failed.

**Lemma 1-1:**

(a) \( \text{JRI}_{ij} = \text{MRI}_i(1_j) - \text{MRI}_i(0_j) \)

(b) \( \text{JRI}_{ij} = \text{MRI}_j(1_i) - \text{MRI}_j(0_i) \)

where we use \( \text{MRI}_i(1_j) \) to indicate that MRI is to be calculated while fixing \( x_j = 1 \), etc.

**Proof:**

\[
\text{MRI}_i(1_j) - \text{MRI}_i(0_j) \\
= \mathbb{E}\{ S(1_i, 1_j, x) - S(0_i, 1_j, x) \} - \mathbb{E}\{ S(1_i, 0_j, x) - S(0_i, 0_j, x) \} \\
= \mathbb{E}\{ S(1_i, 1_j, x) + S(0_i, 0_j, x) - S(0_i, 1_j, x) - S(1_i, 0_j, x) \} \\
= \text{JRI}_{ij}
\]

This proves part (a). The choice of \( i \) and \( j \) are arbitrary, so the proof of part (b) is identical.
This lemma suggests that the JRI of two components describes the relative importance of one component when the other is functioning. In particular:

(a) a positive JRI indicates that one component becomes more important when the other is functioning (the "synergy" case), as with components 1 and 2 in our example system;
(b) a negative JRI indicates that one component becomes less important when the other is functioning ("diminishing returns"), as with components 2 and 3 in the example; and,
(c) a zero JRI indicates that one component's importance is unchanged by the functioning of the other.

The following shows that the range of JRI is limited and that it is possible to find the sign of JRI without calculation in two special cases.

**Lemma 1-2:** JRI is bounded as follows:

(a) for any two components \(i\) and \(j\), \(\text{JRI}_{ij} \in [-1, 1]\);
(b) if two components \(i\) and \(j\) are in parallel, then \(\text{JRI}_{ij} \leq 0\);
(c) if two components \(i\) and \(j\) are in series, the \(\text{JRI}_{ij} \geq 0\).

**Proof:**

Part (a). The structure function \(S(x)\) only assumes values of 0 or 1, thus

\[
E\{ S(x) \} \in [0, 1]
\]

The system is coherent, thus \(E\{ S(I_j, x) \} \geq E\{ S(0_j, x) \}\)

Now \(\text{MRI}_j = E\{ S(I_j, x) - S(0_j, x) \} \in [0, 1]\)

So \(\text{JRI}_{ij} = \text{MRI}_i(1_j) - \text{MRI}_i(0_j) \in [-1, 1]\).

Part (b). If \(i\) and \(j\) are in parallel and \(j\) is working (i.e. \(x_j = 1\)), then \(i\) is irrelevant and \(\text{MRI}_i(1_j) = 0\). Using this with Lemma 1-1 gives:

\(\text{JRI}_{ij} = \text{MRI}_i(1_j) - \text{MRI}_i(0_j) = 0 - \text{MRI}_i(0_j) \leq 0\)

Part (c). If \(i\) and \(j\) are in series and \(j\) is failed (i.e. \(x_j = 0\)), then \(i\) is irrelevant and \(\text{MRI}_i(0_j) = 0\). Using this with Lemma 1-1 gives:

\(\text{JRI}_{ij} = \text{MRI}_i(1_j) - \text{MRI}_i(0_j) = \text{MRI}_i(1_j) - 0 \geq 0\)
More generally, there are four possible situations for the structural relationship of two components \( i \) and \( j \):

(i) \( i \) and \( j \) appear together on at least one minimum path, but not on any minimum cut;
(ii) \( i \) and \( j \) appear together on at least one minimum cut, but not on any minimum path;
(iii) \( i \) and \( j \) appear together on at least one minimum path and on at least one minimum cut; or,
(iv) \( i \) and \( j \) do not appear together on any minimum path nor on any minimum cut.

The last case (iv) is not of interest here, because it implies that at least one component is irrelevant and hence the system is not coherent. A special case of (i) is where components are in series, while a special case of (ii) is where components are in parallel; the previous lemma gives the sign of JRI for these two special cases. The following theorems show that the sign of JRI can be readily determined more generally for both cases (i) and (ii).

**Theorem 1-1:** The following two conditions are equivalent:

(a) there is no minimum path set containing both \( i \) and \( j \);
(b) \( \text{JRI}_{ij} \leq 0 \) for all possible assignments of probabilities.

**Proof:**

(We have two different ways of proving Theorems 1-1 and 1-2, so we will show one type of proof here for Theorem 1-1 and the other type subsequently for Theorem 1-2. Since the theorems are in a sense duals of each other, the proofs can easily be adapted from one theorem to the other by interchanging paths for cuts, etc.)

Assume (a) holds. Begin with an arbitrary system, and choose any two components \( i, j \) within that system. Create the minimum path representation of the system. This will be a collection of paths; the paths are in parallel with each other, and each path consists of one or more components in series. Some of these paths contain component \( i \), some contain component \( j \), and some contain neither; by hypothesis, no path contains both components. Any component which appears on multiple paths can be
thought of as multiple components \((comp_1, comp_2, \ldots)\) with identical reliabilities \((p_{comp_1} = p_{comp_2} = \ldots)\) and perfect correlations \((\text{Cor}_{x_{comp_1}, x_{comp_2}} = 1)\). See Figure 1-2 for a pictorial example.

**Figure 1-2: Minimum Path Representation of an Arbitrary System**

![Diagram of a minimum path representation of an arbitrary system with components labeled i, j, and other.]

Since each path is a simple series, the order of components does not affect the value of the structure function; so arbitrarily re-arrange the components so that \(i\) and \(j\) are the last components on any path they appear on. See Figure 1-3.

**Figure 1-3: Components Re-Ordered**

![Diagram with components i, j, and other re-ordered.]

Combine components into arbitrary modules as follows:

(i) all occurrences of \(i\) are recombined into a single \(i\) component;
(ii) all occurrences of \(j\) are recombined into a single \(j\) component;
(iii) all other components on paths containing \(i\) are combined to form module 1;
(iv) all other components on paths containing $j$ are combined to form module 2; and,
(v) all other components on paths containing neither $i$ nor $j$ are combined to form a third module.

See Figure 1-4 for the resulting structure.

**Figure 1-4: New Modules Formed**

The structure function of this system of components and modules has the form

$$S(x_i, x_j, Y) = Y_1x_1 + Y_2x_j + Y_3 - Y_1x_1Y_2x_j - Y_1x_1Y_3 - Y_2x_jY_3 + Y_1x_1Y_2x_jY_3$$

where $Y_j$ is the indicator random variable for the functioning of module $i$.

So now the expression for JRI is:

$$JRI_{ij} = E\{ S(1_i, 1_j, Y) + S(0_i, 0_j, Y) - S(0_i, 1_j, Y) - S(1_i, 0_j, Y) \}$$

$$= E\{ Y_1 + Y_2 + Y_3 - Y_1Y_2 - Y_1Y_3 - Y_2Y_3 + Y_1Y_2Y_3 \} + (Y_3)$$

$$- (Y_1 + Y_3 - Y_1Y_3) - (Y_2 + Y_3 - Y_2Y_3) \}$$

$$= E\{ Y_1Y_2Y_3 - Y_1Y_2 \} + E\{ Y_1Y_2Y_3 \} - E\{ Y_1Y_2 \} \leq 0$$

This shows that (a) $\Rightarrow$ (b) holds for any pair of components, regardless of any covariances (or for any assignment of probabilities in general). Get (b) $\Rightarrow$ (a) by reversing the order of the steps.

**QED**

**Theorem 1-2:** The following two conditions are equivalent:

(a) there is no minimum cut set containing both $i$ and $j$;

(b) $JRI_{ij} \geq 0$ for all possible assignments of probabilities.
**Proof:**

The proof of this theorem makes use of the properties of supermodularity, as described in [Lo83]. A supermodular set function \( M() \) is one which satisfies the following inequality for any sets \( C \) and \( D \):

\[
M(C \cup D) + M(C \cap D) \geq M(C) + M(D).
\]

Let \( A_1, A_2, ..., A_k \) denote \( k \) random events. For every subset \( C = \{ i(1), ..., i(c) \} \) of the index set \( \{ 1, ..., k \} \), define the joint probability function \( M(C) = \Pr\{ A_i(1) \cap A_i(2) \cap ... A_i(c) \} \). Then \( M(C) \) is a supermodular function.

First show that \( (a) \Rightarrow (b) \). Define the following random events.

\( A_1 \) the event that all minimum cut sets containing component \( i \) are functional

\( A_2 \) the event that all minimum cut sets containing component \( j \) are functional

\( A_3 \) the event that all minimum cut sets containing neither \( i \) nor \( j \) are functional

By hypothesis, there is no minimum cut set containing both components \( i \) and \( j \). Also define \( C = \{ 1, 3 \} \) and \( D = \{ 2, 3 \} \).

By definition, \( JRI_{ij} \) is

\[
= E\{ S(i_1, l_j, x) + S(0_i, 0_j, x) - S(l_i, l_j, x) - S(0_i, l_j, x) \}
\]

\[
= Pr\{S(l_i, l_j, x)=l\} + Pr\{S(0_i, 0_j, x)=l\} - Pr\{S(l_i, l_j, x)=l\} - Pr\{S(0_i, l_j, x)=l\}
\]

\[
= Pr\{A_3\} + Pr\{A_1 \cap A_2 \cap A_3\} - Pr\{A_1 \cap A_3\} - Pr\{A_2 \cap A_3\}
\]

\[
= M(C \cap D) + M(C \cup D) - M(C) - M(D) \geq 0
\]

where the inequality follows from the supermodularity of the joint probability function.

Next show that \( \neg (a) \Rightarrow \neg (b) \).

Suppose there exists a minimum cut set \( C \) containing both \( i \) and \( j \). Since \( C \) is minimal, there exists a path set \( V_i \) (respectively, \( V_j \)) with \( V_i \cap C = \{ i \} \) (respectively, \( V_j \cap C = \{ j \} \)). If we define the component reliabilities \( p_k \) according to \( p_k = 0 \ \forall \ k \in C \setminus \{ i, j \} \) and \( p_k = 1 \ \forall \ k \notin C \), then we get \( JRI_{ij} \) as follows:

\[
= E\{ S(1_i, 1_j, x) + S(0_i, 0_j, x) - S(0_i, 1_j, x) - S(1_i, 0_j, x) \}
\]
\[= \Pr\{S(1,1,x) = 1\} + \Pr\{S(0,0,x) = 1\} - \Pr\{S(0,1,x) = 1\} - \Pr\{S(1,0,x) = 1\}\]
\[= 1 + 0 - 1 - 1 = -1 \leq 0\]

**QED**

We note that the qualitative analysis of networks done by Granot & Veinott in another context in [GV85] produces what seems to be related results regarding the sign of qualitative changes. Also note that the corresponding theorems in [HL93] showed that (a) \(\Rightarrow\) (b) assuming independence, whereas now we say (a) \(\Leftrightarrow\) (b) in general. The JRI values in our example system coincide with the theorem results: \(\text{JRI}_{12} > 0\) and \(\text{JRI}_{23} < 0\).

We can further show that JRI is strictly non-zero in many cases.

**Proposition 1-2:** Suppose that \(0 < p_m < 1\) and \(-1 < \text{Cor}\{x_m, x_n\} < 1\) for all components \(m, n\), then:

(a) under the conditions of Theorem 1-1 part (a), \(\text{JRI}_{ij} < 0\);

(b) under the conditions of Theorem 1-2 part (a), \(\text{JRI}_{ij} > 0\);

(c) in a system with a parallel-series structure, \(\text{JRI}_{ij} \neq 0\) for all pairs \(ij\).

**Proof:**

Part (a): Examine the last line of the proof for Theorem 1-1. Our hypothesis permits us to state that
\[E\{Y_1Y_2Y_3\} < E\{Y_1Y_2\}\] and so \(\text{JRI}_{ij} < 0\).

Part (b): Follows the same type of reasoning as in part (a) but using Theorem 1-2.

Part (c): Recall that in a series-parallel network, each module can be broken down into sub-modules which are either all in series or all in parallel within that module. This can be repeated with sub-sub-modules, etc., until all modules have been broken down into components.

Begin with the system as a single module linking \(s\) and \(t\). Start breaking-down the modules until components \(i\) and \(j\) first appear in different sub-modules. If these two sub-modules are in series with each other, then \(i\) and \(j\) cannot appear on the same minimum cut; apply part (b) above. If the sub-modules are in parallel, then \(i\) and \(j\) cannot appear on
the same minimum path; apply part (a) above. Of course these relative locations of \( i \) and \( j \) are preserved in any subsequent decompositions.

**QED**

The above results describe the intuitive idea that in most real systems every component's importance is affected by every other component's reliability to some degree. The extra conditions are not very restrictive; in a sense, they only rule out "trivial" components (e.g. \( p_m = 1 \) implies that a component is always working; in the unlikely event that this were true in a real system, it would not be of interest to reliability analysis in the conventional sense). Note that even when none of these propositions hold, JRI is exactly zero only for particular combinations of component reliabilities (typically, at a combination of \( p_m \) values where the JRI is crossing over from positive to negative, or vice versa).

There seems to be no direct relationship in general between JRI and measures of importance other than MRI; e.g. \( \text{JRI}_{ij} > 0 \) implies nothing about \( \text{LI}_i \) and \( \text{LI}_j \) in general.

**1.05 Numerical Example: The Sign of JRI**

To illustrate the relationship between JRI's sign and a system's topology, in Tables 1-1, 1-2, and 1-3, we show selected JRI values for a system which has the same bridge structure as the first example in [HL93]; this structure is shown in Figure 1-5.

**Figure 1-5: Structure of Five Component System**
The reliabilities of all the components are assumed to have the same value $p_m$, and all covariance terms are zero except for Cov{$x_2,x_5$} which is specified above each table.

Note that by symmetry, JRI$_{12}$ = JRI$_{45}$ and JRI$_{13}$ = JRI$_{43}$.

Table 1-1: Calculated JRI with Cov{$x_2,x_5$} = 0.00

<table>
<thead>
<tr>
<th>$p_m$</th>
<th>.1</th>
<th>.3</th>
<th>.5</th>
<th>.7</th>
<th>.9</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRI$_{12}$</td>
<td>.972</td>
<td>.784</td>
<td>.500</td>
<td>.216</td>
<td>.028</td>
<td>.007</td>
</tr>
<tr>
<td>JRI$_{14}$</td>
<td>.018</td>
<td>-.126</td>
<td>-.250</td>
<td>-.294</td>
<td>-.162</td>
<td>-.090</td>
</tr>
<tr>
<td>JRI$_{13}$</td>
<td>.072</td>
<td>.084</td>
<td>.000</td>
<td>-.084</td>
<td>-.072</td>
<td>-.043</td>
</tr>
</tbody>
</table>

Table 1-2: Calculated JRI with Cov{$x_2,x_5$} = +0.02

<table>
<thead>
<tr>
<th>$p_m$</th>
<th>.1</th>
<th>.3</th>
<th>.5</th>
<th>.7</th>
<th>.9</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRI$_{12}$</td>
<td>.972</td>
<td>.784</td>
<td>.500</td>
<td>.216</td>
<td>.028</td>
<td>.007</td>
</tr>
<tr>
<td>JRI$_{14}$</td>
<td>-.014</td>
<td>-.114</td>
<td>-.230</td>
<td>-.266</td>
<td>-.126</td>
<td>-.052</td>
</tr>
<tr>
<td>JRI$_{13}$</td>
<td>.056</td>
<td>.076</td>
<td>.000</td>
<td>-.076</td>
<td>-.056</td>
<td>-.025</td>
</tr>
</tbody>
</table>

Table 1-3: Calculated JRI with Cov{$x_2,x_5$} = -0.02

<table>
<thead>
<tr>
<th>$p_m$</th>
<th>.1</th>
<th>.3</th>
<th>.5</th>
<th>.7</th>
<th>.9</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRI$_{12}$</td>
<td>.972</td>
<td>.784</td>
<td>.500</td>
<td>.216</td>
<td>.028</td>
<td>.007</td>
</tr>
<tr>
<td>JRI$_{14}$</td>
<td>-.022</td>
<td>-.138</td>
<td>-.270</td>
<td>-.322</td>
<td>-.198</td>
<td>-.128</td>
</tr>
<tr>
<td>JRI$_{13}$</td>
<td>.088</td>
<td>.092</td>
<td>.000</td>
<td>-.092</td>
<td>-.088</td>
<td>-.061</td>
</tr>
</tbody>
</table>

As an example of how these table entries are calculated, consider JRI$_{12}$ when $p_m = 0.5$. Then we have:

$$ JRI_{12} = E[ S(1_i, 1_j, x) + S(0_i, 0_j, x) - S(0_i, 1_j, x) - S(1_i, 0_j, x) ] $$

$$ = (1) + (0.5^2) - (0.5^2 + 0.5^2 - 0.5^3) - (0.5^2 + 0.5^2 - 0.5^3) $$

$$ = 0.500 $$

Note that the inclusion of non-zero covariances changes the values calculated but not the sign of the values. We see that the signs of these values agree with those predicted by our results. Component pair (1,2) has no minimum cut in common, and so JRI$_{12}$ is positive. Component pair (1,4) has no minimum path in common, and so JRI$_{14}$ is negative. Component pair (1,3) has both a minimum cut and a minimum path in common; our theorems do not indicate what the sign will be, and as the table shows JRI$_{13}$ can be
either positive or negative depending on the $p_i$ values. The fact that $\text{JRI}_{13}$ can have either sign indicates that the relationship between components 1 and 3 depends on their reliabilities; in other words, the priorities for component upgrading are different if the current reliabilities are all quite low than if they are all very high.

This example obviously is simple enough that the conclusions of the above paragraph could easily be obtained without recourse to our theorems. Imagine, however, a much more complicated system being analyzed to determine the effect of multiple failures on system performance. In this kind of study, it would be important for reliability engineers to understand not only the effect of individual component failures on the system reliability, but also the effect on the criticality of other components in the system. The engineers can use Theorems 1-1 and 1-2 to get an initial guide as to how the components interact, and later calculate the actual JRI and MRI values to understand more precisely the relationships.

1.06 Systems of Dual Failure Mode Components

Our discussion so far has made the conventional assumption that a component has only a single failure mode and so can occupy one of two mutually exclusive states, i.e. working or failed; such components are often referred to as binary or binary-state components. In the remainder of this chapter, we consider models which allow for two failure modes and hence three mutually exclusive states: working, failed open, or failed short; we refer to these as dual failure mode (DFM) components. DFM models arise naturally in the study of flows through electrical circuits and pipelines, where the way in which a component fails can be quite important. A DFM component is working ($x_i = 1$) if it functions as designed (e.g. it can regulate the flow of electricity through itself); it is failed open ($x_i = fo$) if there is no flow through it at all (e.g. an open electrical circuit); and it is failed short ($x_i = fs$) if it allows an infinite or uncontrolled flow through itself (e.g. a shorted electrical circuit). We indicate the probability that a component $i$ is failed short by $qs_i = \Pr[x_i = fs]$, and the probability that component $i$ is failed open by $qo_i = \Pr[x_i = fo]$. 
Similarly, an entire system is said to be failed short if there is any path set consisting exclusively of shorted components, it is failed open if there is at least one cut set of open components, and it is working if it is neither failed short nor failed open (i.e. there exists no shorted path and no open cut). Defined in this way, the three system states are also mutually exclusive. Note that there are other ways to define system performance for DFM components; see [SSYY93] and [SS90] for discussions on this point. We shall see that at both the system and component levels, the failed open state is equivalent and has similar behaviour to a conventional failure in the binary-state model, while the failed short state is quite different in its effect. Note that the reliability function \( R(p) \) and the structure function \( S(x) \) have the same interpretation with DFM components as was the case with binary components, though their form becomes more complicated.

Page & Perry [PP87] showed that one could calculate the reliability of a DFM system using two binary-state systems with the same topology.

**Theorem for the Calculation of DFM reliability from [PP87]:**

\[
R(p) = R(p'') - R(p')
\]

where \( R(p'') = \text{Pr}[\text{no open cut}] \) is calculated using \( p_i'' = p_i + q_s i \) and \( q_i'' = q_o i \), while \( R(p') = \text{Pr}[\text{short path}] \) is calculated using \( p_i' = q_s i \) and \( q_i' = p_i + q_o i \). \( R(p'') \) and \( R(p') \) are calculated using the binary-state reliability function corresponding to a system with the same network topology as the original DFM system, but with the probabilities adjusted as shown. Note that \( R(p'') \) is a reliability in the conventional sense, whereas \( R(p') \) is effectively a form of unreliability figure for the DFM system.

The preceding definition is used in [PP87] for systems of independent components, and so is given in terms of reliabilities. The following version is in terms of expectations and allows for failures of components to be associated.

**Definition 1-3:** DFM reliability can be calculated as follows:

\[
E[S(x)] = E[S(x'')] - E[S(x')]
\]
where \( E[S(x')] = \Pr[\text{no open cut}] \) is calculated using a binary-state system in which \( 1_i' = 1_i \cup f s_i \) and \( 0_i' = f o_i \), while \( E[S(x')] = \Pr[\text{short path}] \) uses \( 1_i' = f s_i \) and \( 0_i' = 1_i \cup f o_i \).

We can use this definition and its "state translations" to convert some other terms of DFM reliability into binary-state terms as follows.

**Lemma 1-3:**

(a) \( E[S(1_i,x)] = E[S(1_i''',x'')] - E[S(1_i'',x'')] \);  
(b) \( E[S(f s_i,x)] = E[S(1_i''',x'')] - E[S(1_i'',x'')] \);  
(c) \( E[S(f o_i,x)] = E[S(0_i''',x'')] - E[S(0_i'',x'')] \).

**Proof:**

These equivalencies are obtained by application of Definition 1-3 as follows.

(a) \( E[S(1_i,x)] = E[S(1_i''',x'')] - S(S(1_i',x'')) = E[S(1_i''',x'')] - S(0_i',x'') \)  
(b) \( E[S(f s_i,x)] = E[S(1_i''',x'')] - S(S(f s_i',x'')) = E[S(1_i''',x'')] - S(1_i',x'') \)  
(c) \( E[S(f o_i,x)] = E[S(0_i''',x'')] - S(S(f o_i',x'')) = E[S(0_i''',x'')] - S(0_i',x'') \)

**QED**

These equivalencies will be useful in subsequent calculations and derivations.

### 1.07 Importance of Components for DFM Models

Binary components have only two states, working and failed; thus when we speak of improving the reliability of a component, it is obvious that we are increasing the probability of it being in the working state while reducing the probability of it being in the failed state. In discussing DFM components, we must be more explicit in defining what we mean by "improving the reliability".

From now on when we speak of improving a component (making it more likely to be in the working state), we will specify whether it is improving with respect to open failures (hence decreasing the probability of being failed open, without changing the probability of being failed short), or the with respect to short failures (hence decreasing the probability of being failed short, without changing the probability of being failed open). We will ignore the possibility of simultaneously improving a component with respect to
both types of failures, though such a feature could be handled at the expense of more complicated notation and formulae.

We now develop importance measures for our DFM components. We begin by re-defining marginal reliability importance to handle the two types of failures.

**Definition 1-4:** For systems of DFM components:

(a) \[ \text{MRIO}_i = E \left[ S(1_i, x) - S(f_0, x) \right] \]
\[ = R(1_i, p) - R(f_0, p) \]
if components are independent;

(b) \[ \text{MRIS}_i = E \left[ S(1_i, x) - S(f_s, x) \right] \]
\[ = R(1_i, p) - R(f_s, p) \]
if components are independent.

MRIO is the MRI corresponding to reduction in the probability of an open failure, while MRIS is the MRI for reduction in the probability of a short failure. For the Figure 1-1 example, assume that there is an equal likelihood of either type of failure, so that \( q_0 = 0.1 \), and \( q_s = 0.1 \). Then MRIO \(_1 = 0.99 \) and MRIO \(_2 = 0.09 \), so MRIO \(_1 > \) MRIO \(_2 \) and MRIO \(_2 = \) MRIO \(_3 \); while MRIS \(_1 = 0.19 \) and MRIS \(_2 = 0.09 \), so MRIS \(_1 > \) MRIS \(_2 \) and MRIS \(_2 = \) MRIS \(_3 \). Here we see that although the probability of each component failure type (open or short) is the same, their effect on importance is quite different. For example, MRIO \(_1 \) is much higher than MRIS \(_1 \) because an open failure in component 1 always causes system failure, whereas a short failure often allows the system to keep functioning (depending on the other component states).

We can redefine structural importance in a similar fashion.

**Definition 1-5:**

(a) \[ \text{SIO}_i = 3^{- (n-1)} \sum [ S(1_i, x) - S(f_0, x) ] \]
where the summation is over \( x \) such that \( x_i = 1 \);

(b) \[ \text{SIS}_i = 3^{- (n-1)} \sum [ S(1_i, x) - S(f_s, x) ] \]
where the summation is over \( x \) such that \( x_i = 1 \).

For the example, SIO \(_1 = 0.888 \) and SIO \(_2 = 0.333 \), so SIO \(_1 > \) SIO \(_2 \) and SIO \(_2 = \) SIO \(_3 \); while SIS \(_1 = 0.555 \) and SIS \(_2 = 0.333 \), so SIS \(_1 > \) SIS \(_2 \) and SIS \(_2 = \) SIS \(_3 \).

As with SI and MRI, there can be a direct equivalency between SIO and MRIO, and between SIS and MRIS. This occurs when all components are independent and all
probabilities equal 1/3, i.e. when we have $p_i = q_{o_i} = q_{o_j} = 1/3$ for all $i$, then clearly the following relationships will hold:

(a) $\text{SIO}_i \geq \text{SIO}_j \iff \text{MRIO}_i \geq \text{MRIO}_j$;
(b) $\text{SIO}_i = \text{SIO}_j \iff \text{MRIO}_i = \text{MRIO}_j$;
(c) $\text{SIS}_i \geq \text{SIS}_j \iff \text{MRIS}_i \geq \text{MRIS}_j$;
(d) $\text{SIS}_i = \text{SIS}_j \iff \text{MRIS}_i = \text{MRIS}_j$.

We next construct our JRI equivalents, again in two variations.

**Definition 1-6:**

(a) $\text{JRI}_{ij} = E\left[ S(i,1, j, x) + S(f_{o_i}f_{o_j} x) - S(f_{o_i} 1, j, x) - S(1, j, f_{o_j} x) \right]$

$$= R(1, i, j, p) + R(f_{o_i} f_{o_j} p) - R(f_{o_i}, 1, j, p) - R(1, i, f_{o_j} p)$$

when components are independent;

(b) $\text{JRI}_{ij} = E\left[ S(1, i, j, x) + S(f_{s_i} f_{s_j} x) - S(f_{s_i}, 1, j, x) - S(1, j, f_{s_j} x) \right]$

$$= R(1, i, j, p) + R(f_{s_i} f_{s_j} p) - R(f_{s_i}, 1, j, p) - R(1, i, f_{s_j} p)$$

when components are independent.

For the system in our example, $\text{JRI}_{12} = +0.1$ and $\text{JRI}_{23} = -0.9$; $\text{JRI}_{12} = -0.9$ and $\text{JRI}_{23} = +0.1$.

Both MRI and JRI for binary components can also be expressed as derivatives of the reliability function when component failures are independent. This is also true for the DFM equivalents.

**Lemma 1-4:** When all components are independent, we have:

(a) $\text{MRIO}_i = \frac{\partial}{\partial p_i} R(p) = - \frac{\partial}{\partial q_{o_i}} R(p)$;
(b) $\text{MRIS}_i = \frac{\partial}{\partial p_i} R(p) = - \frac{\partial}{\partial q_{s_i}} R(p)$;
(c) $\text{JRI}_{ij} = \frac{\partial^2}{\partial p_i^2} R(p) = \frac{\partial^2}{\partial q_{o_i}^2} R(p)$;
(d) $\text{JRI}_{ij} = \frac{\partial^2}{\partial p_i^2} R(p) = \frac{\partial^2}{\partial q_{s_i}^2} R(p)$.

**Proof:**

Part (a). Decompose the reliability function (as described in e.g. [BP75]) and then make the substitution of $1-p_{i}-q_{s_{j}}$ in place of $q_{o_{i}}$. 
Now take the derivative with respect to $p_i$.
\[
\frac{\partial}{\partial p_i} R(p) = \frac{\partial}{\partial p_i} \left[ p_i R(1_i, p) + qo_i * R(fo_i, p) + qs_i * R(fs_i, p) \right]
\]
\[
= p_i * R(1_i, p) + (1-p_i-qo_i) * R(fo_i, p) + qs_i * R(fs_i, p)
\]

Part (b): This follows a parallel argument, using $1-p_i-qo_i$ in place of $qo_i$.
\[
R(p) = p_i * R(1_i, p) + qo_i * R(fo_i, p) + qs_i * R(fs_i, p)
\]
\[
\frac{\partial}{\partial p_i} R(p) = \frac{\partial}{\partial p_i} \left[ p_i R(1_i, p) + qo_i * R(fo_i, p) + qs_i * R(fs_i, p) \right]
\]
\[
= R(1_i, p) - R(fo_i, p)
\]

Part (c): Start with the derivative from part (a) and repeat the decomposition, substitution, and differentiation.
\[
\frac{\partial}{\partial p_i} R(p) = \frac{\partial}{\partial p_i} \left[ p_i * R(1_i, p) + (1-p_i-qo_i) * R(fo_i, p) + qs_i * R(fs_i, p) \right]
\]
\[
= p_i * R(1_i, p) + (1-p_i-qo_i) * R(fo_i, p) + qs_i * R(fs_i, p)
\]
\[
= \frac{\partial}{\partial p_i} \left[ p_j * R(1_j, p) + (1-p_j-qs_j) * R(fo_j, p) + qs_j * R(fs_j, p) \right]
\]
\[
= p_j * R(1_j, p) + (1-p_j-qs_j) * R(fo_j, p) + qs_j * R(fs_j, p)
\]
\[
= \frac{\partial}{\partial p_i} R(1_i, p) - R(fo_i, p)
\]

Part (c): Start with the derivative from part (a) and repeat the decomposition, substitution, and differentiation.
\[
\frac{\partial^2}{\partial p_i^2} R(p) = \frac{\partial^2}{\partial p_i^2} \left[ p_i * R(1_i, p) + (1-p_i-qo_i) * R(fo_i, p) + qs_i * R(fs_i, p) \right]
\]
\[
= R(1_i, p) + R(fo_i, p) - R(fo_i, p) - R(1_i, p)
\]
\[
= JRI \delta_i^2
\]
To get the second equality, follow the same procedure above but instead substitute $1-qo_i-qsi$ in place of $p_i$ and take derivatives with respect to $qo_i$.

Part (d): This follows a parallel argument to that in (c) above.

\[
\frac{\partial}{\partial p_i} R(p) = R(1_i,p) - R(fs_i,p) \\
= p_j R(1_i,1_j,p) + (1-p_j-qo_j) R(1_i,fs_j,p) + qo_j R(1_i,fo_j,p) \\
- p_j R(fs_i,1_j,p) - (1-p_j-qo_j) R(fs_i,fs_j,p) - qo_j R(fs_i,fo_j,p)
\]

\[
\frac{\partial^2}{\partial p_i^2} R(p) = 1*R(1_i,1_j,p) + (0-1-0)*R(1_i,fs_j,p) + 0*R(1_i,fo_j,p) \\
- 1*R(fs_i,1_j,p) - (0-1-0)*R(fs_i,fs_j,p) - 0*R(fs_i,fo_j,p) \\
= R(1_i,1_j,p) + R(fs_i,fs_j,p) - R(fs_i,1_j,p) - R(1_i,fs_j,p)
\]

To get the second equality, follow the same procedure above but instead substitute $1-qo_j-qsi$ in place of $p_i$ and take derivatives with respect to $qsi$.

These same results can be obtained by first converting DFM reliability into its equivalent binary component terms (via Lemma 1-2), and then taking the appropriate derivatives.

**QED**

It is easy to show that the definitions for MRI and JRI with DFM components have direct equivalents in terms of binary state components.

**Proposition 1-3:**

(a) $\text{MRIO}_i = \text{MRI}_i''$, i.e. $\text{MRI}_i$ calculated using $x''$;

(b) $\text{MRIS}_i = \text{MRI}_i'$, i.e. $\text{MRI}_i$ calculated using $x'$;

(c) $\text{JRI}_ij = \text{JRI}''_{ij}$, i.e. $\text{JRI}_i$ calculated using $x''$;

(d) $\text{JRI}_{ij} = -\text{JRI}'_{ij}$, i.e. $-\text{JRI}_i$ calculated using $x'$ (note the negative sign).

**Proof:**

These equivalencies are obtained by application of Lemma 1-3 term-by-term to Definitions 1-4 & 1-5.
(a) $\text{MRIO}_j = E[S(1_i, x)] - E[S(0_i, x)]$

$= E[S(1_i, x') - S(0_i, x')] - E[S(0_i, x') - S(0_i, x')]$

$= E[S(1_i, x') - S(0_i, x')] = \text{MRIO}_j$

(b) $\text{MRIS}_j = E[S(1_i, x)] - E[S(f_{ij}, x)]$

$= E[S(1_i, x') - S(0_i, x')] - E[S(1_i, x') - S(1_i, x')]$

$= E[-S(0_i, x') + S(1_i, x')] = \text{MRIS}_j$

(c) $\text{JRIO}_{ij} = E[S(1_i, 1_j, x) - S(f_{oi}, 1_j, x) - S(1_i, f_{oj}, x) + S(f_{oi}, f_{oj}, x)]$

$= E[S(1_i, 1_j, x') - S(0_i, 0_j, x')] - E[S(0_i, 1_j, x') - S(0_i, 0_j, x')]$

$- E[S(1_i, 0_j, x') - S(0_i, 0_j, x')] + E[S(0_i, 0_j, x') - S(0_i, 0_j, x')]$

$= E[S(1_i, 1_j, x')] - E[S(0_i, 1_j, x')] - E[S(1_i, 0_j, x')] + E[S(0_i, 1_j, x')]$

$= \text{JRIO}_{ij}$

(d) $\text{JRIS}_{ij} = E[S(1_i, 1_j, x) - S(f_{si}, 1_j, x) - S(1_i, f_{sj}, x) + S(f_{si}, f_{sj}, x)]$

$= E[S(1_i, 1_j, x') - S(0_i, 0_j, x')] - E[S(1_i, 1_j, x') - S(1_i, 0_j, x')]$

$- E[S(1_i, 1_j, x') - S(0_i, 1_j, x')] + E[S(1_i, 1_j, x') - S(1_i, 1_j, x')]$

$= -E[S(0_i, 0_j, x')] + E[S(1_i, 0_j, x')] + E[S(0_i, 1_j, x')] - E[S(1_i, 1_j, x')]$

$= -\text{JRIS}_{ij}$

**QED**

We offer below a DFM definition of LI in which we follow the idea behind the original version: to be ordered, one component must be at least as important as another component regardless of which state both are assumed to be in.

**Definition 1-7:** $\text{LID}_i \geq \text{LID}_j$ (read, "component $i$ is at least as important in terms of LID as component $j$") if all of the following are true for any assignment of probabilities such that the marginal probabilities for each component are all identical:
(a) \( E[S(f_{ij}, x)] \geq E[S(f_{ij}, x)] \);
(b) \( E[S(f_j, x)] \geq E[S(f_j, x)] \);
(c) \( E[S(1_j, x)] \leq E[S(1_j, x)] \).

If both \( LID_i \geq LID_j \) and \( LID_i \leq LID_j \), then \( LID_i \approx LID_j \).

For the example system \( LID_2 \approx LID_3 \), but \( LID \) cannot give an ordering for components 1 and 2.

Defined in this way, \( LID \) has similar implications to those that \( LI \) had in binary models.

**Proposition 1-4:** \( LID \) implies the following about \( SIO, SIS, MRIO \) and \( MRIS \):

(a) \( LID_i \geq LID_j \Rightarrow SIO_i \geq SIO_j \) and \( SIS_i \geq SIS_j \);
(b) \( LID_i \approx LID_j \Rightarrow SIO_i = SIO_j \) and \( SIS_i = SIS_j \);
(c) \( LID_i \geq LID_j \Rightarrow MRIO_i \geq MRIO_j \) and \( MRIS_i \geq MRIS_j \) for identical components;
(d) \( LID_i \approx LID_j \Rightarrow MRIO_i = MRIO_j \) and \( MRIS_i = MRIS_j \) for identical components.

**Proof:**

First consider part (c). \( LID_i \geq LID_j \) by definition means that both \( E[S(f_{ij}, x)] \geq E[S(1_i, x)] \) and \( E[S(f_{ij}, x)] \leq E[S(1_i, x)] \). From this it follows that

\[
E[S(1_i, x)] + E[S(f_{ij}, x)] \geq E[S(1_i, x)] + E[S(f_{ij}, x)]
\]
\[
E[S(1_i, x)] - E[S(f_{ij}, x)] \geq E[S(1_i, x)] - E[S(f_{ij}, x)]
\]

\( MRIO_i \geq MRIO_j \)

Similarly, \( E[S(f_{ij}, x)] \geq E[S(f_{ij}, x)] \) and \( E[S(1_j, x)] \leq E[S(1_j, x)] \), and so

\[
E[S(1_j, x)] + E[S(f_{ij}, x)] \geq E[S(1_j, x)] + E[S(f_{ij}, x)]
\]
\[
E[S(1_j, x)] - E[S(f_{ij}, x)] \geq E[S(1_j, x)] - E[S(f_{ij}, x)]
\]

\( MRIS_i \geq MRIS_j \).

For part (d), \( LID_i \approx LID_j \) means that both: \( LID_i \geq LID_j \), hence \( MRIO_i \geq MRIO_j \); and \( LID_i \leq LID_j \), hence \( MRIO_i \leq MRIO_j \). If these inequalities for \( MRIO \) are both true, then we must have \( MRIO_i = MRIO_j \). The same logic shows that we must have \( MRIS_i = MRIS_j \).
By setting \( p = 1/3 \), \( qs = 1/3 \), and \( qo = 1/3 \), we have \( MRIO = SIO \) and \( MRIS = SIS \); then (a) follows from (c) and (b) follows from (d).

**QED**

Since the original PCI was defined strictly in terms of the network topology, in principle it can be used as is. Unfortunately, it seems that it no longer is as sensible a measure as before; for example, a short minimum path is beneficial when one is only worried about conventional (open) failures, but it is a liability when dealing with short failures. We do not define a new version herein, as there is no obvious way to adapt its rationale to DFM systems. We still have \( PCI_1 \geq PCI_2 \) and \( PCI_2 \approx PCI_3 \) for the example system.

In Theorems 1-1 and 1-2 we derived some structural properties for the JRI of binary components. Below we present the equivalent results for our DFM model.

**Proposition 1-5:** The following three conditions are equivalent:

(a) there is no minimum path set containing both \( i \) and \( j \);

(b) \( JRIO_{ij} \leq 0 \) for all possible assignments of probabilities;

(c) \( JRIS_{ij} \geq 0 \) for all possible assignments of probabilities.

**Proof:**

Proposition 1-3 showed that \( JRIO = JRI" \) and \( -JRIS = JRI' \). Substitute these into part (b) of Theorem 1-1 as follows.

\[
JRIO_{ij} \leq 0 \iff JRI_i" \leq 0 \\
JRIS_{ij} \geq 0 \iff -JRI_i \geq 0 \iff JRI_i' \leq 0
\]

**QED**

**Proposition 1-6:** The following three conditions are equivalent:

(a) there is no minimum cut set containing both \( i \) and \( j \);

(b) \( JRI_{ij} \geq 0 \) for all possible assignments of probabilities;

(c) \( JRIS_{ij} \leq 0 \) for all possible assignments of probabilities.
**Proof:**

Proposition 1-3 showed that JRIO = JRI" and -JRIS = JRI'. Substitute these into part (b) of Theorem 1-2 as follows.

\[
\begin{align*}
JRIO_{ij} \geq 0 &\iff JRI_{ij}'' \geq 0 \\
JRIS_{ij} \leq 0 &\iff -JRI_{ij}' \leq 0 \iff JRI_{ij}' \geq 0
\end{align*}
\]

**QED**

These two results make intuitive sense. JRIO behaves just like JRI does because these both deal with the same kind of failures (i.e. those due to open cuts); for example, if we had two components in parallel (hence no common minimum path) we would expect that reducing the risk of an open failure in one would tend to make the other less important with respect to causing a system open failure. On the other hand, JRIS behaves in the opposite fashion: in the same situation, making one component less likely to fail short would tend to make the other one more critical with respect to system short failures. This is the reason for the negative sign in part (d) of Proposition 1-3. Our result in Proposition 1-2 concerning the strictness of the inequalities clearly continues to hold for these DFM equivalents.

Looking back we can see that all of the DFM importance measures (except LID, which was unable to give a ranking) ranked component 1 as being more important than components 2 and 3 in the example system in Figure 1-1; this is due in part to the simple structure of the system and our choice of probabilities. Note that the strength of this ranking varied considerably depending on the measure used. Consider for example the MRI measures: we see that component 1 is much more important than component 2 regarding open failures (MRIO\(_1\) = 0.99 and MRIO\(_2\) = 0.09), but is only slightly more important regarding short failures (MRIS\(_1\) = 0.19 and MRIS\(_2\) = 0.09). This means that if we wanted to improve the reliability of a component, we would get much better gains from reducing the chance of an open failure in component 1 than from either reducing the
chance of an short failure in component 1 or reducing the chance of either failure in component 2.

1.08 Future Research

This chapter has dealt with components which are either binary-state or DFM three-state. An obvious goal would be to extend these notions of component importance to systems of general multi-state components, or perhaps even continuous-state components. The problem in doing so is that there seem to be many differing ways to define such a system's reliability (or performability, see [SS90]), so any importance measures and properties may depend heavily on the definition used. Technically, PCI can be used with components with any number of states as it only considers network topology; however, the usefulness of the PCI ordering may depend on the nature of the system (e.g. PCI considers a long s-t path to be undesirable, but as we saw with the failed short state of DFM components this is not always appropriate).

We should note that we have completely ignored one important issue, which is the actual calculation of the measures we have discussed. As other authors have discussed (see e.g. [Co87]) some of these measures involve calculations which are NP-Complete (e.g. the work required to calculate SI grows exponentially in the number of components), and many require information on component reliabilities and covariances that may not be readily available in practice. For example, MRI calculations require the knowledge of every component's reliability as well as the covariance between every pair of components, whereas in reality component reliabilities are only estimates and most covariances are completely unknown. Addressing these problems is beyond the scope of our work herein, and so we leave them for future research to deal with.

1.09 Concluding Remarks

In this chapter, we reviewed several existing measures of reliability importance for binary components and discussed how they are related. We then showed that structural results previously derived in [HL93] for JRI of independent components continue to hold
for correlated components. We proposed extensions to all of these measures to cover DFM components, and showed that their original properties largely continue to hold.
Chapter 2

Maintenance and Inventory for a Single Machine

2.01 Chapter Abstract

We consider a system containing a single machine subject to random failure, for which we wish to determine a maintenance policy and a spare ordering policy. We examine the solvability and desirability of jointly optimizing these two traditionally separate policies. We discuss why the general form of the problem is largely intractable and then proceed to examine several special cases which are easier to analyze. Special attention is paid to systems which are only allowed to contain a single spare; we show that for these systems our problem has some convexity properties which make it amenable to cost minimization. We subsequently consider systems which allow more than one spare, but make use of other restrictions or assumptions to somewhat simplify analysis.

2.02 Introduction

Consider a system containing one machine subject to random failure, for which we want to determine a maintenance (age replacement) policy and an inventory (spare ordering) policy that will permit the system to operate at minimum cost. The traditional approach to finding a maintenance policy is to assume that spares are always available as needed, and then select a replacement policy which minimizes the expected maintenance costs (i.e. ignoring any costs due to inventory). Similarly, the traditional approach to inventory policy is to make some assumption about the distribution of demand, and then select an ordering policy which minimizes the expected inventory costs while ignoring the maintenance costs. The selection of these two policies in sequence ignores the effect that each can have on the other. In this paper we consider the problem of joint optimization of replacement and ordering policies, paying particular attention to solvability (is it tractable?) and desirability (is it worthwhile?).

This work is motivated in part by the author's previous work experience in aircraft maintenance. At his place of work, maintenance decisions and inventory decisions were
made by two distinct groups of people, and a significant amount of the local manager's
time was spent resolving problems between these two groups. It was also a feature of the
organization that on occasion there were arbitrary limits placed on the maximum level of
inventory that could be stocked. Sometimes this was due to a physical space limit; e.g.,
when deploying to a remote location, only one of any large spare would be taken. More
generally, the limits arose from policy decisions made by a higher level headquarters; e.g.,
due to limits in the purchasing budget or a production shortfall at the manufacturer, no site
was permitted to hold more than one of a certain high value spare, and an additional spare
could only be obtained in exchange for a failed part. Thus a single spare system could
arise due to a spare rationing policy that was beyond the local manager's control.

One of the difficulties in jointly considering the maintenance and inventory policies
for a machine is caused by the interaction between maintenance demand and inventory
supply. In traditional maintenance analysis a spare is assumed to always be available; thus
the performance of the inventory policy has no effect on when maintenance may be carried
out. Similarly, in traditional inventory analysis the demand is assumed to arise from an
exogenous process; this process continues to make demands regardless of the past success
(or lack thereof) of the inventory policy. In a reality, however, maintenance decisions will
necessarily be influenced by whether or not a spare is currently in stock; similarly,
whenever there is an inventory stock-out that prevents us from running the machine, the
demand process will stop until a spare is delivered. Thus we must explicitly consider the
effect each sub-system has on the other if we are to develop optimal policies.

There is a large body of literature dealing with maintenance issues in general; refer
to [VF89] for a survey of papers dealing with the maintenance of a single machine. The
earliest work that considered the joint optimization of maintenance and inventory policies
dealt with Markovian systems; see for example Ross [Ro69], Derman & Lieberman
[DL67], and Allan & D'Esopo [AD68]. Kao in [Ka73] works with semi-markov
processes, but does not consider inventory issues. Falkner [Fa68] worked with a non-
Markovian (i.e. deteriorating) machine but handled inventory quite differently from our work: he used a periodic inventory policy with immediate replenishment at the beginning of each period. His approach allowed a wide variety of costs and structural features. He showed that the optimal replacement time was a function of the stock on hand and the time until next replenishment, and that an optimal order quantity could be calculated by an iterative procedure. Falkner's work in [Fa69] which dealt with inventory only, and in [Fa70] which dealt with deterministic machines, was also along these lines. Park & Park [PP85] similarly dealt with systems using periodic review inventory policies. There are also a number of papers which have studied variations on (s-1,s) continuous review inventory policies that are partially relevant to our problem: see for example Feeney & Sherbrooke [FS66], Posner & Yansouni [PY72], Smith [Sm77], and Schmidt & Nahmias [SN85]. We will later in this paper make use of some results by Atkins & Katircioglu [AK95] and by Schultz [Sc89], who studied delayed (s-1,s) inventory policies.

There have been quite a few papers dealing with single spare models, one of the earliest of which was [NO74] by Nakagawa & Osaki. They considered a system like ours but without any breakage cost; consequently, age replacement as such was unnecessary. Subsequent papers, e.g. [KO78a], [KO78b], [KO78c], [KO78d], [KO79], [KO88], [KH81], [OKY81] [Os77], [PP86a], [PP86b], [SL92], [Sr92], [SS89], [TO78a], and [TO78b], have added other costs and/or maintenance actions to the model but have continued to neglect a breakage cost, and so their results could not be directly compared to classic age replacement theory either.

We might classify these papers based on structural concepts. The earlier papers tended to assume deterministic lead times, while later work allowed for random lead times and minimal repairs. The following allowed for random lead times: [KH81], [PP86a], [PP86b], [Sr92], and [SS89]. A distinction between order types (scheduled versus unscheduled) was made in [TO78b], [KO78a], [KO78b], [KO78c], [KO78d], [KO79], [KO88], [KH81], [Os77], [SL92], [Sr92], [SS89], and [TO78a]. Minimal repairs were
included in some form in [SL92], [Sr92], [KO88], and [PP86a]. Many papers required the use of a replacement upon receipt maintenance policy, i.e. always install a spare as soon as it arrives: this assumption was made in [KO78a], [KO78b], [KO78c], [KO78d], [KO79], [KH81], [Os77], [SL92], [Sr92], [SS89], and [TO78a]. We note that most of these papers quite naturally have tended to emphasize the case where the machine has an increasing failure rate; consequently, they have neglected to consider the option of returning an unused spare component to its origin if the machine has a decreasing failure rate.

Mine & Kawai [MK77] examined a similar system which did have a breakage cost but no holding or explicit shortage charges; their order lead time was random and their objective used discounted total cost rather than average cost rate. They showed that under reasonable conditions a unique optimal ordering time and replacement time existed, and they derived some properties of these times. [OKY81] also used discounted total cost as their objective. [SLT92] considered a variation in which policies were based upon the cumulative number of failures since installation rather than age. [NO78] considered a variation in which the main decision was whether to repair or replace, using the estimated cost or estimated time to repair as the decision criteria.

The most general form of the problem that we might consider in this chapter would allow the lead times for receipt of spares to be random, with a distribution that might depend on whether the order was scheduled or not, and would allow inventory to contain as many spares in stock or on order as desired. We might hope to develop an equation using renewal principles that looks something like the following, which we could then analyze to find jointly optimal maintenance and inventory policies.

\[
E \text{[cost rate of policy]} = \frac{E \text{[maintenance costs per cycle]}}{E \text{[length of cycle]}} + \frac{E \text{[inventory costs per cycle]}}{E \text{[length of cycle]}}
\]
This equation determines the long run expected cost rate of operating the system by dividing the expected cost per renewal cycle by the expected length of such a cycle. This requires that we be able to determine clear renewal points and easily describe the system's behaviour between these points. In the general problem such renewal points can be specified, but the behaviour of the system between these points is complicated to the point of being intractable for exact analysis, in part because of the previously mentioned interaction between maintenance demand and inventory supply. In this chapter our approach is to make some restrictive assumptions that allow us to define and solve simpler versions of the problem, particularly for the case where the system contains only one spare at a time. Typically, such restrictions facilitate analysis by partially removing the maintenance-inventory interaction, or by restricting the choice of allowable policies.

We begin this chapter by more precisely describing our modeling assumptions and notation. We then consider systems which are allowed to contain at most one spare, first requiring that the order lead time be deterministic and then allowing it to be random; our most detailed analytical results are obtained in these sections. Next to be examined are systems which allow more than one spare, but make use of other restrictions or assumptions to somewhat simplify analysis. In each case, we proceed by defining a renewal system, specify an equation which describes its expected cost rate, and then derive some of the properties of the equation. We conclude with a suggested approach to tackling the general problem in actual practice. Proofs for our results are contained in Appendix B.

2.03 Model Description & Notation

We examine a system which contains one machine subject to random failure. We can shut down the installed machine at any time, and we can then replace it provided that we have a new spare machine in stock. Between placement of an order and receipt of a spare there is a random lead time with distribution $G(t)$, density $g(t)$, and mean $L > 0$. The replacement action itself is assumed to be perfect (i.e. it returns the machine to good-as-
new status) and requires a random time with mean $R$. The stochastic machine lifetimes are assumed to follow a distribution function $F(t)$ which has finite mean $\mu$, density $f(t)$, and hazard rate $z(t)$. We say that $F(t)$ is IFR (increasing failure rate) when $z(t)$ is non-decreasing, and strictly IFR when $z(t)$ is strictly increasing; similarly DFR or strictly DFR indicates $z(t)$ is non-increasing or strictly decreasing, respectively. In most of this chapter it is assumed that the machine is IFR; we will indicate whenever we depart from this assumption. We use $t_r$ to represent the scheduled replacement time of the component, and $t_o$ to denote the scheduled time at which we order a spare.

Throughout this chapter we assume that the failure distribution $F(t)$ is known, whereas in practice it is typically unknown and must be estimated. This estimation process is a non-trivial problem in itself, and can become quite complicated if it is performed while trying to operate the system efficiently, i.e. if we are learning as we go. This is an interesting and important problem but is beyond the scope of the work herein.

Our system incurs a cost $cv(t)$ whenever a part is replaced, where $c$ is a cost coefficient and $v(t)$ is a convex function of the age of the machine, i.e. the cost of a replacement depends on the machine age; this is the cost of a preventive replacement, and includes the cost of the part itself, labour charges for the actual installation, and the administrative cost of placing an order (which is assumed to be negligible, so that ordering of individual spares is appropriate). The simplest case is where $v(t) = 1$ for all $t$, so that the replacement cost is a constant; allowing $v(t)$ to be convex permits consideration of more general cost expressions, e.g. a fixed parts cost plus a age-varying labour charge.

The system incurs a breakage or failure cost $b$ if the replacement is done after failure; this is the additional cost of a corrective replacement, and can include charges for extra labour, repair of damage to the rest of the system, etc. If the system is unavailable, i.e. while awaiting delivery of a needed spare, a shortage penalty cost at rate $k$ per unit
time is incurred; this can represent missed profits, the cost of renting a temporary replacement, etc. A cost rate of $h$ per unit time is incurred while holding a spare in inventory; this represents the opportunity cost of capital, depreciation of the spare, etc.

An operating cost is also incurred at rate $aq(t)$ whenever the system is operational, where $a$ is a cost coefficient and $q(t)$ is a non-decreasing function of machine age, i.e. the machine gets more expensive to operate as it ages. This term is used to represent direct costs of operation such as for fuel use, etc., as well as many of the other costs that previous authors have included in their models. One such cost is that of declining revenues due to deteriorating production quality or yield, as discussed in [MB95]. Another is the decreasing expected salvage value of the installed part, as in e.g. [KH81]. Probably the most important is the increasing expected cost of minimal repairs (which return the machine to bad-as-old status) as in [KO88]. There are several different ways to model minimal repairs; see [Be93] for a thorough discussion of the topic. Our work takes a fairly general approach, in that we assume that there are two kinds of failure, minor and major. A minor failure is always rectified by a minimal repair, which is instantaneous and does not incur the breakage penalty $b$. A major failure does incur the breakage penalty and can only be rectified by replacing the failed machine with a new one. Thus the term $aq(t)$ in our model is an aggregate of whichever of these other costs may be relevant to the system being modeled.

We will introduce other cost factors and decision variables as required later in this chapter. All costs are assumed to be non-negative unless otherwise indicated.

### 2.04 Single Spare with Basic Structure

In this section our system is constrained to allow only one spare to be in stock or on order at any time, and by requiring that the order lead time for both scheduled and unscheduled orders be deterministic and equal to $L$; we naturally require $0 \leq t_o \leq t_r - L \leq \infty$. A single spare is obviously a severe restriction on inventory policies, though it may be approximately true for real systems due to storage space limitations, and in some cases it
none the less leads to an optimal policy. We further restrict our model structure by assuming \( v(t) = 1 \) (so replacement cost is a constant value \( c \)), \( q(t) = 0 \) (no operating cost), and \( R = 0 \) (replacement is instantaneous). This restricted system gives us the most tractable model that we consider and hence permits our most detailed results. Previous papers have analyzed many variations of this restricted system, but we believe our choice of cost structure is less well examined and generally more interesting. We cover both the two major costs generally considered in the maintenance literature (replacement and breakage) as well as the two major costs widely considered in the inventory literature (holding and shortage). Our results therefore can be directly compared to those of classic models of age replacement (such as in chapter 3 of [JMR67]) which largely ignore inventory issues.

Our performance objective is to minimize the system's expected cost per unit time (expected cost rate). Since replacement times are regeneration points of the system, this cost rate can be obtained by looking at one cycle of the system; that is, from the time of one replacement until the time of the next. The expected cost rate thus equals the expected sum of costs incurred in one cycle divided by the expected length of a cycle.

During our discussion, it may be helpful for the reader to refer to the time-lines shown in Figure 2-1. These show the possible orders of events that can occur in the system during each renewal cycle.

**Figure 2-1: Timing of Events**

(a) Machine survives until scheduled replacement

\[
\begin{array}{c}
0 \quad \text{to} \quad \text{to} + L \quad \text{tr}
\end{array}
\]

(b) Machine fails after spare arrival

\[
\begin{array}{c}
0 \quad \text{to} \quad \text{to} + L \quad \text{X} \quad \text{tr}
\end{array}
\]
(c) Machine fails after spare is ordered

\[ \begin{array}{cccc}
0 & \text{to} & X & \text{to} + L & \text{tr} \\
\end{array} \]

(d) Machine fails before spare is scheduled for ordering

\[ \begin{array}{cccc}
0 & X & \text{to} & \text{to} + L & \text{tr} \\
\end{array} \]

where

\[
\begin{align*}
X & \quad \text{machine failure} \\
\text{machine operation} & \\
\text{spare on order} & \\
\text{shortage period} & \\
\text{spare in stock} &
\end{align*}
\]

Let us begin by looking at the traditional methods of making the replacement and ordering decisions sequentially. The traditional maintenance approach is concerned with the replacement cost \( c \) and the breakage cost \( b \); the idea is to balance the cost of using too many parts against the increasing expected costs due to operation and breakage. Spares are assumed to be always available. What we must decide is when to replace the part. A replacement is made on failure or possibly as a preventive measure before failure at some planned time \( t_r \). The form of our optimal policy is an age replacement policy defined by this single number \( t_r \).

In the derivations which follow we make use of the relation that

\[
\mu = \int_0^\infty tf(t)dt = \int_0^\infty F(t)dt \tag{2-2}
\]

The expected cost rate for this policy sums the expected cost of parts plus the expected cost of breakage, both per cycle;

\[
c + bF(t_r) \tag{2-3}
\]
and divides by the expected cycle length

\[
\int_{t_r}^{\infty} x f(x) \, dx + \int_{t_r}^{\infty} t_r f(x) \, dx = \int_{0}^{t_r} F(x) \, dx
\]

Combining these gives an expression for the expected cost rate which we call the Replacement Cost Function \( RCF \).

\[
RCF(t_r) = \frac{c + bF(t_r)}{\int_{0}^{t_r} F(x) \, dx}
\]

Since \( F(t) \) is IFR, \( RCF \) is a pseudo-convex (and unimodal) function of \( t_r \) with a unique minimum, as was shown in chapter 4 of [JMR67] (and as we show more generally in Theorem 2-1 which follows). It is not necessarily convex. The principle benefit of pseudo-convexity in this context is that any local minimum of the function is also a global minimum; furthermore, if the pseudo-convexity is strict this implies that the minimum is a unique point (otherwise there may be an interval over which the function has the same minimum value). Refer to [ADSZ88] or to [BSS93] chapters 3 and 4 for more details on the subject of convexity.

The optimal \( t_r \) will be one of \( t_r' \) or \( \infty \) as follows, where \( z(t) \) is the hazard rate:

(a) \( 0 < t_r' \leq \infty \) if \( bz(t_r') - RCF(t_r') = 0 \);

(b) \( t_r = \infty \) if \( bz(\infty) - RCF(\infty) \geq 0 \).

This equation is saying that the optimal scheduled replacement age is that at which the marginal expected cost rate of operating equals the average expected cost rate. In the event that no such \( t_r' \) exists, i.e. if \( bz(\infty) < RCF(\infty) \), then this implies that it is optimal to replace only at failure, so set \( t_r = \infty \). To actually solve for \( t_r \), we can use (a) and (b) above in conjunction with some form of line search procedure (see [BSS93] chapter 7 for examples).
Now consider the traditional inventory approach, which is concerned with balancing the shortage cost rate $k$ incurred while a demand is unmet against the cost rate $h$ of holding a spare in stock. The decision to be made is when to order a spare so as to minimize these costs. In our system, we cannot order a spare if there is already one in stock or on order. The order instant $t_o$ will therefore lie between the time the last spare was installed (time zero), and a time $L$ before the next definitely will be needed (at $t_r$); thus $0 \leq t_o \leq t_r - L$.

In general for this model, the expected shortage cost per cycle is

$$
k \left[ \int_0^{t_o} L f(x)dx + \int_{t_o}^{t_o + L} (t_o + L - x) f(x)dx \right] = k \int_{t_o}^{t_o + L} F(x)dx$$

and the expected holding cost per cycle is

$$
h \left[ \int_{t_o}^{t_o + L} (x - t_o - L) f(x)dx + (t_r - t_o - L) F(t_r) \right] = h \int_{t_o}^{t_o + L} \overline{F}(x)dx$$

The expected cycle length is not quite the same as for RCF due to the possibility of a stock-out prolonging the cycle. It is equal to the sum of the expected part life plus the sum of the expected shortage time.

$$
\int_0^{t_o} (x + L) f(x)dx + \int_{t_o}^{t_o + L} (t_o + L) f(x)dx + \int_{t_o}^{t_o + L} x f(x)dx + \int_{t_o}^{t_r} f(x)dx + \int_{t_o}^{t_o + L} F(x)dx
\quad = \int_0^{t_o} \overline{F}(x)dx + \int_{t_o}^{t_o + L} \overline{F}(x)dx
$$

Combining these terms yields the Ordering Cost Function $OCF$.

$$
OCF(t_o) = \frac{k \int_{t_o}^{t_o + L} F(x)dx + h \int_{t_o}^{t_r} \overline{F}(x)dx}{\int_{t_o}^{t_o + L} F(x)dx + \int_0^{t_o} \overline{F}(x)dx}
$$
Since $F(t)$ is IFR, $OCF$ is a pseudo-convex function of $t_o$ and so has a single minimum (see Theorem 2-2 below for details).

The $OCF$ equation assumes that the inventory manager knows and makes use of all relevant information about the distribution of demand, but typically this is not the case. We would commonly expect the manager to make a simplifying assumption about the form of the demand distribution, while taking account of whatever data was available about historical demand rates. Probably the simplest assumption is to treat demand as a Poisson process, with mean time between demands equal to $\mu$. This ignores the effect of preventive replacements and simplifies the ordering decision so that orders are only placed at demand epochs, i.e. choose $t_o$ to be either 0 or $t_r - L$ (whichever is least costly). This also simplifies implementation because the calculation only requires the average component lifetime without preventive replacements. Other approximations are possible; for example, one could assume a Poisson demand distribution with a mean inter-demand time $\mu^* < \mu$ which reflects historical demand from an existing maintenance operation. We will return to this issue later in the chapter.

To illustrate such a simplified case we offer what we call the Simplified Ordering Factor $SOF$, which chooses the ordering policy in complete disregard of maintenance demand (beyond using $\mu$).

$$SOF = kL - h\mu$$

The first term on the right-hand side represents the cost of a shortage each cycle, while the second is the cost of holding a spare throughout an average cycle. If $SOF$ is negative, one should order on demand; if positive, one should order immediately after replacement; or if zero, then either option is equally good.

Now we consider the joint optimization of maintenance and inventory, for which an operating policy is described by the pair $(t_o, t_r)$. If the machine fails before $t_r$, it is replaced immediately if a spare is available, or else as soon as a spare arrives; otherwise it is replaced at $t_r$. If the machine fails before $t_o$, an order for a spare is placed immediately;
otherwise an order is placed at \( t_o \). We want to find a feasible \( t_o, t_r \) pair that minimizes the expected cost rate for the entire system.

The expected cost function for the joint problem is built using (2-9) for the expected cycle length, (2-3) for the expected replacement and breakage costs per cycle, (2-7) for the expected shortage cost per cycle, and (2-8) for the expected holding cost per cycle. Combining these terms gives the expected operating cost per unit time, which we call the Joint Cost Function \( JCF \).

\[
JCF(t_o, t_r) = c + k \int_{t_o}^{t_o+L} F(x) dx + h \int_{t_r}^{t_o+L} \bar{F}(x) dx + bF(t_r)
\]

This function has some useful properties.

**Theorem 2-1:** \( JCF \) is unimodal and pseudo-convex in \( t_r \); hence for a given \( t_o \), \( JCF(t_o, t_r) \) has no interior \((t_o+L < t_r < \infty)\) maximum with respect to \( t_r \) and at most one interior minimum at \( t_r' \), and the optimal \( t_r \) will be one of \( t_o+L, \infty, \) or \( t_r' \) as follows:

(a) \( t_r = t_o+L \) if \( h + bz(t_o+L) - JCF(t_o, t_o+L) \geq 0 \);
(b) \( t_o+L \leq t_r' \leq \infty \) if \( h + bz(t_r') - JCF(t_o, t_r') = 0 \);
(c) \( t_r = \infty \) if \( h + bz(\infty) - JCF(t_o, \infty) \geq 0 \).

If there is no \( t_r' \), then \( JCF \) is monotone in \( t_r \). Case (a) indicates that the installed machine should be replaced immediately upon receipt of a spare (at \( t_o+L \)), case (c) implies that the machine should be replaced only at failure. Note that case (b) is simply saying that optimality occurs when the marginal expected cost rate of running equals the average expected cost rate overall.

**Theorem 2-2:** \( JCF \) is unimodal and pseudo-convex in \( t_o \); hence for a given \( t_r \), \( JCF(t_o, t_r) \) has no interior \((0 < t_o < t_r - L)\) maximum and at most one interior minimum point \( t_o' \), and so the optimal \( t_o \) will be one of \( 0, t_r - L, \) or \( t_o' \) as follows:
(a) $t_o = 0$ if $k - h \left[ \frac{\bar{F}(L)}{F(L)} \right] - JCF(0, t_r) \geq 0$

(b) $0 \leq t_o' \leq t_r - L$ if $k - h \left[ \frac{\bar{F}(t_o' + L)}{F(t_o' + L) - F(t_o')} \right] - JCF(t_o', t_r) = 0$

(c) $t_o = t_r - L$ if $k - h \left[ \frac{\bar{F}(t_r)}{F(t_r) - F(t_r - L)} \right] - JCF(t_r - L, t_r) \leq 0$

If there is no $t_o'$, then $JCF$ is monotone in $t_o$. Case (a) indicates that a spare should be ordered as soon as possible (at time 0), while case (c) implies that one should order only on demand (at $t_r - L$).

Proposition 2-1: A change in cost coefficients will tend to change the optimal policy values as follows:

(a) for any given $t_o$, the optimal $t_r$ is non-decreasing in the shortage cost rate $k$, and non-increasing in the holding cost rate $h$, and the breakage cost $b$;

(b) for any given $t_r$, the optimal $t_o$ is non-decreasing in $c$, $b$, and $h$, and non-increasing in $k$.

Part (a) implies that a previously optimal replacement policy of "replace at failure" remains unchanged if $k$ or $c$ increases, or if $b$ or $h$ decrease. Conversely, a previously optimal policy of "replace on receipt" remains unchanged if $k$ or $c$ decreases, or if $b$ or $h$ increase.

Part (b) implies that a previously optimal ordering policy of "just in time" remains unchanged if $c$ or $b$ or $h$ increase, or if $k$ decreases. Conversely, a previously optimal policy of "always stock" remains unchanged if $c$ or $b$ or $h$ decrease, or if $k$ increases.

For $t_o$ to be increasing in the maintenance costs ($b$ and $c$) is not surprising; a later order lengthens the expected cycle length, hence allowing the amortization of these costs over a longer period of time. The effect of the inventory costs ($k$ and $h$) on $t_r$ is readily explained by examining the derivatives of the cost function, but does seem somewhat counter-intuitive from the perspective of a maintenance manager: if the penalty for not
having a spare at time of failure is high, the machine is replaced less often, which would seem to risk more failures; if holding costs are high, it is replaced more often, which would seem to require more inventory be carried. The rationale for this behaviour becomes clear when we take the inventory manager's perspective: if the shortage cost is high, we wait longer to replace so that we have more time to get a spare in stock; if the holding cost is high, we replace sooner so as to get the spare off our shelf.

The jointly optimal pair \((t_o, t_r)\) can be calculated for a given problem instance using non-linear programming (NLP) methods, which typically search for Karush-Kuhn-Tucker (KKT) points. The NLP problem statement is to minimize \(JCF(t_o, t_r)\) subject to \(0 \leq t_o \leq t_r - L\). Figure 2-2 shows the feasible region for the NLP problem in which the four possible types of KKT point locations are shown.

**Figure 2-2:** Possible locations of KKT solution points.

![Diagram showing possible KKT solution points](image)

We know that the KKT conditions are necessary for a point to be a minimum (see chapter 4 of [BSS93]), but they are not sufficient for global optimality in all problem instances because although \(JCF\) is pseudo-convex in each variable individually it is not in general pseudo-convex in both variables simultaneously, as the following example shows.
Jointly Pseudo-Convex Counter-Example: Let \( F(t) = 1 - e^{t^3}, L = 0.6, c = 900, b = 1000, k = 1000, h = 10. \) Then there are two KKT points on the diagonal boundary, one of which is a minimum while the other is a saddle point.

In the next theorem we shall give a sufficient condition for global optimality; but first, the following definition is provided to aid in discussion.

**Definition 2-1:** We say that a feasible policy is *viable* if it makes economic sense to operate the system under that policy. If the policy is not viable, then we would be better off not operating the system at all; it would be cheaper to constantly incur the shortage cost.

For the problem at hand, a viable solution \((t_o, t_r)\) is one for which \( JCF(t_o, t_r) < k \). The next theorem makes use of this fact to characterize a globally optimal solution.

**Theorem 2-3:** If \((t_o, t_r)\) is a KKT point of \( JCF \) and is viable, then \((t_o, t_r)\) is a global minimum. In particular, any KKT point with \( t_o + L < t_r \) is always viable and hence always a global minimum.

Refer again to Figure 2-2. The diagonal boundary line represents the constraint \( t_r - L > t_o \), while the lower (horizontal) boundary line represents \( t_o \geq 0 \). Each circle represents one of the four possible types of KKT points, and the arrows shows the improving direction associated with each. Theorem 2-3 indicates that a KKT point located in the interior (labeled 1) or on the lower boundary (2) is always a global minimum, while a point on the \( t_r = t_o + L \) diagonal boundary (3 or 4) might instead be a local minimum or a saddle point.

When the minimum occurs at an interior point, combining Theorems 2-1 and 2-2 yields the relation

\[
bz(t_r', t_o') = k - h \left[ \frac{F(t_o')}{F(t_o' + L) - F(t_o')} \right].
\]
To compute an optimal pair \((t_0, t_r)\), rather than using generic NLP software we could take advantage of the preceding theorems and corollaries to customize both the KKT conditions and a search algorithm. We do not know whether the potential benefits of using such a customized NLP approach in practice would be worth the effort of the customization, but in Appendix A we describe this idea and outline how such an algorithm might proceed (we did not actually use it for the numerical studies in this chapter).

Having discussed both separate and joint optimization, we can now compare our joint approach to the more traditional approach in which \(RCF\) is used to choose \(t_r\), and then another function is used to determine \(t_0\). Note that

\[
JCF = \left[ OCF \right] + \left[ RCF \right] \int_{0}^{t} \frac{\bar{F}(x) \, dx}{\int_{0}^{t} \bar{F}(x) \, dx + \int_{t}^{t+r} F(x) \, dx}.
\]

The multiplier for \(RCF\) adjusts for the longer cycle time due to the possibility of a stock-out.

The comparison of joint to sequential optimization involves two types of cost rate increases. The first arises in using \(RCF\) to calculate \(t_r\), followed by \(JCF\) to calculate \(t_0\) given \(t_r\) (call this combination \(JCF:RCF\)). Here we lose by using \(RCF\), which ignores inventory considerations in determining maintenance policy. Optimizing \(JCF:RCF\) will determine the best possible sequential policy from the system's point of view.

The second cost increase comes from using some method of determining the inventory policy (given a maintenance policy) that is more myopic, easier to calculate, and/or easier to implement. Consider the following examples:

(a) One can ignore maintenance costs (due to \(b\) and \(c\)) in determining inventory policy, by using \(OCF\) to calculate \(t_0\) given \(t_r\) (call this combination \(OCF:RCF\)). Optimizing
**OCF:RCF** will determine the best possible sequential policy from the inventory manager's point of view;

(b) If the inventory manager does not want to keep track of the age of the machine, a policy of (s-1,s) form can be adopted (with \( s \leq 1 \)). **XJCF:RCF** is the same as **JCF:RCF** except that we restrict the choice of \( t_o \) to either extreme of 0 or \( t_r - L \). Optimizing **XJCF:RCF** will determine the best possible sequential (s-1,s) policy from the system's point of view;

(c) One can ignore both maintenance costs and the aging of the machine by using **XOCF:RCF**, which is the same as **OCF:RCF** except that we restrict the choice of \( t_o \) to either extreme of 0 or \( t_r - L \). Optimizing **XOCF:RCF** will determine the best possible sequential (s-1,s) policy from the inventory manager's point of view; or,

(d) Inventory optimization can ignore completely the existence of planned maintenance: its costs, the aging of the installed machine, and the use of preventative replacement. This allows the use of a simplified approximation such as **SOF** to determine \( t_o \) given \( t_r \) (call this combination **SOF:RCF**). This is the simplest method for an inventory manager to calculate and operate.

All of these inventory methods can be thought of as variations on a continuous time \((s-1,s)\) type of inventory policy. **JCF** and **OCF** imply that we use a (-1,0) policy until \( t = t_o \) when we switch to a (0,1) policy. Using **XJCF**, **XOCF**, or **SOF** to choose an endpoint for ordering is equivalent to choosing only a (-1,0) or a (0,1) policy with no switching. Since the optimal \( t_o \) is generally not an endpoint, a standard \((s-1,s)\) continuous time policy is generally not the optimal form of ordering policy for the simple system we consider; see [AK95] for more about the sub-optimality of standard \((s-1,s)\) policies.

We can also think of these different methods in terms of the cooperation and information exchange that can or must occur between the maintenance manager and the inventory manager. Joint optimization implies that the two managers must cooperate in making the two decisions together for the benefit of the overall system, in spite of the
natural incentives for each manager to make their respective decisions so as to obtain maximum benefits for their own sub-system. Separate optimization does not require such cooperation, but merely requires some flow of information from the maintenance manager to the inventory manager; the maintenance manager would not use any information about the inventory process (i.e. the lead time and the cost structure). The use of \(JCF:RCF\) requires that the inventory manager have access to and make use of full information about maintenance costs, the preventive maintenance schedule, and the machine's lifetime distribution. The inventory manager needs to know the preventive maintenance schedule and the machine’s lifetime distribution in order to use \(OCF:RCF\), while only the mean of the lifetime distribution needs be known in order to use \(SOF:RCF\). We naturally expect that the use of less information will tend to make the inventory decision process simpler but result in higher costs.

While it is difficult to make generalizations about the size of these cost increases, we can offer the following.

**Proposition 2-2:** Sequential optimization can be arbitrarily worse than joint optimization; i.e. the ratio \(JCF:RCF / JCF\) can be arbitrarily large.

**Proposition 2-3:** Using \(OCF:RCF\) instead of \(JCF:RCF\) to find \(t_o\) always gives an earlier ordering time; that is, \((t_o \text{ from } OCF:RCF) \leq (t_o \text{ from } JCF:RCF)\).

**Proposition 2-4:** Not only does \(SOF\) ignore interior points \((t_o : 0 < t_o < t_r - L)\) which may be optimal, but it is possible for \(SOF\) to choose the wrong endpoint \((t_o = 0 \text{ or } t_o = t_r - L)\).

Our model in this section does not explicitly deal with the issue of how much could be saved if we could keep multiple spares in stock or on order, but it does offer one indicator in the optimal value found for the ordering time. If the optimal \(t_o = 0\), we would suspect that further cost reductions could be obtained if we allowed for advance ordering of spares (i.e. relax the constraint \(t_o \geq 0\) to \(t_o \geq -L\)) and/or multiple spares in stock.
Conversely, an optimal $t_0 > 0$ indicates that nothing would be gained by relaxing the ordering constraint and that having more than one spare in stock would be sub-optimal.

In our modeling of the problem so far, we have taken as our objective the minimization of the expected cost rate; that is, the cost rate calculated over calendar time. We believe that in general this is the most reasonable objective. Some other authors, however, have also considered the maximization of the cost effectiveness, which they define as the expected proportion of time that the system is functioning (i.e. not short) divided by the expected cost rate during system functioning. This is equivalent to minimizing the expected cost rate calculated over operating time (rather than calendar time). If we were to use this as our objective, then our equation for $JCF$ would become slightly simpler, in that there would be one less term in the denominator.

$$JCF_1(t_o, t_r) = \frac{c + k \int_{t_o}^{t_o+L} F(x)dx + h \int_{t_o}^{t_r} -F(x)dx + bF(t_r)}{\int_{t_r}^{t_o+L} \overline{F}(x)dx}$$  \hspace{1cm} (2-15)

One could easily show that all of the results previously derived for the original objective (2-12) would continue to hold for this new objective (2-15).

Our modeling effort so far has also assumed that the system manager has reasonable estimates for all of the required cost coefficients; this may not be true, especially for the shortage cost rate $k$ and the breakage cost $b$ which often should include such intangibles as loss of customer good will. It might therefore be more appropriate to specify constraints on the solution that enforce certain levels of performance, instead of or in addition to the use of cost parameters; these constraints are often called service constraints. This technique is often used in inventory modeling, where the same problem can arise, but to our knowledge has not previously been considered in the literature for the single spare model. For example, in place of specifying a breakage cost, we might require
that the expected proportion of machines which fail in use be less than some maximum limit. Similarly, instead of specifying a shortage cost we might set a minimum on the expected proportion of demands for spares which are filled from stock (i.e. a minimum fill rate) or a maximum on the expected wait to receive a back-ordered spare. We can add these service constraints to the NLP minimization problem as follows.

**Minimize:** \( JCF(t_o, t_r) \)

**Subject to:**

\[
\begin{align*}
  t_o & \geq 0 \quad t_r \geq t_o + L \\
  F(t_r) & \leq \text{maximum expected failure proportion} \quad (2-16) \\
  \overline{F}(t_o + L) & \geq \text{minimum expected fill rate} \quad (2-17) \\
  \int_{t_o}^{t_o+L} F(x) \, dx & \leq \text{maximum expected wait} \quad (2-18)
\end{align*}
\]

Specified in this way, service constraints are easily handled in our search for an optimal solution pair \((t_o, t_r)\). We say this because each of them only acts on one of the two decision variables; in terms of the solution space illustrated in Figure 2-1, each constraint is either a horizontal or a vertical line. We can therefore use the results of Theorems 2-1 and 2-2 to show that Theorem 2-3 still holds, and so our search procedure should proceed as easily with service constraints as without. Note that if a service constraint is used instead of the shortage cost \(k\), i.e. so that \(k = 0\), then the optimal ordering time will be as late as is feasible, i.e. the optimal \(t_o\) will be a boundary solution.

We next turn to numerical studies on two small sets of test problems to illustrate the kind of performance that can result from sequential versus joint optimization. We did not use any service constraints. First we tried a small parameter set obtained from an existing business; the goals were simply to see what could happen, and to do a "reality check" on our thinking (i.e. do the cost functions behave as expected?). Table 2-1 shows the cost coefficients and calculated optima for 12 combinations of distributions and costs.
These 12 cases were derived in the following manner. Case 1 uses informed estimates for cost coefficients and for $L$ and $\mu$ for a computer used by a business consultant based in a major city (i.e. in a climate controlled office with ready access to spare parts); the failing part is a hard disk drive. The failure distribution $F(t)$ is arbitrarily assumed to be of Weibull form $F(t) = 1 - e^{-\lambda t^\eta}$ with shape parameter $\eta = 3$, $\mu = 2000$ days, $L = 1$ day. Cases 2 and 3 use the same situation but with the inventory cost coefficients doubled and halved. Cases 4-6 are duplicates of 1-3 but with $\eta = 1.5$. Cases 7-12 are for the same consultant when doing business in a remote northern community (harsher environment, no local source of spares); $\mu = 1000$ days, $L = 3$ days. All cases use $c = 500$ and $b = 950$.

The columns in the table show calculated optimal times and expected cost rates of joint optimization, compared to sequential optimization using each of $JCF$, $OCF$, and $SOF$ to find $t_j$. The columns labeled "%A" show the percentage increase in the expected cost rate over the calculated joint optimum; i.e. $\%A = \frac{\text{sequential cost} - \text{joint cost}}{\text{joint cost}}$. Calculations were done using the Maple mathematics software package. Note that the calculated $t_j$ and $t_r$ values for all the sequential methods were obtained by simple univariate optimization, so the proven pseudo-convexity properties of $JCF$ are sufficient to guarantee that these are in fact optima.

Table 2-1: Example Solutions

<table>
<thead>
<tr>
<th>#</th>
<th>$L$</th>
<th>$\eta$</th>
<th>$k$</th>
<th>$h$</th>
<th>$t_r$</th>
<th>$t_j$</th>
<th>$t_r$</th>
<th>$%A$</th>
<th>$t_j$</th>
<th>$%A$</th>
<th>$t_j$</th>
<th>$%A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3.0</td>
<td>300</td>
<td>1.0</td>
<td>1239</td>
<td>1238</td>
<td>0.621</td>
<td>1360</td>
<td>0.6</td>
<td>1359</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3.0</td>
<td>600</td>
<td>2.0</td>
<td>1151</td>
<td>1150</td>
<td>0.665</td>
<td>1360</td>
<td>2.4</td>
<td>1359</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3.0</td>
<td>150</td>
<td>0.5</td>
<td>1294</td>
<td>1293</td>
<td>0.596</td>
<td>1360</td>
<td>0.3</td>
<td>1359</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1.5</td>
<td>300</td>
<td>1.0</td>
<td>2030</td>
<td>2029</td>
<td>0.877</td>
<td>2536</td>
<td>0.6</td>
<td>2535</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1.5</td>
<td>600</td>
<td>2.0</td>
<td>1718</td>
<td>1717</td>
<td>1.001</td>
<td>2536</td>
<td>1.8</td>
<td>2535</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1.5</td>
<td>150</td>
<td>0.5</td>
<td>2249</td>
<td>2248</td>
<td>0.813</td>
<td>2536</td>
<td>0.1</td>
<td>2535</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>3.0</td>
<td>300</td>
<td>1.0</td>
<td>542</td>
<td>539</td>
<td>1.405</td>
<td>680</td>
<td>4.7</td>
<td>653</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>3.0</td>
<td>600</td>
<td>2.0</td>
<td>475</td>
<td>472</td>
<td>1.597</td>
<td>680</td>
<td>13.0</td>
<td>653</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>3.0</td>
<td>150</td>
<td>0.5</td>
<td>596</td>
<td>593</td>
<td>1.285</td>
<td>680</td>
<td>1.5</td>
<td>653</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>1.5</td>
<td>300</td>
<td>1.0</td>
<td>851</td>
<td>641</td>
<td>2.222</td>
<td>1268</td>
<td>1.1</td>
<td>635</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>1.5</td>
<td>600</td>
<td>2.0</td>
<td>562</td>
<td>559</td>
<td>2.866</td>
<td>1268</td>
<td>4.8</td>
<td>635</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>1.5</td>
<td>150</td>
<td>0.5</td>
<td>1058</td>
<td>648</td>
<td>1.862</td>
<td>1268</td>
<td>0.3</td>
<td>635</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>
The results in Table 2-1 suggest that in most cases although the $t_0$ and $t_r$ values may differ noticeably between joint and sequential optimization, the actual cost rates are tend to be quite similar. This is particularly true of the sequential methods, which tend to give results very similar to each other. The cost differences tended to be larger for the cases where the inventory costs were large and the lead time relatively long (i.e. where inventory considerations were most relevant); the largest difference was observed (in case 8) where both of these were true. Note that as predicted by Proposition 2-3, $OCF:RCF$ always gives an earlier ordering time than does $JCF:RCF$.

The next step in our numerical work was intended to explore the behaviour of our heuristics over a wider range of coefficients, and to try to characterize when the cost differences would or would not be large. We again chose a failure distribution $F(t)$ of Weibull form with shape parameter $\eta = 3$, scale parameter $\lambda = 1$; so $\mu \approx 0.9$. For all cases $c = 1$ while the other parameters were varied as follows: $L = \{0.5, 0.05, 0.005\}$, the ratio of $b/c = \{10, 1, 0.1\}$, the ratio of $h/c = \{5, 0.5, 0.05\}$, and the ratio of $k/h = \{10000, 1000, 100, 10, 1\}$. In total this gave 135 different cases or combinations of parameters. In Table 2-2 we summarize the joint versus sequential differences for these cases by giving the mean and other statistics of the cost rate differences relative to the joint optima.

<table>
<thead>
<tr>
<th>Sequential Procedure</th>
<th>$JCF:RCF$</th>
<th>$OCF:RCF$</th>
<th>X$JCF:RCF$</th>
<th>X$OCF:RCF$</th>
<th>$SOF:RCF$</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Cost Increase - Mean</td>
<td>2.6</td>
<td>3.3</td>
<td>4.3</td>
<td>4.6</td>
<td>5.1</td>
</tr>
<tr>
<td>% Cost Increase - Standard Deviation</td>
<td>6.5</td>
<td>7.5</td>
<td>9.5</td>
<td>9.8</td>
<td>10.9</td>
</tr>
<tr>
<td>% Cost Increase - Largest</td>
<td>28.0</td>
<td>29.3</td>
<td>52.6</td>
<td>52.6</td>
<td>52.6</td>
</tr>
<tr>
<td># Cases with increase &lt; 1%</td>
<td>110</td>
<td>106</td>
<td>94</td>
<td>93</td>
<td>92</td>
</tr>
<tr>
<td># Cases with increase 1-10%</td>
<td>9</td>
<td>10</td>
<td>20</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td># Cases with increase &gt; 10%</td>
<td>16</td>
<td>19</td>
<td>21</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>

As expected, the sequential combination $JCF:RCF$ gives the closest results to joint optimization. On average the differences were quite small, but they varied greatly from case to case: the vast majority of cases had negligible differences, but about one in ten had
a difference >10%. Moving across the table we see that both the frequency of large
differences and their magnitude increase somewhat as we use less and less information
about maintenance in optimizing inventory. The methods increasingly behave in a "hit-or-
miss" fashion, tending to give results that are either very good or very bad; there did not
seem to be any pattern as to which method did better in any given instance (except as
discussed in the next paragraph). $OCF:RCF$ generally does better than $XOCF:RCF$, which
implies that it is more important for inventory to allow scheduled orders (due to part
aging) than to account for maintenance costs. $XOCF:RCF$, which ignores both of those
factors, generally performs worse than both of these methods. The simplistic $SOF:RCF$ is
the worst performer on all counts, though of course it is also the easiest to use.

What Table 2-2 does not show is under what circumstances the large differences
occurred. Our numerical work did not reveal any simple algebraic rules for this, but it did
reveal a pattern that we can describe qualitatively and which we think makes good
intuitive sense. Large differences most commonly occurred when all four cost coefficients
were in some sense balanced; we might think in terms of the potential costs that could be
incurred over the life of the part (i.e. the life-cycle costs). When maintenance costs were
much larger than inventory costs, they drove the optimization process and made the joint
versus sequential issue irrelevant at the system level. Similarly, if any one life-cycle cost
was very large (or very small), its presence (respectively, absence) made the optimization
process much more straightforward. It was when all four life-cycle costs were
operationally significant that it became important to trade them off in an optimal manner;
and so then joint optimization showed a material advantage.

A comment about the ordering lead time can also be made. The primary effect of
increasing lead time relative to the part life is to increase the potential shortage cost
incurred; so indirectly the lead time also plays a role in creating or avoiding the balanced
life-cycle cost structure described above. The lead time had a more direct effect when it
was very large; e.g. $L = 0.5$ was more than half the expected life of the part. In these
circumstances, it was more difficult to find a viable solution (because the part was often failing before spare delivery), and so working around this lead time problem came to dominate the solution process.

The most obvious limitation of these numerical results is that they are based upon a particular choice of failure distribution. We expect that the qualitative issues we describe would hold in general, though the size and frequency of the cost differences would vary. Consequently, we feel able to conclude that reasonable gains can be made using joint optimization, and that if sequential optimization is unavoidable one can benefit from using as much maintenance information as is available when making the subsequent inventory decision.

2.05 Early Ordering with Replacement on Receipt

The previous section restricted $0 \leq t_o$ but some other authors, e.g. Thomas & Osaki in [TO78a] and [TO78b], have allowed $-L \leq t_o \leq 0$ as an alternative, so that a spare could be ordered before the start of the current cycle. That this could be appropriate would be indicated by a solution to $JCF$ in which the optimal $t_o = 0$ (a boundary solution); conversely, $t_o > 0$ indicates that there is no value in having a spare arrive any earlier. Such a boundary solution will often arise when the order lead times are relatively long compared to the machine life (e.g. if $L > \mu$). To keep the model tractable these authors have also restricted the maintenance policy to performing replacement on receipt, i.e. always install a new spare as soon as it arrives, regardless of the installed machine's functioning or age (hence fix $t_r = t_o + L$). This is obviously a severe restriction on maintenance flexibility, but it does allow for considerable simplification in the analysis. The replacement on receipt restriction enables us to know the exact spare arrival times and cycle lengths in advance; without this restriction, the timing uncertainty of future cycles makes analysis intractable. [TO78b] also tried to analyze the model without this restriction but erroneously treated the future arrival time of a spare as being independent of the choice of
ordering time. Replacement on receipt also implies that the system may have multiple
spares on order but never any stock on hand, and hence no holding costs will be incurred.

Simplifying $JCF$ to fit the new assumptions above yields the following equation
$JED$ for the expected cost rate, with $-L \leq t_0 \leq 0$ or equivalently $0 \leq t_0 + L \leq L$.

$$JED(t_0) = \frac{c + bF(t_0 + L) + k \int_{0}^{t_0 + L} F(x) dx}{t_0 + L}$$

(2-19)

This equation need not be pseudo-convex over its entire range, but fortunately it is
over the range of interest to us.

**Theorem 2-4:** Over its viable range (i.e. all $t_0 : JED(t_0) < k$), $JED(t_0)$ is pseudo-convex in
$t_0$ with a unique minimum at:

(a) $-L < t_0 \leq 0$ if $bf(t_0 + L) + kF(t_0 + L) - JED(t_0) = 0$;

(b) $t_0 = 0$ if $bf(L) + kF(L) - JED(0) \leq 0$.

Note that (b) implies there is no benefit to ordering early.

For example, let $c = $ 10, $b = $ 20, $h = $ 2, $k = $ 100, $L = 1$, and $F(t) = 1 - e^{-t}$, so
that the delivery lead time is longer than the average machine life ($L < \mu \approx 0.9$).

Optimization of $JCF$ from the previous section with these parameters gives its best
solution of $JCF(0.00, 0.80) = $ 41.03. If we restrict maintenance to replacement on
receipt but allow early ordering of spares, then we can optimize $JED$ in this section to find
a solution of $JED(-0.48) = $ 27.65. In this example, it is optimal to order the spare one
renewal cycle in advance of its use, but in general the best ordering time may be several
periods in advance.

### 2.06 Single Spare With Returns

The previous section considered policies in support of a machine which was
deteriorating with age, i.e. was IFR, but in this section we consider instead a machine
whose reliability is improving with age, i.e. is DFR. There is clearly no requirement for
preventive replacement of the machine, but there is still a need to provide a spare for
eventual use as a corrective replacement once the original fails. Furthermore, since over
time the installed part is less and less likely to fail, it may be beneficial to return a spare
from our stock to its source once the risk of failure has sufficiently declined. Thus to run
such a system at minimal cost, we must determine an optimal policy for the order and
return of spares.

Why should we be interested in such a system? First, while one may not always
have the option of returning a spare to its source, this option is certainly not unheard of;
within a large corporation for example, a regional repair shop could ship back an
unneeded spare to the company's central warehouse. Second, while it is true that the
reliability of most machines deteriorates with age, there are examples where the reverse is
true, at least during the early part of the operating life when so-called infant mortality may
be a problem such as for components with a bathtub-shaped hazard rate curve. To our
knowledge this option of returning an unused spare has not previously appeared in the
literature for the single spare model.

We need to add some notation to model this situation. We use $t_s$ to represent the
scheduled time at which to return a spare from stock to the external source; we require $0 
\leq t_0 \leq t_s - L \leq \infty$. A restocking cost $r$, representing shipping and handling etc., will be
incurred if a spare is returned unused to its source. The form of our spare policy is as
follows. At time zero, the newly installed machine begins operation. If it fails before $t_0$,
then immediately place an order for a new one. If the machine is still operating at $t_0$, then
place an order for a spare at that time. Once a spare arrives, use it to replace the installed
machine if it has previously failed; otherwise hold the spare in stock until it is needed. If
the installed machine has not failed by time $t_s$, then return (send back) the spare to its
source. If the installed machine fails after $t_s$, then immediately order a new spare and
install it upon arrival. Our decision problem is to find a pair of values $(t_0, t_s)$ that
minimizes the expected cost rate. We continue to assume that we can deal with only one
spare at a time, and that the order lead time is deterministic.
We note as an aside that our order and return policy is somewhat related to an continuous review inventory system using an \((s-1, s)\) ordering policy, except that our \(s\) changes twice over time. We set \(s = 0\) initially, then at \(t_o\) switch to \(s = 1\), and finally at \(t_s\) switch back to \(s = 0\).

For this model, the expected shortage (or stock-out) time per cycle is

\[
\int_0^{t_o} Lf(t)dt + \int_{t_o}^{t_o+L} (t_o + L - t)f(t)dt + \int_{t_o}^{\infty}Lf(t)dt = \int_{t_o}^{t_o+L} F(t)dt + L\overline{F}(t_s)
\]

and the expected holding time per cycle is

\[
\int_{t_o+L}^{t_s} (t - t_o - L)f(t)dt + (t_s - t_o - L)\overline{F}(t_s) = \int_{t_o+L}^{t_s} F(t)dt
\]

The cycle length is the sum of the machine lifetime plus any shortage time, so the expected cycle length is

\[
\mu + \int_{t_o}^{t_o+L} F(t)dt + L\overline{F}(t_s)
\]

The replacement cost \(c\) will be incurred every cycle, while the restocking cost \(r\) will be paid only if we return the spare, i.e. only if the installed machine fails after the scheduled return time. Combining these terms with our other cost parameters gives an equation for the system's expected cost rate, which we call \(ORS\).

\[
ORS(t_o, t_s) = \frac{c + k \int_{t_o}^{t_o+L} F(t)dt + kL\overline{F}(t_s) + h \int_{t_o}^{t_s} F(t)dt + r\overline{F}(t_s)}{\mu + \int_{t_o}^{t_o+L} F(t)dt + L\overline{F}(t_s)}
\]

This function has some useful properties (recall that \(0 \leq t_o \leq t_s - L \leq \infty\)).
**Theorem 2-5:** For any given \( t_s \), ORS is pseudo-concave in \( t_o \); hence the optimal ordering time will be either \( t_o = 0 \) (order immediately) or \( t_o = \infty \) (order on failure).

**Theorem 2-6:** For any given \( t_o \), ORS is pseudo-convex in \( t_s \); hence there will be a unique optimal return time at:

(a) \( t_s = t_o + L \) if \( r + kL - h / z(t_o + L) - ORS(t_o, t_o + L)L \leq 0 \);
(b) \( t_o + L < t'_s < \infty \) if \( r + kL - h / z(t'_s) - ORS(t_o, t'_s) L = 0 \);
(c) \( t_s = \infty \) if \( r + kL - h / z(\infty) - ORS(t_o, \infty) L \geq 0 \).

Case (a) implies that return upon receipt is optimal, whereas case (c) implies that it is optimal to keep the spare until failure.

These results allow us to readily determine an optimal spare policy \((t_o, t_s)\). We proceed as follows, noting that if \( t_o = \infty \) then we must also have \( t_s = \infty \):

Step 1: compute the value of ORS\((\infty, \infty)\), i.e.

\[
ORS(\infty, \infty) = \frac{c + kL}{\mu + L}
\]  

(2-24)

Step 2: fix \( t_o = 0 \) and optimize ORS over \( t_s \) to find \( ORS(0, t'_s) \);

Step 3: compare \( ORS(\infty, \infty) \) with \( ORS(0, t'_s) \) and choose the least costly policy.

In effect, we need to perform only a univariate optimization plus one extra function evaluation of ORS in order to determine our optimal policy.

Note that we allow \( t_s = t_o + L \), i.e. the spare can be returned as soon as it arrives.

This "sneaky" policy avoids incurring any holding costs, while ensuring that a spare is "on the way" during the initial period of machine operation (when the risk of failure is highest).

If the restocking cost \( r \) is sufficiently high (as we would expect in most applications) then this "sneaky" policy will not be optimal.

The following describes other properties of ORS.

**Proposition 2-5:** A change in the cost coefficients will tend to change the optimal solution as follows:
(a) for a given $t_o$, the optimal return time $t_s$ is non-decreasing in the restocking cost $r$ and the shortage cost rate $k$, and non-increasing in the holding cost rate $h$ and the replacement cost $c$;
(b) for a given $t_s$, the optimal ordering time $t_o$ is non-decreasing in $c$, $r$, and $h$, and non-increasing in $k$.

This proposition reflects what we would intuitively expect to happen if the system costs were to change. For example, when the holding cost $h$ increases we tend to avoid holding a spare by ordering later and returning earlier; conversely, when the shortage cost $k$ increases we tend to avoid the risk of shortage by ordering a spare earlier and holding it longer.

Now suppose that we begin at time zero with a spare already in stock, so that we don't need to place an order but have only to decide when (if ever) to return this spare. This obviously would not happen on an ongoing basis, but could apply for example upon start-up of a new system or when a replacement is performed before failure for reasons of technological change. The mathematics which describe this situation are similar but slightly simpler than those for our original model. The expected shortage time per cycle, the expected holding time per cycle, and the expected cycle length are respectively

$$
L \overline{F}(t_s), \quad \int_0^{t_s} \overline{F}(t) dt, \quad \mu + L \overline{F}(t_s).
$$

The replacement cost $c$ will still be incurred every cycle, while the restocking cost $r$ will be paid if the installed component fails after the spare has been returned. Combining these terms with our other cost parameters gives an equation for the system's expected cost rate, which we call RS.

$$
RS(t_s) = \frac{c + kL \overline{F}(t_s) + h\int_0^{t_s} \overline{F}(t) dt + r \overline{F}(t_s)}{\mu + L \overline{F}(t_s)}
$$

(2-25)
Not surprisingly, this equation RS has similar properties to ORS.

**Theorem 2-7:** RS is pseudo-convex in $t_s$; hence there will be a unique optimal return time, either $t_s = 0$ (return immediately), $0 < t_s' < \infty$, or $t_s = \infty$ (keep until needed at failure).

- (a) $t_s = 0$ if $r + kL - h/z(0) - RS(0)L \leq 0$;
- (b) $0 < t_s' < \infty$ if $r + kL - h/z(t_s') - RS(t_s')L = 0$;
- (c) $t_s = \infty$ if $r + kL - h/z(\infty) - RS(\infty)L \geq 0$.

**Proposition 2-6:** The optimal return time $t_s$ is non-decreasing in $r$ and $k$, and non-increasing in $h$ and $c$.

Next we present a hypothetical numerical example for our ORS model and compare the cost of our $(t_o, t_s)$ policy to the cost of using a simpler policy that does not allow returns (in effect, $t_s = \infty$). Suppose we have a machine which fails according to a DFR Weibull distribution with shape parameter $\eta = 0.5$ and scale parameter $\lambda = 1$, so $\mu = 2$ years. Let the lead time $L = 0.05$ years, the cost of replacement $c = $1000, the restocking cost $r = $100, the holding cost rate $h = $200/year, and the shortage cost rate $k = $10,000/year.

If we use a policy which does not allow returns, then our options are to order either immediately and hold until failure, or to not order until failure. If we order immediately, we will incur a holding cost until failure but no restocking cost. Setting $t_o = 0$ and $t_s = \infty$ in (2-23) gives the expected cost rate.

\[
1000 + 10000 \int_0^{0.05} \left[ 1 - e^{-t_s^3} \right] dt + 10000(0.05)(0) + 200 \int_0^{\infty} e^{-t_s^3} dt + 100(0) \\
\approx 728
\]
If instead we do not order until failure, then both the holding and the restocking costs are avoided but we expect to pay more for shortage costs. Setting $t_o = \infty$ and $t_s = \infty$ in (2-23) gives the expected cost rate for this policy.

$$\frac{1000 + 10000(0.05) + 10000(0.05)(0) + 200(0) + 100(0)}{2 + 0.05 + 0} \approx 732$$

(2-27)

We would of course chose the less expensive of the two and so order immediately and hold until failure, at a cost rate of $728 \text{ / year}$.

If we now allow the return of spares, then we may also choose to order immediately and return at some time $t_s$. Optimizing ORS$(0, t_s)$ over $t_s$ gives an optimal $t_s' = 2.0$ years, at a cost rate of

$$\frac{1000 + 10000 \int_0^{0.05} [1 - e^{-t^3}] dt + 10000(0.05)e^{-2.00t^3} + 200 \int_0^{2.00} e^{-t^3} dt + 100e^{-2.00t^3}}{2 + \int_0^{0.05} [1 - e^{-t^3}] dt + (0.05)e^{-t^3}} \approx 679$$

So in this particular example, we can expect to save approximately 7% by returning an unused spare (at $679 \text{ / year}$) rather than holding it until failure (at $728 \text{ / year}$).

In formulating both versions of this model, we use the minimization of the expected cost rate per unit total time as our objective. This is a common objective and is arguably the best one. In some circumstances, however, the minimization of the expected cost rate per unit operating time (equivalently, the maximization of the so-called cost effectiveness) may be a more reasonable objective. We therefore note in passing that all of the results in this section still apply to this alternate objective, albeit in a slightly modified (and simplified) form.

### 2.07 Single Spare Extended Cost Structure

In Section 2.04 of this chapter we considered a model with a simplified cost structure so as to keep our analysis relatively straight forward. In this section we show
that the same results apply to a somewhat more general cost structure; we retain the restriction of a single spare with deterministic order lead times.

First allow the replacement cost \( c v(t) \) to be a non-increasing convex function of the age of the installed machine. The expected replacement cost per cycle then becomes

\[
cC(t_r) = c \left\{ \int_0^{t_r} v(x) f(x) \, dx + v(t_r) \overline{F}(t_r) \right\}
\] (2-29)

We also allow for an increasing operating cost at rate \( a q(t) \), which will add an expected cost per cycle of

\[
aA(t_r) = a \int_0^{t_r} q(x) \overline{F}(x) \, dx
\] (2-30)

Finally we allow the replacement action to take a random amount of time, with a mean of \( R \). This will increase the expected cycle length by \( R \) and add an additional shortage cost \( kR \) per cycle. Adding these three features to the equation \( JCF \) gives the extended cost rate \( JOD \).

\[
JOD(t_o, t_r) = \frac{cC(t_r) + k \int_{t_o}^{t_o+L} F(x) \, dx + kR + h \int_{t_o}^{t_r} \overline{F}(x) \, dx + bF(t_r) + aA(t_r)}{\int_{t_o}^{t_r} \overline{F}(x) \, dx + \int_{t_o}^{t_o+L} F(x) \, dx + R}
\]

This extended function has the same properties as does \( JCF \), so Theorems 2-1, 2-2, 2-3, and 2-4 still apply (with appropriate additions in notation). For this reason, the proofs in Appendix B for these theorems are written using the extended cost function \( JOD \) rather than the simpler \( JCF \).

### 2.08 Scheduled versus Unscheduled Orders

In practice there can be a difference in the lead times and delivery costs between orders which are scheduled (i.e. placed at \( t_o \)) and those which are unscheduled (i.e. placed...
on failure before \( t_o \). Unscheduled orders may involve larger administrative costs to place, and since they are of a more urgent nature they may be delivered by a faster but more expensive method of transport (e.g. by air instead of by land). This feature has been included in many previous works on this subject, including [Os77], [TO78a], [KO78a], [KO78b], [KO78d], [KO81], [KH81], [SS89], and [Sr92].

Allowing for multiple lead times and delivery costs in our model involves a fairly straightforward addition to our notation, but can complicate the analysis considerably. That a two number \((t_o, t_r)\) policy is generally not optimal can be argued as follows. If the installed machine fails well before \( t_o \), then we will order immediately, even though this will involve paying some premium for expedited delivery; and if the machine fails after \( t_o \), an order will have already been placed and so we can only wait for its arrival (assuming of course that we cannot retroactively hasten its delivery). If however the machine fails just slightly before \( t_o \), then depending on the relative costs and delivery speeds of the two order types we may wish to wait until \( t_o \) to place a less expensive scheduled order rather than immediately making a more costly emergency order; thus an optimal policy form will now require more decision variables in order to specify this waiting procedure.

Whatever analysis we do for a system like this, we must be clear in our assumptions about what types of orders can or must be placed and their characteristics. Given a particular set of such assumptions, we can adapt our equations from the previous sections to distinguish amongst order types.

Suppose that lead times remain deterministic but are now distinct: scheduled orders still take \( L \) units of time to arrive, but unscheduled orders (i.e. on failure) require only \( L_x \leq L \) units of time to arrive and incur a cost premium \( c_x u(t) \), where \( c_x \) is a cost coefficient and \( u(t) \) is a non-decreasing function of the component age (e.g. \( u(t) = 1 \) for all \( t \) implies a constant cost premium). We will use a *double age* \((t_x, t_o)\) ordering policy as follows. If the installed machine fails before the *expediting age* \( t_x \) then immediately place an unscheduled order. If failure occurs after \( t_x \) but before \( t_o \) then wait and let the
scheduled order occur at \( t_0 \). If there has been no failure by \( t_0 \) then place a scheduled order at that time. Preventive replacement will occur as before at \( t_r \) if failure has not occurred by then. Thus the system policy is specified by the three values \((t_x, t_0, t_r)\), where \( 0 \leq t_x \leq t_0 \leq t_r - L \leq \infty \). Note that our use of two critical ages for ordering seems not to have appeared in the single spare literature before.

The expected shortage time per cycle will now be

\[
K(t_x, t_0) = R + \int_{t_x}^{t_x+L} F(t)dt - F(t_x)[(t_0 + L) - (t_x + L_x)]
\]

(2-32)

The expected cycle time also changes to

\[
\int_0^{t_x} \overline{F}(t)dt + R + \int_{t_x}^{t_x+L} F(t)dt - F(t_x)[(t_0 + L) - (t_x + L_x)]
\]

\[
= \int_0^{t_x} \overline{F}(t)dt + K(t_x, t_0)
\]

(2-33)

The expected ordering cost premium is

\[
c_x \int_0^{t_x} u(t)f(t)dt
\]

(2-34)

The following equation describes the expected cost rate \( JT_D(t_x, t_0, t_r) \).

(2-35)

\[
cC(t_x) + bF(t_r) + aA(t_r) + kK(t_x, t_0) + h \int_{t_x+L}^{t_x+L} \overline{F}(t)dt + c_x \int_0^{t_x} u(t)f(t)dt
\]

\[
\frac{\int_0^{t_x} \overline{F}(t)dt + K(t_x, t_0)}{t_x+L}
\]

Note that setting \( L_x = L, \ t_x = t_0, \) and \( c_x = 0 \) transforms \( JT_D \) back into the \( JOD \) equation of the previous section.
Solving for optimal values of $t_x, t_o$ and $t_r$ is fairly straightforward, as the following suggests.

**Theorem 2-8:** $JTD(t_x, t_o, t_r)$ is pseudo-convex in each of $t_x, t_o$ and $t_r$, so that there will be a unique minimum for each variable (given values for the other two) as described below:

(a) Given $t_x$ and $t_o$, the optimal $t_r$ is at:
   (i) $t_r = t_o + L$ if $h + b z(t_o + L) + a q(t_o + L) + c v(t_o + L) - JTD(t_x, t_o, t_o + L) \geq 0$;
   (ii) $t_o + L \leq t_r \leq \infty$ if $h + b z(t_r) + a q(t_r) + c v(t_r) - JTD(t_x, t_o, t_r) = 0$;
   (iii) $t_r = \infty$ if $h + b z(\infty) + a q(\infty) + c v(\infty) - JTD(t_x, t_o, \infty) \leq 0$.

Case (i) implies that a spare should be installed upon receipt, while (iii) implies that replacements should be performed only at failure;

(b) Given $t_x$ and $t_r$, the optimal $t_o$ is at:
   (i) $t_o = t_x$ if $k - h = \frac{F(t_x + L)}{F(t_x) - F(t_x + L)} - JTD(t_x, t_x, t_r) \geq 0$
   (ii) $t_x \leq t_o \leq t_r - L$ if $k - h = \frac{F(t_o + L)}{F(t_x) - F(t_o + L)} - JTD(t_x, t_o, t_r) = 0$
   (iii) $t_o = t_r - L$ if $k - h = \frac{F(t_r)}{F(t_x) - F(t_r)} - JTD(t_x, t_r - L, t_r) \leq 0$

Case (i) implies that a spare should be ordered as soon as possible, while (iii) implies that a spare should be ordered just in time for scheduled replacement;

(c) Given $t_o$ and $t_r$, the optimal $t_x$ is at:
   (i) $t_x = 0$ if $\frac{c_x u(0)}{(t_o + L) - (L_x)} - k + JTD(0, t_o, t_r) \geq 0$
   (ii) $0 \leq t_x \leq t_o$ if $\frac{c_x u(t_x)}{(t_o + L) - (t_x + L_x)} - k + JTD(t_x, t_o, t_r) = 0$
(iii) \( t_x = t_o \) if \( \frac{c_x u(t_o)}{L - L_x} - k + JTD(t_o, t_o, t_r) \leq 0 \)

Case (i) implies that unscheduled orders should never be placed, while (iii) implies that a spare should always be ordered as soon as the installed machine fails.

Note that a two number policy \((t_o, t_r)\) is equivalent to fixing \( t_x = t_o \), which will not generally lead to an optimal solution for this problem. As an example, let \( c = 10 \), \( b = 20 \), \( h = 2 \), \( k = 100 \), \( L = 0.05 \), \( L_x = 0.05 \), \( c_x = 5 \), \( u(t) = 1 \), and \( F(t) = 1 - e^{-t^3} \), so that unscheduled orders require the same lead time as scheduled orders but are more expensive (assume other costs are zero). Optimization of \( JTD \) with these parameters but fixing \( t_x = t_o \) (a single age ordering policy) gives its best solution of \( JTD(0.39, 0.39, 0.62) = 25.78 \). If we remove the restriction on \( t_x \) and so use a double age ordering policy, then we can optimize \( JTD \) to find a lower cost solution of \( JTD(0.32, 0.39, 0.62) = 25.67 \).

Our treatment here generalizes much of the past work dealing with deterministic lead times, so that [Os77], [TO78a], [KO78a], [KO78b], and [KO78d] can be viewed as special cases of the work herein.

2.09 Single Spare, Random Lead Time

In this section we again assume that scheduled and unscheduled orders are identical in both cost and lead time, but now the lead time is allowed to be random with arbitrary distribution \( G(t) \); this was the case in [KH81], [PP86a], [PP86b], [SS89], and [Sr92]. In addition to making formulae somewhat complicated, the randomness also has an effect on the optimal form of the age replacement policy: a conventional single critical age is not optimal. Due to the randomness of its delivery time, an ordered spare may not arrive in time for the scheduled preventive replacement at \( t_r \); so a policy must indicate what to do when this occurs. The new maintenance policy is a double age replacement policy \((t_r, t_w)\): if the machine survives until age \( t_r \) replace it if a spare is available at that time or as soon as possible thereafter, but shut down the machine at the wear-out age \( t_w \geq t_r \), even if the spare has still not arrived (thus incurring shortage costs until it does arrive).
Note that the single age replacement policy from the previous sections is a special case of this in which \( t_w = t_r \). Since scheduled and unscheduled orders are identical, the inventory policy remains a single critical ordering age \( t_0 \), so our optimal system policy has the three decision variables \((t_0, t_r, t_w)\), where \( 0 \leq t_0 \leq t_r \leq t_w \leq Q \). Note that our use of two critical ages for replacement seems not to have appeared in the single spare literature before.

To derive the expected cost rate for this system, we need to consider two random events: the time of failure of the installed machine, and the time of arrival of an ordered spare. The system behaviour and hence costs incurred depend on when each of these events occurs, and in particular which one occurs first. Figure 2-3 shows the possible orders of events that can occur in the system during each renewal cycle.

Figure 2-3: Timing of Events with Random Lead Times

(a) Machine survives until scheduled replacement and spare arrives before then

\[
\begin{array}{cccc}
0 & \rightarrow & S & tr & tw \\
\end{array}
\]

(b) Machine fails before scheduled replacement but after spare arrival

\[
\begin{array}{cccc}
0 & \rightarrow & S & X & tr & tw \\
\end{array}
\]

(c) Machine fails after spare is ordered but before it arrives

\[
\begin{array}{cccc}
0 & \rightarrow & X & S & tr & tw \\
\end{array}
\]

(d) Machine fails before spare is scheduled for ordering

\[
\begin{array}{cccc}
0 & X & \rightarrow & S & tr & tw \\
\end{array}
\]

(e) Spare arrives after scheduled replacement and machine survives until then

\[
\begin{array}{cccc}
0 & \rightarrow & tr & S & tw \\
\end{array}
\]
The terms in the cost rate equation for this random lead time model cover the same types of costs as before but are complicated by the need to consider all of the possible combinations of events. The expected replacement cost per cycle is $c$ times

\[ C(t_o, t_r, t_w) = \int_0^{t_r} v(t) f(t) dt + \int_{t_r}^{t_w} \overline{F}(t_r) G(t_r - t_o) v(t_r) + \overline{F}(t_w) \overline{G}(t_w - t_o) v(t_w) \]

\[ + \overline{G}(t_w - t_o) \int_0^{t_x} v(t) f(t) dt + \int_{t_w}^{t_o} \overline{F}(t_w) \int_{t_r}^{l_r} v(t_o + l) g(l) dl \]

\[ + \int_{t_r}^{t_w} \int_{t_r}^{t_x} v(t) f(t) dt g(l) dl + \int_{t_r}^{t_o} v(t_o + l) \int_{l_r}^{l_w} f(t) dt g(l) dl \]

The probability of breakage each cycle is

\[ B(t_o, t_r, t_w) = F(t_r) + \int_{t_r}^{t_w} \overline{G}(t - t_o) f(t) dt \]

The expected operating cost per cycle is $a$ multiplied by

\[ A(t_o, t_r, t_w) = G(t_r - t_o) \int_0^{t_r} q(x) \overline{F}(x) dx + \overline{G}(t_w - t_o) \int_0^{t_w} q(x) \overline{F}(x) dx \]

\[ + \int_{t_w}^{t_r} \int_0^{t_r} q(x) \overline{F}(t) dx g(l) dl \]

The expected holding time per cycle $H(t_o, t_r, t_w)$ is

\[ \int_{t_r}^{t_w} \left[ t - t_o - l \right] g(l) dl f(t) dt + \overline{F}(t_r) \int_0^{t_r} \left[ t_r - t_o - l \right] g(l) dl \]

\[ \int_0^{t_r} \left[ t - t_o - l \right] g(l) dl f(t) dt + \overline{F}(t_r) \int_0^{t_r} \left[ t_r - t_o - l \right] g(l) dl \]
The expected shortage time per cycle $K(t_o, t_w)$ is

\[ R + LF(t_o) + \int_{t_o}^{t_o+L} g(l) dl f(t) dt + \frac{\int_{t_o}^{t_o+L} g(l) dl}{t_o} + \int_{t_o}^{t_o+L} g(l) dl \]

\[ = R + LF(t_o) + \int_{t_o}^{t_o+L} g(l) dl f(t) dt + \frac{\int_{t_o}^{t_o+L} g(l) dl}{t_o} \]

The expected length of a cycle is

\[ R + \int_{t_o}^{t_o+L} f(t) dt + \frac{\int_{t_o}^{t_o+L} g(l) dl}{t_o} + \int_{t_o}^{t_o+L} g(l) dl f(t) dt \]

\[ = \int_{t_o}^{t_o+L} F(t) dt + K(t_o, t_o) \]

where the last term is $K(t_o, t_w)$ but fixing $t_w = t_r$.

The sum of the six cost terms divided by the expected cycle length gives the expected cost rate $J(t_o, t_r, t_w)$.

\[ \frac{cC(t_o, t_r, t_w) + bB(t_o, t_r, t_w) + aA(t_o, t_r, t_w) + hH(t_o, t_r) + kK(t_o, t_w)}{\int_{t_o}^{t_o+L} F(t) dt + K(t_o, t_o)} \]

We delay until the next section a discussion of the properties of this equation.

\section*{2.10 Single Spare, Two Random Lead Times}

This section combines the results of the previous two sections, in that unscheduled orders may be more expensive and faster on average than scheduled orders and that both of these lead times may be random. This is the most general single-spare model that we
consider in this chapter, and is similar to the ones considered in [KH81], [PP86a], and [PP86b]. To run such a system requires the use of a double age maintenance policy \((t_r, t_w)\) combined with a double age ordering policy \((t_x, t_o)\). Thus the system will use a four number policy \((t_x, t_o, t_r, t_w)\), where \(0 \leq t_x \leq t_o \leq t_r \leq t_w \leq \infty\).

For this model the expected costs per cycle of replacement, breakage, operating, and holding are the same as those in the previous section. The expected ordering cost is

\[
c_x \int_0^t u(t) f(t) \, dt \tag{2-43}
\]

The expected shortage time per cycle changes to

\[
K(t_x, t_o, t_w) = R + L_x F(t_x) + \int_{t_o}^{t_x} \left[ t_o + L - t \right] f(t) \, dt \\
+ \int_{t_o}^{t_x} \int_{t_o}^{t_x} \left[ t_o + L - t \right] g(l) \, dF(t_x) \, dt + \int_{t_o}^{t_w} \left[ t_o + L - t_w \right] g(l) \, dl \\
= R + \int_{t_o}^{t_x} F(t) \, dt + L F(t_o) - \left[ t_o + L - t_x - L_x \right] F(t_x) \\
+ \int_{t_o}^{t_x} \int_{t_o}^{t_x} \overline{G}(l) \, dF(t) \, dt + \overline{F}(t_w) \int_{t_o}^{t_w} \overline{G}(l) \, dl
\]

The average cycle time is simply

\[
\int_{0}^{t} \overline{F}(t) \, dt + K(t_o, t_r, t_r) \tag{2-45}
\]

where the last term is \(K(t_x, t_o, t_w)\) but fixing \(t_w = t_r\).

Combining these yields the cost function \(JTR(t_x, t_o, t_r, t_w)\) shown below.
\[ cC(t_x, t_r, t_w) + bB(t_o, t_r, t_w) + aA(t_o, t_r, t_w) + hH(t_o, t_r) + kK(t_x, t_o, t_w) + c_x \int_0^t u(t)f(t)dt \]

\[ \int_0^t \overline{F}(t)dt + K(t_x, t_o, t_r) \]

Note that setting \( L_x = L \), \( t_x = t_o \), and \( c_x = 0 \) transforms \( JTR \) back into the \( JOR \) equation of the previous section. This means that both equations have the properties described below.

**Theorem 2-9:** \( JTR(t_x, t_o, t_r, t_w) \) is pseudo-convex in each of \( t_x \), \( t_r \), and \( t_w \); thus there exists a unique optimum for each variable (given values for the others) as described below:

(a) For all \( t_x \), \( t_o \), and \( t_r \), the optimal \( t_w \) is:
   (i) \( t_w = t_r \) if \( cv(t_r) + bz(t_r) + aq(t_r) - k \geq 0 \);
   (ii) \( t_r \leq t_w \leq \infty \) if \( cv(t_w) + bz(t_w) + aq(t_w) - k = 0 \);
   (iii) \( t_w = \infty \) if \( cv(\infty) + bz(\infty) + aq(\infty) - k \leq 0 \).

Case (iii) implies that the machine should not be shut down unless a spare is available to perform an immediate replacement. Case (i) means we shut down the machine when the scheduled replacement age is reached, regardless of whether a spare is available. It is worthwhile noting that the expressions above do not include the other decision variables, nor do they include such other factors as the lead time distribution \( G(l) \) or mean repair time \( R \); this indicates that the optimal wear-out age is independent of all of these other parameters. Later in Chapter 3 we will encounter this same property in another context;

(b) Given \( t_x \), \( t_o \), and \( t_w \), the optimal \( t_r \) is:
   (i) \( t_r = t_o + L \) if \( cv(t_o + L) + bz(t_o + L) + aq(t_o + L) + h - JTR(t_x, t_o, t_o + L, t_w) \geq 0 \);
   (ii) \( t_o + L \leq t_r \leq \infty \) if \( cv(t_r) + bz(t_r) + aq(t_r) + h - JTR(t_x, t_o, t_r, t_w) = 0 \);
   (iii) \( t_r = t_w \) if \( cv(t_w) + bz(t_w) + aq(t_w) + h - JTR(t_x, t_o, t_w, t_w) \leq 0 \).

Case (i) implies that a spare should be installed upon receipt, while (iii) implies that replacements should be performed only at failure.
(c) Given \( t_o, t_r, \) and \( t_w, \) the optimal \( t_x \) is:

(i) \( t_x = 0 \) if \( \frac{c_x u(0)}{(t_o + L) - (L_x)} - k + JTR(0, t_o, t_r, t_w) \geq 0 \)

(ii) \( 0 \leq t_x \leq t_o \) if \( \frac{c_x u(t_x)}{(t_o + L) - (t_x + L_x)} - k + JTR(t_x, t_o, t_r, t_w) = 0 \)

(iii) \( t_x = t_o \) if \( \frac{c_x u(t_o)}{L - L_x} - k + JTR(t_o, t_o, t_r, t_w) \leq 0 \)

Case (i) implies that unscheduled orders should never be placed, while (iii) implies that a spare should always be ordered as soon as the installed machine fails.

We have been unable to derive an equivalent optimality result for \( t_o; \) we conjecture that the behaviour of \( JTR \) with respect to \( t_o \) depends on the particular choice of lead time distribution \( G(l). \) This means that in practice we need to make careful use of non-linear programming techniques to search for a good value for the ordering time, as there may be multiple local minima.

Given the results above, one can use a procedure like the following to determine a good and hopefully optimal set of parameters \((t_x, t_o, t_r, t_w).\)

Step 1: Initially set \( t_x = t_o = t_r = t_w = 0; \)

Step 2: Find the optimal \( t_w \) via NLP line search methods, using Theorem 2-9 to verify optimality;

Step 3: Find a \( t_r \leq t_w \) via NLP line search methods, using Theorem 2-9 to verify optimality (given the other current values of \( t_x, t_o \) and \( t_w );\)

Step 4: Find a \( t_o \leq t_r - L \) via NLP line search methods (this is more difficult, as \( JTR \) may not be pseudo-convex in \( t_o, \) i.e. there may be multiple local minima);

Step 5: Find a \( t_x \leq t_o \) via NLP line search methods, using Theorem 2-9 to verify optimality (given the other current values of \( t_o, t_r \) and \( t_w );\)

Step 6: Calculate \( JTR(t_x, t_o, t_r, t_w); \)
Step 7: Repeat steps 3 through 6 to find improved solutions, until satisfied that no material improvement in $JTR(t_x, t_o, t_r, t_w)$ is possible (using standard stopping rules for non-linear search).

Instead of using this coordinate search procedure, after finding $t_w$ in step 2 one can use standard NLP algorithms to search the interval $[0, t_w]$ for good values of $t_x, t_o$ and $t_r$.

Once a solution has been found, we can say something about how it would change if a cost coefficient were to be changed.

**Proposition 2-7:** A change in the cost coefficients will tend to change the optimal solution as described below:

(a) the optimal wear-out age $t_w$ is non-decreasing in the shortage cost rate $k$, and non-increasing in the operating cost rate $a$, the breakage cost $b$, and the replacement cost $c$. It is constant in the holding cost rate $h$ and the expediting cost $c_x$;

(b) for given $t_x, t_o$ and $t_w$, the optimal replacement age $t_r$ is non-decreasing in $h$, and non-increasing in $k$ and $c_x$; and,

(c) for given $t_x, t_o$ and $t_w$, the optimal expediting age $t_x$ is non-decreasing in $k$, and non-increasing in $c, b, a, h$, and $c_x$.

The work in this section generalizes the results of [PP86b], [PP86a], and [KH81], except that [PP86a] makes different use of minimal repairs in its system. To some extent it also generalizes some of our work in previous sections, as illustrated in Figure 2-4 where the different models are listed in decreasing generality.

**Figure 2-4:** Relations Amongst Single Spare Models

```
Two Random Leadtimes (2.10)

<table>
<thead>
<tr>
<th>One Random Leadtime (2.09) Two Deterministic Leadtimes (2.08)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Deterministic Leadtime (2.04) &amp; (2.07)</td>
</tr>
</tbody>
</table>
```
2.11 Multiple Spares without Stock-outs

In this section we allow our system to order and hold multiple spares, but without the full complexity of the general problem that we introduced at the beginning of the chapter. To our knowledge this multiple spare model has not been investigated before in the context of continuous review ordering policies, though Falkner considered similar issues in the context of a periodic review stocking policy; see [Fa68], [Fa69], and [Fa70].

We avoid the interaction problem between maintenance demand and inventory supply by considering two special cases: the borrowed spare case and the instantly expedited delivery case. Each special case eliminates the possibility of a stock-out, which means that inventory deficiencies will not stop the maintenance demand process. We revert in this section to our earlier assumptions that the lead time is deterministic and identical for both scheduled and unscheduled orders, that replacements are instantaneous ($R = 0$), and that the only costs are simply replacement $c$, breakage $b$, operating $aq(t)$, holding $h$, and shortage $k$.

Suppose a point in time has been reached at which the maintenance manager wants (at $t_r$) or needs (on failure) to perform a replacement, but there is no spare in stock. It is sometimes the case that a spare can be obtained by borrowing it from an alternative source; for example, it is not unheard of for airlines which are not direct competitors to loan spare parts back and forth amongst themselves, so as to avoid canceling flights due to shortages. This works as follows. To perform a replacement with no spare in stock, borrow a spare from an alternative source and immediately install it in our system. When our spare arrives from the regular source, give it to the alternative source; in the meantime, we incur a penalty cost at rate $k$ as a rental fee to the alternative source for the privilege of borrowing.

In our analysis, this borrowing arrangement is modeled as follows. Since a spare can be borrowed to avoid stock-out we will assume that maintenance uses a single age policy. Note that this is not necessarily optimal, as it might be better to delay a preventive
replacement instead of paying a premium rate to borrow a spare; but this would indicate
the use of a double age maintenance policy which creates the kind of maintenance-
inventory interaction that we are trying to avoid. Given the use of a single age
maintenance policy, we can use the delayed \((s-1, s)\) inventory policy developed by Atkins
and Katircioglu [AK95] for IFR demand processes. This gives the following equation for
the expected cost rate, \(CB(t_o, s, t_r)\).

\[
(2-47)
\]

\[
C + bF(t_r) + a \int_0^t q(x)F(x)dx + \frac{\int_0^t \Phi_{t_r,t_o,s}(x)dx}{1 - \Phi_{t_r,t_o,s}(t)} + F(t_o) \left[ k \int_0^L \Phi_{t_r,t_0,s-1}(x)dx + h \int_0^\infty \Phi_{t_r,t_0,s}(x)dx \right] + \int_0^t F(x)dx
\]

The first three terms in the numerator are the replacement, breakage, and operating cost
per cycle or machine-life. The first square-bracketed term gives the holding and shortage
costs of the policy assuming that the currently installed machine lasts at least until we
place the next order, while the second bracketed term gives those costs assuming that the
installed machine fails before the next order is due to be placed. Here \(\Phi_{t_r,t_0,s}(t)\)
represents the distribution function for the time until the \(s\)-th replacement; that is, the sum
of \(s-1\) lifetimes of future machines each with distribution \(F(t)\) truncated at \(t_r\), plus the
length of the current cycle which would have the same distribution except that we delay
ordering by a time \(t_o\). Thus \(\Phi_{t_r,t_0,s}(t)\) is an \(s\)-fold convolution. Similarly \(\phi_{t_r,t_0,s}(t)\)
represents the corresponding density function. The decision problem is to find values for
\(t_o, s,\) and \(t_r\) that minimize the expected cost rate for the system. To help with part of this,
we can adapt the following result from [AK95].
Inventory Solution Procedure from [AK95]:

Find \( t_o \) and \( s \) by the following procedure:

(a) first find \( s \):

\[
\Phi_{t_r,0,s}(L) \leq \frac{h}{h+k} < \Phi_{t_r,0,s-1}(L) \tag{2-48}
\]

(b) second, find \( t_o \):

\[
\Phi_{t_r,t_o,s}(L) = \frac{h}{h+k} \tag{2-49}
\]

**Lemma 2-1:** For a given \( t_r \), the \( t_o \) and \( s \) values from the above procedure are optimal in \( CB(t_o, s, t_r) \).

This means that given an age replacement policy \( t_r \) (perhaps sub-optimal) we can at least in principle find an optimal inventory policy to support it.

Instead of borrowing a spare, it may be that in the event of stock-out a spare part can be delivered very rapidly by paying a premium cost. If this delivery time is short enough relative to the machine life so as to be negligible, then we may treat this as an instantaneous delivery. This is similar to the lost sales model often used in other inventory work; in fact, we extend the results from [Sc89] for this version of our model. To account for the difference in stock-out penalties we need to make one change from the previous borrowed spare equation: we now use \( c_x \) to represent the fixed cost premium (not a cost rate) for expediting. This gives the following equation \( CE(t_o, s, t_r) \) for the expected cost rate.

\[
\begin{align*}
&
\left\{ \begin{array}{l}
\begin{aligned}
&
c + bF(t_r) + a \int_0^{t_r} q(x) \overline{F}(x) \, dx + \overline{F}(t_o) \left[ c_x \, \Phi_{t_r,t_o,s}(L) + h \int_L^{t_r} \Phi_{t_r,0,s-1}(x) \, dx \right] \\
&
\end{aligned}
\end{array}
\right. \\
&
\begin{aligned}
&
+ F(t_o) \left[ c_x \, \Phi_{t_r,0,s-1}(L) + h \int_L^{t_r} \Phi_{t_r,0,s-1}(x) \, dx \right]
\end{aligned}
\right. \\
&
\int_0^{t_o} \overline{F}(x) \, dx
\end{align*}
\]
As with the borrowed spare case, we would like to characterize the optimal inventory policy for this instantly expedited case. We conjecture that this could be done in a manner such as the following. Given \( t_r \), find optimal \( t_0 \) and \( s \) by the following procedure:

(a) first find \( s \):
\[
\frac{\phi_{t_r,0,s}(L)}{\Phi_{t_r,0,s}(L)} \leq \frac{h}{c_x} < \frac{\phi_{t_r,0,s-1}(L)}{\Phi_{t_r,0,s-1}(L)}
\]

(b) second, find \( t_0 \):
\[
\frac{\phi_{t_r,t_0,s}(L)}{\Phi_{t_r,t_0,s}(L)} = \frac{h}{c_x}
\]

Note that whereas in the borrowed spare case we compared a cost ratio to a distribution function, here we compare a cost ratio to a hazard function. Schultz in [Sc89] proved a result equivalent to part (b) for the special case where \( s = 1 \), but we have been unable to prove our conjecture for the general case of any \( s \).

Finding the actual solution in an application of either the borrowed spare case or the expedited spare case is unfortunately not a trivial exercise, as the convolutions of truncated lifetime distributions are not easy to calculate either analytically or numerically. We can improve things somewhat if we restrict our maintenance policy to replacement at failure, so that \( t_r = \infty \). Aside from making our maintenance policy trivial, this assumption also means that the machine lifetime distributions are untruncated, so that at least for some choices of distribution functions (e.g. the normal distribution) the convolutions become much more tractable.

With \( t_r = \infty \) the equation \( CB(t_0, s, t_r) \) for the borrowed spare case would become \( CB(t_0, s, \infty) \) as follows.
Lemma 2-1 still applies to this simpler equation.

As an example, let \( c = $10, b = $20, a = $0, k = $100, h = $2, L = 0.05, \) and \( F(t) = 1 - e^{-t^2} \). Then the critical ratio \( h / [h + k] = 0.02 \), which using (2-48) and (2-49) leads to optimal values \( s = 1 \) and \( t_o = 0.22 \). Use these with (2-53) to find \( CB(0.22, 1, \infty) = $35. \)

### 2.12 The General Problem

We now briefly return to the general problem which allows the lead times for receipt of spares to be random (with the distribution possibly dependent on the type of order) and the number of spares in stock or on order is unconstrained. This general problem appears to be intractable, in part because of the maintenance-inventory interaction. We conjecture that an optimal policy form would have four critical parameters: \( s, t_o, t_r, \) and \( t_w \). Maintenance would use a double age replacement policy, whereby it would shut down the installed machine when its age was greater than \( t_r \) if a spare was in stock, or in any event when the age reached \( t_w \). Inventory would use a delayed \((s-1,s)\) policy, where \( s \) would be the target inventory level, and new orders would be placed after a delay of \( t_o \) or at the time of the next demand, whichever occurred first. The complication is that optimal \( t_o \) and \( t_r \) would need to be dynamic functions of the inventory-on-order (represented by a vector of demand times) rather than static numbers.

One way to approach this problem would be to define some bounds on the minimum cost solution, and then within these bounds use heuristics to search for a good solution for a particular problem instance. We could start by trying the solution methods
given in the previous sections for more restricted versions (e.g. single spare) of the problem. Any of these which have the appropriate structural assumptions (e.g. lead time being deterministic or random) will give feasible policies with known cost rates and hence can provide candidate solutions and act as upper bounds on the unknown optimum.

For example, the following could be used to start the solution procedure in applications where the lead time is deterministic:

Step 1: restrict the system to a single spare with a non-negative ordering time, and then develop an initial solution by optimizing $JOD(t_o, t_r)$;

Step 2: if the calculated ordering time $t_o > 0$, then this indicates that a single spare is in fact optimal, so stop with the optimal solution;

Step 3: if $t_o = 0$, then next restrict the maintenance policy to replacement on receipt and optimize $JED(t_o)$. Compare this cost rate to that from step 1 and use whichever policy gives the best results.

2.13 Future Research

Obviously many applications of the joint maintenance and inventory optimization problem will require policies of a more general form than those contained herein, i.e. most real inventory systems must handle multiple spares with the possibility of suffering stock-outs. The general form of the problem unfortunately appears to be intractable, at least for the method of analysis we use in this chapter. Thus we recommend this more difficult but important problem for future work.

Another avenue for further development is the merger of the work dealing with systems of one machine and one spare to handle the entire life-cycle of a machine. Ideally this might result in a model which can deal with the classic bathtub-shaped hazard rate curve, wherein we could return a spare after the initial burn-in phase, take no action during the primary service life, and then order a replacement at some point in the machine's wear-out phase.
2.14 Concluding Remarks

For the general form of the problem we considered, we have only been able to suggest some heuristic solution procedures; exact optimality appears to be intractable to our style of analysis. When we assumed that we were to use a single age maintenance policy and could either borrow a spare or instantly expedite one, then we were able to write down the cost function and specify some of its properties. When we restricted our inventory policy to having a single spare in stock or on order with general lead times, then we were able to specify the optimal policy form, write down the cost function, and determine some of this function's optimality conditions.

When we further restricted the lead times to be deterministic, we were able to derive a full set of optimality conditions. We found that the objective function is pseudo-convex in each decision variable and any economically viable KKT point is a global minimum. We also showed that joint optimization can be worthwhile in comparison to traditional separate optimization. Separate optimization can be arbitrarily bad in extreme cases, while for the numerical data we considered, joint optimization gave an average improvement of 3% over the best form of sequential optimization with one tenth of the cases having a difference greater than 10%.

Overall, our work enhances the existing single-spare optimization literature by generalizing and combining the work of previous authors, and by introducing the notions of breakage costs, return of spares, service constraints, and double age policies.
Chapter 3
The Machine Repairman Problem with Preventive Maintenance

3.01 Chapter Abstract

Although the machine repairman problem has been a subject of research for quite some time, it seems that all existing studies have made the reasonable but restrictive assumption that repairs are made only after machine failure, i.e. preventive maintenance is ignored. Our work extends the machine repairman model by allowing for preventive repair under the assumption that the machines have an increasing hazard function. We derive analytical results for the case where we restrict the form of the repair policy to a single critical age, and then investigate other heuristic policies which can improve the system's cost effectiveness. We conclude by comparing the performance of different policies using two numerical studies.

3.02 Introduction

Consider a system consisting of \( n \) repair shops which support \( m \) machines subject to stochastic failure. Whenever a machine breaks down, it is repaired by a repair shop; each repair keeps a shop busy for a random amount of time, during which it cannot service other broken machines. The problem of efficiently operating such a system is commonly referred to as the machine repairman problem (MRP), or more generally the machine interference problem. This type of problem can arise not only in maintenance and repair situations, but also in situations where workers tend several automatic machines or a computer processes jobs from different input sources.

In our system, we will also allow the system manager to have the option of shutting down an old but still operational machine before it breaks, so as to effect a preventive repair. To take advantage of this ability, the manager must decide when to best perform preventive shut downs. This decision is typically based upon the analysis of a single machine as described in e.g. Barlow & Proschan [BP75] or Jorgenson et al [JMR67], which would prescribe a single critical age for the preventive repair policy.
Such a policy, while optimal for a single machine, is not in general optimal for a system of many machines sharing common repair facilities; this is the issue that we explore herein.

There is a fairly extensive literature regarding the MRP. The work by Cho & Parlar [CP91] provides a review of multiple machine reliability and maintenance in general, while Stecke & Aronson [SA85] provide a comprehensive review of the machine interference literature in particular. Agnihothri [Ag89] considers the relationships amongst system parameters that hold true for any MRP. Most of the other existing MRP papers (e.g. [Cr75], [Go77], [Wi77], [GI81], [Le84]) examine the performance characteristics of a system under a variety of assumptions regarding failure distributions, number of repair shops, costs incurred, etc., but seem to always assume that repair will occur only after failure i.e. that there is no preventive maintenance. One exception is the work by Albright [Al80], which allows for some preventive maintenance in that the rate of failure for the machines can be reduced by paying a cost which represents oiling, tuning, and other minor adjustments which do not take the machine out of service. We also recently became aware of some work by Moinzadeh & Berke [MB95] which studies a system that appears quite similar to ours, except that the machines in their system get less productive as they age but never actually fail; hence their mathematical model and method of analysis are quite different from ours. They are essentially working on an enhancement to what is sometimes referred to as a tool wear model; see e.g. [JY92]. We do not discuss the kind of economic interdependence (versus stochastic interdependence) amongst machines that leads to what are often called opportunistic maintenance models; this topic has a large body of research of its own (for an example, see [Oz88]).

In our work, we make use of the $GI/M/n$ finite population queuing results developed for the MRP by Bunday and Scraton [BS80] and later generalized by Sztrik in [Sz85] and [Sz88]. Bunday and Scraton showed that the steady state probabilities for the number of machines operating could be calculated using only the means of the failure and
repair distributions. Although they did not explicitly consider it, their formulae can be used to analyze certain kinds of preventive maintenance policies.

This work is motivated in part by the author's previous experience in aircraft maintenance. At his place of work, each of a group of aircraft underwent a periodic inspection by a single maintenance shop after having flown a predetermined number of hours. The rate of flying was sufficiently high that the shop was most often busy; whenever one aircraft had been completed, the next was generally ready to be brought in. However, it occasionally happened that the next aircraft still had a considerable number of hours left in its rated life; this prompted the question of whether it was best to bring it in anyway, or to leave the shop empty until this life was exhausted. Conversely, it often happened that an airplane would reach its rated life before the shop was finished with the previous one; the question then was whether to allow it to fly some more or to require it to sit idle until the shop was ready for it. It was never clear in either situation what the optimal decisions might be.

We begin our discussion by describing more precisely our continuous time system model. We then examine single age repair policies and their properties, followed by more general policies; we shall see that single age repair is tractable to analyze, but is generally sub-optimal. We next compare the performance of all of these policies in a numerical study using discrete time systems; we switch to discrete time here because it allows us to readily calculate the optimal policies for comparison purposes. There follows a second numerical study, which uses simulation to evaluate some of our heuristics in the continuous time model. We conclude with a summary of our results and a discussion of desirable extensions to our work. Proofs for our results are given in Appendix C.

### 3.03 Basic Model and Notation

Our continuous time system consists of \( m \) identical machines serviced by \( n \) identical repair shops, \( n \leq m \). The operating life of each machine has distribution \( F(t) \), density \( f(t) \), and hazard function \( z(t) \). We assume that \( z(t) \) is strictly increasing in \( t \), so that
the machines have a strictly increasing failure rate (are IFR). A machine can be repaired either before (preventive repair) or after (corrective repair) it has failed; for both forms of repair, the repair duration has a negative exponential distribution $\Gamma(t)$ with density $\gamma(t)$ and mean $R$. Repairs are assumed to be perfect, and any shop can fix any machine.

There are four costs charged to our system. A shortage cost is incurred at rate $k$ per unit time per machine unavailable (while waiting for or undergoing repair). We incur a repair cost $c$ every time we perform a repair (corrective or preventive), and in addition pay a breakage cost $b$ for every machine which fails in use (i.e. this is a cost premium for corrective repair). While a machine is running, it incurs an operating cost at rate $a q(t)$, where $a$ is a cost coefficient and $q(t)$ is non-negative non-decreasing function of the machine age; this cost can represent such things as e.g. increased fuel consumption, quality losses, and some types of minimal repair. The sunk cost of having the repair shops is common to all policies and so is ignored in this analysis. Our objective is to determine a preventive maintenance policy which will minimize the long run expected cost $C(policy)$ of operating the system.

We may think of this system as a finite population $n$ server queue in which service times are independent and identically distributed (iid) markovian. The arrivals to this queue are iid IFR if the decision to do preventive maintenance is based only on the age of that particular machine; but in general we will want to take into account the state of the whole system in making such a decision, so that arrivals will not necessarily be independent of each other or of service times.

More generally, we may describe our system as a simple form of closed queuing network (CQN). This particular CQN has two queuing nodes (sometimes referred to as stations): an $n$-server repair node whose service times are iid negative exponential, and an $m$-server "running" node whose service times are IFR and depend on the maintenance policy chosen. A fixed number $m$ of machines circulates between these two nodes; whenever a machine finishes service at one node, it next enters the queue at the other
node. Since the machines are identical and repairs have a negative exponential distribution, we may assume that the queuing discipline for each station is either first-come first-serve (FCFS) or last-come first-serve preempt-resume (LCFS.PR) according to our needs; we could also treat the running node as having an infinite number of servers, if that were useful. Our goal is to find a maintenance policy which leads to a minimum cost steady state for this CQN. See Figure 3-1 for an illustration of such a CQN.

**Figure 3-1:** A CQN with four machines and two repair shops.

CQN like this but with all queuing stations having independent negative exponential service distributions have been extensively studied, as in e.g. [SY87] and [SY88]. Unfortunately our machine running times are not negative exponential and need not even be independent, so there seem to be few previous results from the CQN literature that apply to this problem.

### 3.04 The Simple Single Age Repair Policy

The most straightforward form of preventive maintenance policy is an age repair policy. Let us assume for the moment that the system consists of only one machine plus one repair shop. We will specify a single critical age $t_j$, such that if the machine reaches age $t_j$ without having failed it will be shut down at that time for preventive maintenance. We can easily model this small system as a renewal process, in that whenever the machine
finishes repair it is as good as new and the system is renewed. Accordingly, we can calculate the cost rate of a policy by taking the expected cost per renewal period and dividing by the expected length of a renewal period (as done in e.g. [BP75] and [JMR67]) to create the equation below.

\[
C_1(t) = \frac{c + kR + bF(t) + a \int_0^t q(x) \overline{F}(x) dx}{R + \int_0^t \overline{F}(x) dx}
\]  

Here we have the sum of the repair cost, the expected shortage cost per cycle, the expected failure cost, and the expected operating cost, divided by the sum of the mean time to repair plus the expected machine life.

**Lemma 3-1:** \( C_1(t) \) is pseudo-convex in \( t \), thus it will have a unique minimum at \( t_1, 0 \leq t_1 \leq \infty \) as follows:

(a) \( t_1 = 0 \) if \( bz(0) + aq(0) - C_1(0) \geq 0 \);

(b) \( 0 \leq t_1 \leq \infty \) if \( bz(t_1) + aq(t_1) - C_1(t_1) = 0 \);

(c) \( t_1 = \infty \) if \( bz(\infty) + aq(\infty) - C_1(\infty) \leq 0 \).

Case (a) implies that the system is not viable and therefore should not operate, while case (c) means that preventive replacement should never be performed.

This critical age policy is clearly optimal for a single machine or whenever \( n \geq m \), and would be the policy used if one were to naively apply the age repair principles discussed in e.g. [BP75] for a single machine to a system where several machines share a smaller number of repair shops. Suppose we were to calculate a \( t_1 \) from the above formula and use it in a system of \( m \) machines. We might naively expect that the cost rate would be \( mC_1(t_1) \), but it will generally be higher than this due to interference at the repair shops; i.e., a machine will sometimes be idle waiting for repair because all the shops are busy. The rate given by \( mC_1(t_1) \) is instead a lower bound for the actual cost rate resulting
from the use of $t_j$ (or indeed of any feasible policy). In the next section we show how to determine the actual cost of single age policies for systems of multiple machines.

### 3.05 The Joint Single Age Repair Policy

If there are more machines than shops in the system, it is harder to derive a cost rate equation based upon renewal concepts directly, so we take a different approach. Bundy and Scraton in [BS80] derived a set of integro-differential equations which they solved to find the steady state distribution of the system, i.e. the probability that $i$ of $m$ machines in a system are inoperative. For an age $t$ repair policy, the probability of having $i$ machines inoperative is (using [BS80])

\[ \text{prob}_i(t) = \left[ \frac{m!}{(m-i)!i!} \right] R^i \left[ \int_0^t \overline{F}(x) \, dx \right]^{i} \text{prob}_0(t) \quad \text{for } 0 \leq i \leq n \quad (3-2) \]

\[ \text{prob}_i(t) = \left[ \frac{m!}{(m-i)!n!n^{i-n}} \right] R^i \left[ \int_0^t \overline{F}(x) \, dx \right]^{i} \text{prob}_0(t) \quad \text{for } n \leq i \leq m \quad (3-3) \]

where \( \sum_{i=0}^{m} \text{prob}_i(t) = 1 \) \quad (3-4)

which after solving for $\text{prob}_0(t)$ become respectively

\[ \text{prob}_i(t) = \frac{m!}{(m-i)!i!} \left[ \int_0^t \overline{F}(x) \, dx \right]^{i} \sum_{j=0}^{n} \left\{ \frac{m!}{(m-j)!j!} R^j \left[ \int_0^t \overline{F}(x) \, dx \right]^{-j} \right\} + \sum_{j=n+1}^{m} \left\{ \frac{m!}{(m-j)!n!n^{j-n}} R^j \left[ \int_0^t \overline{F}(x) \, dx \right]^{-j} \right\} \]

for $0 \leq i \leq n$
\[
\text{prob}_i(t) = \frac{m!}{(m-i)! \; i! \; n! \; n^t} \left( \int_0^t F(x) \, dx \right)^{-i} \left( \int_0^t \bar{F}(x) \, dx \right)^{n-i} \\
\sum_{j=0}^{n-i} \frac{m!}{(m-j)! \; j! \; n!} \left( \int_0^t \bar{F}(x) \, dx \right)^{j-i} + \sum_{j=n+1}^{m} \frac{m!}{(m-j)! \; n! \; n^t} \left( \int_0^t \bar{F}(x) \, dx \right)^{n-j-i}
\]

for \( n \leq i \leq m \)

while the cost rate \( cost_i(t) \) for each case \( i \) is

\[
\text{cost}_i(t) = ik + [m - i] \text{rate}(t)
\]

where \( \text{rate}(t) = \left[ c + bF(t) + a \int_0^t q(x) \bar{F}(x) \, dx \right] \left[ \int_0^t \bar{F}(x) \, dx \right]^{-1} \)

is the cost rate per machine running. Taking the summation of the products of costs and their probabilities gives us an equation for our expected cost rate:

\[
C_m(t) = \sum_{i=0}^{m} \text{cost}_i(t) \text{prob}_i(t) = \sum_{i=0}^{m} \left[ ik + (m - i) \text{rate}(t) \right] \text{prob}_i(t) \quad (3-7)
\]

This reduces to (3-1) when \( m = 1 \) and \( n = 1 \). We can also re-arrange terms to get

\[
C_m(t) = \text{rate}(t) \left[ m - \sum_{i=0}^{m} i \ \text{prob}_i(t) \right] + k \sum_{i=0}^{m} i \ \text{prob}_i(t) \quad (3-9)
\]

\[
= \text{rate}(t) E[MU(t)] + k E[MD(t)]
\]

where \( MU(t) \) is the number of machines up (running) and \( MD(t) \) is the number of machines down (not running). It will later be useful to have this cost rate in the form of a fraction.

Define \( y_i \) as

\[
y_i = \frac{m!}{(m-i)! \; i! \; n! \; n^t} \left( \int_0^t \bar{F}(x) \, dx \right)^{-i} \quad \text{for } 0 \leq i \leq n \quad (3-10)
\]

\[
y_i = \frac{m!}{(m-i)! \; n! \; n^{t-n}} \left( \int_0^t \bar{F}(x) \, dx \right)^{-i} \quad \text{for } n \leq i \leq m \quad (3-11)
\]
so that it represents the numerator of (3-5) or (3-6) depending on the value of $i$. Then

$$C_m(t) = \frac{N_m}{D_m} = \frac{\sum_{i=0}^{m} [ik + (m-i)rate(t)]y_i}{\sum_{j=0}^{m} y_j} \quad (3-12)$$

We can use any of (3-8), (3-9), or (3-12) to calculate the actual cost rate resulting from the use of $t_j$ in our $m$ machine system. They can also be used to develop a better age repair policy, which we call joint single age repair. We will specify a single age $t_m$, such that any machine that reaches age $t_m$ (i.e. without having already failed) will be shut down for preventive maintenance. The ages $t_m$ and $t_j$ are generally not equal, as $t_j$ is obtained by optimizing (3-1), while $t_m$ results from optimizing (3-8). This latter equation has several useful properties.

**Proposition 3-1:** An interior minimum (i.e. $\partial C_m(t_m) / \partial t = 0$) occurs at $t_m$ such that

$$C_m(t_m) = C_{m-1}(t_m) + bz(t_m) + aq(t_m) \quad \text{where } C_0 = 0 \quad (3-13)$$

or equivalently,

$$\frac{[bz(t_m) + aq(t_m) - rate(t_m)]}{[k - rate(t_m)]} = \frac{\text{Var} [MU(t_m)]}{E[MU(t_m)]} \quad (3-14)$$

**Lemma 3-2:** $C_2(t)$ is pseudo-convex in $t$ with a unique minimum at $t_2$ as follows:

(a) $t_2 = t_j$ if $C_j(t_j) + bz(t_j) + aq(t_j) - C_2(t_j) \geq 0$;

(b) $t_j \leq t_2 \leq \infty$ if $C_j(t_2) + bz(t_2) + aq(t_2) - C_2(t_2) = 0$;

(c) $t_2 = \infty$ if $C_j(\infty) + bz(\infty) + aq(\infty) - C_2(\infty) \leq 0$.

**Proposition 3-2:** For any given $t$, the expected number of machines down is increasing in the number of machines $m$ at a non-decreasing rate, i.e. we have

(a) $E[MD(t)_{m+1}] > E[MD(t)_m]$; and,

(b) $\{E[MD(t)_{m+2}] - E[MD(t)_{m+1}]\} > \{E[MD(t)_{m+1}] - E[MD(t)_m]\}$. 
Theorem 3-1: Over its viable range, i.e. for all \( t: C_m(t) \leq mk \), we have that \( C_m(t) \) is pseudo-convex in \( t \) with a global minimum at \( t_m \) as follows:

\[
(a) \quad t_m = t_{m-1} \quad \text{if} \quad C_{m-1}(t_{m-1}) + bz(t_{m-1}) + aq(t_{m-1}) - C_m(t_{m-1}) \geq 0 ;
\]

\[
(b) \quad t_{m-1} \leq t_m \leq \infty \quad \text{if} \quad C_{m-1}(t_m) + bz(t_m) + aq(t_m) - C_m(t_m) = 0 ;
\]

\[
(c) \quad t_m = \infty \quad \text{if} \quad C_{m-1}(\infty) + bz(\infty) + aq(\infty) - C_m(\infty) \leq 0 .
\]

Case (c) above implies that preventive maintenance should not be performed.

Proposition 3-3: \( C_m(t) \) and its optimal age \( t_m \) have the following properties:

(a) the optimal age \( t_m \) is increasing in \( k \) and \( c \), and decreasing in \( b \) and \( a \);

(b) both the expected number of machines down (in or awaiting repair) and the expected number of machines idle (awaiting repair) are increasing in \( R \) and decreasing in \( t \) and \( n \); and,

(c) for any given \( t \), the ratio \( C_m(t) / mC_1(t) \) is increasing in \( m \), decreasing in \( n \).

To actually calculate \( t_m \) we can use non-linear programming (NLP) methods to search \( C_m(t) \) over the interval \([0, \infty]\), using Proposition 3-1 to identify the optimal solution. If we happened to know the optimal ages for smaller or larger systems, for example \( t_{m-1} \) and \( t_{m+1} \), then we could narrow down the search interval considerably since \( t_{m-1} \leq t_m \leq t_{m+1} \).

The use of \( t_m \) from the above analysis rather than \( t_j \) from traditional maintenance procedures reduces our operating cost rate without an increase in policy complexity (we discuss later the size of this reduction). In fact, within the class of single age policies \( t_m \) gives an optimal policy. As we shall see in the following sections, to reduce the cost rate further requires the use of more complicated policies.

3.06 Better Heuristic Policies

So far the decision to shut down a working machine for preventive maintenance has been based only upon the age of that particular machine. Next we examine policies which consider other state information. In doing so, it may be helpful to illustrate the policy forms in a diagram which represents a system of two machines, in which each axis
indicates the age of one of the machines. Our MRP then consists of finding a boundary for this diagram to indicate the feasible area for our policy, such that whenever we contact the boundary we shut down the oldest machine. For example, Figure 3-2 shows the shape of a single age policy.

**Figure 3-2: Single Age Policy**

The current availability of the repair shops is probably the most obvious additional information to consider. With only one machine, whenever we shut down, a shop is immediately available; but this is not always true with multiple machines. We expect to get interference, in that the machines may be delayed in a queue waiting to be repaired by a shop; thus we might specify a maintenance policy that explicitly considers whether the shops are busy in deciding whether to shut down a machine. Our goal in this is to avoid shutting down a machine if it is just going to sit idle awaiting service. The following is useful in determining such a policy.

**Theorem 3-2:** Whenever all the repair shops are busy, it is optimal to shut down a functioning machine as soon as its age equals the *wear-out age* $t_w$, and not before. The optimal wear-out age $t_w$ can be obtained as follows:

(a) $t_w = 0$ if $bz(0) + aq(0) - k > 0$;

(b) $0 \leq t_w \leq \infty$ if $bz(t_w) + aq(t_w) - k = 0$;

(c) $t_w = \infty$ if $bz(\infty) + aq(\infty) - k \leq 0$. 
Case (a) implies that the system is not viable, while case (c) means that we should always wait until a shop is available before shutting down. Note that searching for $t_w$ is fairly straightforward, as both $z(t)$ and $q(t)$ are monotonically increasing in $t$, as well, $t_w$ is independent of $m$, $n$, and $R$. This wear-out age is similar in principle to that discussed in the previous chapter for Theorem 2-9.

A simple way to use this wear-out age is to specify a repair policy which uses age $t_L$ whenever at least one repair shop is available, and the wear-out age $t_w$ otherwise. Thus we get our first heuristic.

**The Double Age Policy:** Shut down the oldest machine if its age is $\geq t_L$ and a shop is available, or if its age is $\geq t_w$ and all shops are busy.

We could also refer to this as a "can-must" policy, in that we can shut down any time after $t_L$ if the shop is free but we must shut down by $t_w$ regardless. Unfortunately, while we can easily determine the optimal value of the upper age limit $t_w$, we do not know of any exact analytical way to find an optimal value for $t_L$, so the use of the double age policy would imply some numerical optimization via dynamic programming or some kind of search using simulation to evaluate candidate ages. Figure 3-3 shows the form of the double age policy for a system of two machines and one repair shop. The boundary extensions near the vertical and horizontal axis apply when one machine has age = 0, i.e. it is either undergoing or just completed a repair.

Note that a single age policy is a special case of the double age policy in which we arbitrarily set $t_w = t_L$. Another special case of the double age policy is obtained when we arbitrarily fix $t_w = \infty$. This gives a slightly simpler heuristic.

**The Wait Until Available (WUA) Policy:** Shut down the oldest machine at the first point that its age is $\geq t_L$ and a repair shop is available.

So if a machine has reached its critical age $t_L$ but the repair shop is busy, wait until the shop is available before shutting down. This seems like a sensible policy that one might
expect to actually see in use in a real system. Figure 3-4 illustrates the shape of this policy.

**Figure 3-3: Double Age Policy**

![Double Age Policy Diagram]

We could of course generalize the double age policy to an $n+I$ age policy $(t_L(i), t_w)$, which uses a shut down age $t_L(i)$ whenever $i$ shops are available, $i=\{1,...,n\}$, and age $t_w$ whenever no shops are available. As noted in [MB95], this form of policy would be more complicated to optimize and implement, and so we do not pursue it further in this chapter.

As an aside, let us suppose that the decision to shut down each machine and send it for repair is not made by a central manager who knows the state of the entire system, but by a local operator who has limited knowledge about the rest of the system. If the
operator has no information beyond the age of that particular machine, then clearly the optimal policy for the local maintenance decision will be the joint single age $t_m$ policy. If it is also known whether or not at least one repair shop is available, then the optimal policy for the local maintenance decision is the double age $(t_L, t_w)$ policy, since this policy takes into account all the information available to the local operator. If the status of each repair shop is also known, then the optimal policy for the local maintenance decision must be an $n+1$ age $(t_L(i), t_w)$ policy, as again it makes use of all the available information in making the shut down decision.

Next we want to look beyond the availability of repair shops to consider the ages of the machines other than the one that we are considering shutting down. In effect we would like to coordinate the future shut down decisions of all the machines in the system. A policy which explicitly considered all of the possible combinations of all machine ages in making a shut down decision would obviously be quite complicated to optimize and to implement. A simpler way to handle these ages is through a single statistic that summarizes the state of the system, such as the mean age.

**The Average Age Policy:** Shut down the oldest machine when the average age of the machines exceeds some chosen value; i.e. when $[\text{mean age}] \geq t_a$.

As with the double age policy, finding the best value of $t_a$ requires numerical optimization. See Figure 3-5 for an illustration of this policy.

The last heuristic policy we consider is based upon the short term cost of machine operation, rather than the machine ages as such. This myopic policy looks ahead at the expected cost over the next short period of time assuming that a preventive shut down is not done; if this cost exceeds some chosen limit, then we shut down a machine now rather than incur this cost. Thus we get the following policy.

**The Constant Cost Policy:** Shut down the oldest machine when the expected short term operating cost exceeds some chosen critical value, i.e. when $[\text{cost of operating for a further period } \Delta t] \geq cost_{max}$.
If we let $\Delta t \to 0$, then the shut down decision will be based upon the instantaneous expected cost rate. Figure 3-6 shows the approximate shape of this policy form.

**Figure 3-5: Average Age Policy**

**Figure 3-6: Constant Cost Policy**

Unfortunately, we are unable to specify analytically the cost of any of these heuristic policies (let alone determine the overall optimum). Thus later in the chapter we turn to numerical studies to approach this problem.

### 3.07 Service Constraints

As was mentioned in Chapter 2, in some situations a system manager may be unable to easily determine values for some of the cost coefficients, particularly the breakage cost $b$ or the shortage cost rate $k$; in this case it might be more appropriate to
impose service constraints rather than specifying explicit costs. This technique is commonly used in inventory modeling where the same problem arises. Thus instead of including a breakage cost in the cost equation, we would require that on average no more than a certain proportion of machines would fail before shut down; instead of the shortage cost, we could require that our system have on average at least a certain number of machines working, at least a certain number of machines working with a specified probability, or at most a certain expected length of downtime per machine per cycle.

Alternatively, one could include both the cost terms and these service constraints in the decision process. Thus we could rephrase our optimization problem as a constrained problem; the example below shows this for a single age policy.

Minimize:

\[
C_m(t) = \sum_{i=0}^{m} [ik + (m - i) \text{rate}(t)] \text{prob}_i(t)
\]  \hspace{1cm} (3-15)

Subject to:

\[
F(t) \leq \text{maximum expected proportion of machines which fail before shut down}
\]  \hspace{1cm} (3-16)

\[
\sum_{i=0}^{m} (m - i) \text{prob}_i(t) \geq \text{minimum required average # machines up}
\]  \hspace{1cm} (3-17)

\[
\sum_{i=0}^{m-U} \text{prob}_i(t) \geq \text{minimum required probability of having at least U machines up} \hspace{1cm} (3-18)
\]

\[
\frac{\sum_{i=0}^{m} i \text{prob}_i(t) \int_{0}^{i} F(x)dx}{\sum_{i=0}^{m} (m - i) \text{prob}_i(t)} \leq \text{maximum average down time per machine per cycle} \hspace{1cm} (3-19)
\]

The addition of such constraints obviously makes the solution process more difficult in general, but for a single age policy only univariate optimization is required so such constraints should not pose any major problems.
3.08 Discrete Time Model

So far in this chapter we have concentrated on deriving analytical results for our system of machines, but we now turn to numerical studies so as to examine the effectiveness of the various policies that we have introduced. For simplicity, we consider systems with a small number of machines for which the operating cost rate and the repair cost are negligible (i.e. $a = 0$, $c = 0$). Our first numerical study uses a discrete time formulation of our system model. Switching to discrete time allows us to numerically solve for the optimal policy, thus providing a standard to compare our heuristics against. It is also an interesting version of the problem in its own right; see [ST80] for an analysis of a similar but simpler model. In the following paragraphs we describe the discrete time model that is related to our original continuous time one; in some sense, each is an approximation of the other.

In this model, machine age is measure in discrete periods of length $\Delta t$, with a maximum age allowed of $T$ such periods; hence each machine can be in any of $(T+1)$ states, and the whole system of $m$ machines has $(T+1)^m$ states in total. The objective, as with the continuous time model, is to minimize the long run expected cost. We must decide at the beginning of each period what action to take during the period, with the results of that action becoming known at the end of the period. If a machine is operating at the beginning of the period and we do not shut it down, then it has a probability $\text{fail}_i$ of failing at the end of the current period, where $i$ indicates the number of periods it has already run and $\text{fail}_{i+1} > \text{fail}_i$ (this models an IFR lifetime distribution); due to our age limit, we always have $\text{fail}_T = 1$. The machine is assumed to operate throughout the current period regardless. If we do shut down an operating machine, or if a machine is already broken before the start of the period, then that machine will be unavailable throughout the period. A repair shop can try to repair one machine each period; there is a constant probability $\text{rep}$ of it being fixed at period's end (this mimics a CFR repair time
distribution). For each machine that breaks at the end of a period, we incur a cost \( \text{broke} \); for each machine not operating in a period, we incur a shortage cost \( \text{short} \).

This model can be formulated as a discrete time finite state infinite horizon dynamic program, and then solved using linear programming to find the optimal solution for the particular \( T, \Delta t \), and cost parameters chosen. We refer the reader to chapters 8 and 9 of [Pu94] for the details of using a linear program in this manner. In dynamic program form, the state of the system is the vector of machine ages; the actions we can take in each state are either "do nothing" (represented below by \( N \)) or "do a preventive shut down" (represented by \( M \)). We give the linear program below for the case of two machines and one repair shop. The formulation for a system of more than two machines would add more subscripts to the terms in an obvious way. A system with more than one repair shop would require more decision variables to represent the additional decisions possible in each period (e.g. \( L \) could represent the action "do a preventive shut down for two machines").

Minimize

\[
\text{CostRate} = \sum_{i,j} \left[ M_{i,j} \text{cost}\left(M_{i,j}\right) + N_{i,j} \text{cost}\left(N_{i,j}\right) \right]
\]  

(3-20)

Subject to

\[
M_{i,j} + N_{i,j} - \sum_{a,b} \left[ \text{Pr}\left[i, j \mid M_{a,b}\right] M_{a,b} + \text{Pr}\left[i, j \mid N_{a,b}\right] N_{a,b} \right] = 0 \quad \forall \ i, j
\]  

(3-21)

\[
\sum_{i,j} \left[ M_{i,j} + N_{i,j} \right] = 1
\]  

(3-22)

\[
M_{i,j} \geq 0, \ N_{i,j} \geq 0, \ M_{i,j} N_{i,j} = 0 \quad \forall \ i, j
\]

where

\( T \) = maximum age of a machine, measured in periods of length \( \Delta t \)

\( ij \) = machine ages at start of period, \( 0 \leq i \leq T, \ 0 \leq j \leq T \)

\( M_{ij} \) = steady state probability of being in state \( ij \) and doing a shut down
\( N_{ij} \) = steady state probability of being in state \( ij \) and not shutting down
\( \text{cost}(A_{ij}) \) = one period cost of being in state \( ij \) and taking action \( A \in \{M, N\} \)
\( Pr[i,j|\text{A}_{ab}] \) = transition probability of ending the period in state \( ij \) given that we started in state \( ab \) and took action \( A \)

Here, (3-20) gives the expected cost of operating the system in the steady state. The set of constraints defined by (3-21) are sometimes referred to as balance constraints, in that they require the rate of transition into any state be equal the rate of transition out of that state. (3-22) simply requires that the steady state probabilities sum to one. Solution of the LP gives the steady state probabilities of being in a state and taking a particular action when using the optimal policy. A decision variable which is positive indicates that that state-action pair is used in the optimal policy, while a decision variable which is zero indicates that that state-action pair is not used. Hence the pattern of positive values indicates the form of the optimal policy.

One advantage of using an LP formulation is that it is easy to add additional constraints to the model. We in fact do this for each of the heuristics, to force the solution to have the appropriate form (e.g. a single shut down age for the single age policies). One disadvantage is the relatively large size of the state space. The full LP has \((T+1)^{n+2}\) rows and \((n+1)(T+1)^m\) columns, though these can each be reduced by about half by combining "symmetric" states together, e.g. machine ages 2,3 and 3,2 are identical in their behaviour and so can be combined into a single state. Combining states in this way makes the LP smaller but denser.

### 3.09 Performance Study

For our numerical study, we use \( T = 30 \) time periods each of length \( \Delta t \) in our model; thus, a machine age can never be exceed \( 30\Delta t \). We generate failure probabilities \( \text{fail}_i \) with
\[ \text{fail}_i = \frac{F((i + 1)\Delta t) - F(i \Delta t)}{1 - F(i \Delta t)} \]  

(3-23)

except that \( \text{fail}_{30} = 1 \). We choose \( F(t) \) to be a Weibull distribution with scale parameter \( \lambda = 1 \) in all cases, and shape parameter \( \eta = 1.5 \) for "high variance" cases and \( \eta = 4.0 \) for "low variance" cases. For the high variance cases, each period \( \Delta t = 0.10 \) units of time, and for the low variance cases each period \( \Delta t = 0.05 \) units of time. We fix \( \text{broke} = 1 \) and vary \( \text{rep} \) and \( \text{short} \) according to the failure distribution: low variance uses \( \text{rep} = \{0.2, 0.3, 0.4, 0.5\} \) and \( \text{short} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6\} \); high variance cases use \( \text{rep} = \{0.4, 0.5, 0.6, 0.8\} \) and \( \text{short} = \{0.2, 0.4, 0.8\} \). We test these parameters on systems of 2 and 3 machines, each with one repair shop. The parameter ranges give 24 combinations for the low variance case and 12 for the high; in total we have 72 different combinations of costs, repair rates, failure variance, and number of machines.

Solution of the linear program gives the optimal policy and its cost rate. To test the heuristic policies, we add extra constraints to the model which force the solution to have the appropriate structure, and use a simple search procedure to find the critical ages that give the best results for each heuristic in each case. We use AMPL mathematical modeling software to specify the linear program, and then CPLEX optimization software to solve it. Appendix D shows an example of such an AMPL model specification. The largest LP in this study has approximately 15,000 rows and 100,000 non-zero coefficients.

The tables which follow show the cost differences amongst the various policies in our discrete time study. The overall difference is that between the naive lower bound given by \( mC_j(t_j) \), which is what we might expect the cost rate to be if we used the single machine optimal age, and the feasible policy \( C_m(t_j) \), which is what the cost rate will actually be if we use the single machine optimal age. This overall difference can be divided into two parts:
(a) the first gap is that between $C_m(t_f)$ and the calculated optimum. This represents the amount we can potentially recover using other heuristics in place of the simple single age policy. We might also think of this as the value of having more information and coordination available in making maintenance decisions; and, (b) the second gap is that between the calculated optimum and $mC_f(t_f)$. This represents the cost premium due to running a system with multiple machines but only one shop (i.e. the unavoidable cost increase). This can also be thought of as the looseness of our naive lower bound $mC_f(t_f)$.

Tables 3-1 to 3-4 show the average and range of performances of each policy applied to our data set. Each table summarizes the results obtained when we vary the cost parameters as mentioned above while keeping the number of machines and the failure distribution fixed. The values listed are the percentages above the optimal expected cost rate, i.e. $\% = \frac{\text{heuristic cost} - \text{optimal cost}}{\text{optimal cost}}$; the Naive column refers to the lower bound $mC_f(t_f)$, and shows negative percentages because its cost rate is below the feasible optimum. WUA with $t_L = t_f$ represents the use of the wait until available heuristic except that we fix $t_L = t_f$ rather than searching for an optimal $t_L$; a similar simplification is also tried for the double age policy.

**Table 3-1:** Two machines, low failure variance (24 cases)

<table>
<thead>
<tr>
<th>Policy</th>
<th>Naive</th>
<th>Simple Age</th>
<th>Joint Single Age</th>
<th>WUA $t_f = t_f$</th>
<th>WUA $t_f = t_f$</th>
<th>Double Age, $t_f = t_f$</th>
<th>Double Age</th>
<th>Average Age</th>
<th>Constant Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-0.8%</td>
<td>0.6%</td>
<td>0.6%</td>
<td>0.4%</td>
<td>0.3%</td>
<td>0.3%</td>
<td>0.3%</td>
<td>0.5%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Max</td>
<td>-10.5%</td>
<td>6.5%</td>
<td>5.8%</td>
<td>5.6%</td>
<td>5.6%</td>
<td>1.2%</td>
<td>0.7%</td>
<td>0.9%</td>
<td>5.0%</td>
</tr>
<tr>
<td>Mean</td>
<td>-4.2%</td>
<td>4.1%</td>
<td>3.9%</td>
<td>1.1%</td>
<td>0.9%</td>
<td>0.7%</td>
<td>0.5%</td>
<td>0.5%</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

**Table 3-2:** Three machines, low failure variance (12 cases)

<table>
<thead>
<tr>
<th>Policy</th>
<th>Naive</th>
<th>Simple Age</th>
<th>Joint Single Age</th>
<th>WUA $t_f = t_f$</th>
<th>WUA $t_f = t_f$</th>
<th>Double Age, $t_f = t_f$</th>
<th>Double Age</th>
<th>Average Age</th>
<th>Constant Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-2.0%</td>
<td>1.5%</td>
<td>1.4%</td>
<td>0.8%</td>
<td>0.5%</td>
<td>0.6%</td>
<td>0.4%</td>
<td>0.1%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Max</td>
<td>-20.5%</td>
<td>18.1%</td>
<td>18.1%</td>
<td>12.5%</td>
<td>11.7%</td>
<td>2.2%</td>
<td>1.1%</td>
<td>2.1%</td>
<td>6.5%</td>
</tr>
<tr>
<td>Mean</td>
<td>-9.3%</td>
<td>9.1%</td>
<td>8.3%</td>
<td>2.3%</td>
<td>1.5%</td>
<td>1.4%</td>
<td>0.7%</td>
<td>0.9%</td>
<td>3.5%</td>
</tr>
</tbody>
</table>
Table 3-3: Two machines, high failure variance (24 cases)

<table>
<thead>
<tr>
<th>Policy</th>
<th>Naive</th>
<th>Simple Single Age</th>
<th>WUA $t_f = t_1$</th>
<th>WUA $t_f = t_j$</th>
<th>Double Age, $t_f = t_j$</th>
<th>Double Average Age</th>
<th>Constant Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-0.6%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Max</td>
<td>-7.0%</td>
<td>2.7%</td>
<td>2.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Mean</td>
<td>-3.1%</td>
<td>1.2%</td>
<td>1.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

Table 3-4: Three machines, high failure variance (12 cases)

<table>
<thead>
<tr>
<th>Policy</th>
<th>Naive</th>
<th>Simple Single Age</th>
<th>Joint Single Age</th>
<th>WUA $t_f = t_1$</th>
<th>WUA $t_f = t_j$</th>
<th>Double Age $t_j$</th>
<th>Double Average Age</th>
<th>Constant Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-1.4%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Max</td>
<td>-14.4%</td>
<td>5.7%</td>
<td>4.2%</td>
<td>0.2%</td>
<td>0.2%</td>
<td>0.2%</td>
<td>0.5%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Mean</td>
<td>-6.7%</td>
<td>2.8%</td>
<td>2.0%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

Figure 3-7 illustrates typical shapes observed for optimal policies for a two machine system. The extensions along the vertical axis and horizontal axis were a common feature to all the trial cases. On the other hand, the principle shape of the boundary ranged between being almost square (especially when the failure variance was high) to almost diagonal (when the failure variance was low). We might describe this shape as a squashed square, in which the degree of squashing varied according to the case. For the three machine cases, the three dimensional equivalents were observed; that is, the optimal policy resembled a squashed cube.

Figure 3-7: Typical Optimal Policies
By considering the above summaries as well as the individual solutions (not shown), we can develop some assessment of the performance of our policies. First of all note that the overall difference (i.e. Simple versus Naive) is increasing in $m$ and decreasing in the failure variance. The optimal policy cost rate seems on average to lie roughly in the middle of this overall difference. The gap between $C_m(t_j)$ and the optimal repair policies (i.e. the recoverable cost increase) appears to be increasing in the repair rate, roughly concave in the shortage cost, decreasing in the failure variance, and increasing in $m$. The gap between the optimal repair policy and $mC_I(t_j)$ (i.e. the unrecoverable cost increase) appears to be decreasing in the repair rate, roughly convex in the shortage cost, decreasing in the failure variance, and increasing in $m$.

In general, it seems that both of the single age policies are outperformed on average by all of the other heuristic policies considered, while within this latter set the differences on average are fairly small. To put it another way, it seems to be more important to use more than one critical age than it is to get the choice of critical age(s) exactly right. For example, the use of the joint single age $t_m$ does not provide much of a cost rate reduction compared to the simple single age $t_j$.

The double age policy performs best both on average and in most of the examples evaluated. This is not surprising, as it is a generalization of all of the other heuristics tested except for average age and constant cost, and thus will do at least as well in each case considered. When the relative shortage cost is moderate to high, the WUA policies are much better than the single age policies and do almost as well as the respective double age policies. Conversely, when the shortage cost is very low the single age policies perform better than WUA policies and are almost as good as the double age policies. As for the optimal ages themselves, we observe that they are ordered $t_L (3$ machines$) \leq t_L (two$ machines$) \leq t_1 \leq t_2 \leq t_3 < t_w$ in each case, with all the $t$ values quite close to each other except that the $t_w$ values tend to be much larger. That $t_1 \leq t_2 \leq t_3$ follows from Theorem
3-1, and by comparing their respective optimality criteria one can see that \( t_i \leq t_w \) also is to be expected.

The average age policy does almost as well overall as the double age policy, but with more variation in results. We can explain this by thinking of the shapes of these policies on our diagrams. Average age yields its best results when the optimal policy shape is almost a straight diagonal line, just like the shape of the average age policy in Figure 3-5. This occurs when the failure variance is low, the shortage cost is high, and the repair rate is high; in these situations it out-performs the double age policy. Conversely, the double age policy does best under the opposite conditions, when the optimal policy shape is almost a square as in Figure 3-3. The constant cost policy is generally the worst of the heuristic policies, but is still better than single age policies; it is at its best when shortage costs are high.

3.10 Simulation Study

This section describes a simulation study which tests the performance of some of our heuristics for the original continuous time model. The data set we use has \( m = \{2, 4, 6\} \), \( n = 1 \), \( R = \{0.05, 0.10\} \), \( c = 0 \), \( a = 0 \), and \( b = 1 \). We again use two Weibull failure distributions: with shape parameter \( \eta = 4.0 \) we let \( k = \{2, 16\} \), and with \( \eta = 1.5 \) we let \( k = \{4, 8\} \). For each combination of parameters we evaluate the simple single age and joint single age policies, the naive lower bound, and the double age policy with \( t_L \equiv t_I \). We choose to test the latter heuristic because it performed fairly well in our discrete time study and yet requires no numerical searching for parameter values. We evaluate the single age policies and the naive lower bound analytically using (3-8), and use SIGMA simulation software to evaluate the double age policy. The simulation was run for a simulated clock time of 10,000 for the 2 machine systems, and 5,000 for the 4 and 6 machine systems; thus each run consisted of at least 20,000 machine life-cycles. Each parameter combination was run twice using different random number seeds for the two runs; thus the study consists of 48 simulation runs in total. The table below gives the
mean % above our naive lower bound i.e. $\Delta\% = \frac{\text{heuristic cost} - \text{naive cost}}{\text{naive cost}}$ for the three policies, broken down by number of machines and failure variance. Each entry in Table 3-5 is the mean of 4 simulation runs. We are comparing the performance to the lower bound (naive cost) rather than the optimum because the optimum is unknown; from our discrete time study, we know that this lower bound can be fairly loose.

**Table 3-5: Three Policies Compared to Lower Bound**

<table>
<thead>
<tr>
<th></th>
<th>Low Variance $\eta = 4.0$</th>
<th>High Variance $\eta = 1.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy</td>
<td>Simple</td>
<td>Joint</td>
</tr>
<tr>
<td>$m=2$</td>
<td>5.1%</td>
<td>4.9%</td>
</tr>
<tr>
<td>$m=4$</td>
<td>27.0%</td>
<td>24.1%</td>
</tr>
<tr>
<td>$m=6$</td>
<td>52.7%</td>
<td>43.5%</td>
</tr>
</tbody>
</table>

The most obvious feature of these results is their magnitude. The two machine values are similar to what we found with the discrete time case, but as we increase the number of machines the percentages climb quite sharply. This is in part due to the fact that the repair shop is getting busier servicing more machines; for our test set, six machines give close to 100% utilization. As with the discrete time results, the high variance ($\eta = 1.5$) cases have smaller differences than do the low variance cases ($\eta = 4.0$).

We can see that the joint single age begins to provide a material cost saving as $m$ increases; more noticeably, the double age heuristic provides quite a large saving, reducing by roughly half the difference between the simple single age cost and the naive cost. This is an important point, because in our discrete time results we saw that the optimal cost rate tended to be about half-way between the naive and the simple single age costs. The implication here is that the double age policy may be near optimal once the number of machines gets moderately large. The trends in optimal age values in our discrete time results lend support to this conjecture.

One of our main interests is to see how our double age heuristic performs compared to the single age policies. Tables 3-6 and 3-7 show the amount by which the
double age heuristic improves on the cost of the joint single age policy, i.e. we consider $\Delta \% = \frac{\text{joint cost} - \text{double cost}}{\text{joint cost}}$. Table 3-6 shows the average improvements broken down by the number of machines; each entry is the mean of 16 simulation runs. We can see that the cost savings are increasing in the number of machines.

**Table 3-6:** Mean Improvement by Number of Machines

<table>
<thead>
<tr>
<th>$m$</th>
<th>Mean $\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.0%</td>
</tr>
<tr>
<td>4</td>
<td>7.5%</td>
</tr>
<tr>
<td>6</td>
<td>9.2%</td>
</tr>
</tbody>
</table>

Table 3-7 shows the average improvements broken down by the other three factors that were varied; each entry is the mean of 24 simulation runs. We can see that the cost savings tend to be larger when each factor is relatively low.

**Table 3-7:** Mean Improvement by Other Factor

<table>
<thead>
<tr>
<th>Level \ Factor</th>
<th>Failure Variance</th>
<th>Shortage Cost Rate</th>
<th>Mean Repair Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>10.7%</td>
<td>8.5%</td>
<td>7.6%</td>
</tr>
<tr>
<td>High</td>
<td>2.5%</td>
<td>4.7%</td>
<td>5.6%</td>
</tr>
</tbody>
</table>

These two tables indicate that the cost savings due to using the double age heuristic in place of the joint single age policy tend to be higher when the failure variance, shortage cost rate, and mean repair time were relatively small while the number of machines is relatively large. Thus our results here for simulation in continuous time are in general agreement with those we obtained in our discrete time study.

To check whether these results might be due to the randomness in simulation, we performed an analysis of variance (ANOVA) test. We used a fixed effect additive model which considered the main effects of our four factors along with their first-order interactions. Calculation of the F-statistics indicated that all of the main effects were statistically significant ($P < 0.001$), as were most of the interactions: the interactions between failure variance and repair time, and between shortage cost and repair time, were
not significant (both $P = 0.09$). A check of the marginal means showed that even where the interaction effects were statistically significant, there were no counter-intuitive results (e.g. when failure variance and shortage cost were both low, the cost savings was larger than when either or both factors were high). Thus we can reasonably conclude that the improvements seen in Tables 3-6 and 3-7 are not just due to random behaviour of the simulation.

3.11 Future Research

There are several directions in which we would like to be able to extend the work in this chapter. One possibility is to drop our requirement for the machines to be identical, so as to allow each to have a different failure rate or cost structure. Possibly one could build on the results of Sztrik in [Sz85] for such an extension.

In our model we allow for multiple repair shops, but we take their number $n$ as given and so ignore their cost. If we included the shop costs and considered $n$ as another decision variable, then we could address questions of capacity planning, i.e. what is the optimal number of repair shops to build and operate? If both the shops and the machines are heterogeneous then this would raise another issue commonly considered in the MRP, the optimal assignment of machines to shops.

We would also like to allow the repair of machines to be deterministic in duration, as this might better model maintenance of complex machines where a repair action may actually be a relatively standardized set of inspections and adjustments. Equivalent numerical work could readily be done using the same dynamic programming and simulation methods to compare heuristics. The difficulty here is that any analytical work would need to take a different approach in order to be tractable, as to our knowledge there are no existing results for finite population $GI/D/n$ queues. One possible approach would use our discrete time framework as the basis for a semi-markov model, which could handle repair times with general distributions.
3.12 Concluding Remarks

Our analytical and numerical work shows clearly that for the type of system we consider, single age policies are quite tractable to analyze and optimize; but their performance is sub-optimal and can result in potentially meaningful inefficiencies. We can improve on this situation using heuristic policies; we find that these tend to be most worthwhile (i.e. give the largest cost savings) in systems with more machines, low shortage costs, low failure variance, and a fast repairs. Finally, although we are unable to describe the truly optimal policy structure, the heuristics we examine seem to perform well enough to make consideration of more complicated policies unnecessary, at least for the small systems we studied.

For actual applications, we suggest the following guidelines:

(a) if machine lifetimes have a high variance, then it does not greatly matter what policy is chosen, so make the choice based upon other factors (such as administrative convenience, labour considerations, etc.);

(b) if lifetime variance is fairly low and shortage costs are moderate or low relative to breakage costs, then the double age policy would likely be the best policy type, with \( t_L = t_1 \) being a reasonable choice for initial consideration; or,

(c) if lifetime variance is fairly low and shortage costs are high relative to breakage costs, then the average age policy would likely be the best choice, with \( t_a = t_1 / m \) being a good value for initial consideration.

Note that these recommendations are based upon our numerical studies of relatively small systems using only two example failure distributions, and so of course their applicability to general systems is a matter of conjecture.
Bibliography


[BHP63] Barlow RE, Hunter LC, Proschan F, "Optimum redundancy when components are subject to two kinds of failure", Journal of SIAM 11, 64-73, 1963


[Cr75] Crabill TB, "Optimal control of a maintenance system with variable service rates", *Operations Research* 22, 736-745, 1974


[HL93] Hong JS, Lie CH, "Joint reliability importance of two edges in an undirected network", *IEEE Transactions on Reliability* 42, 17-23, 1993


[PP86a] Park KS, Park YT, "Ordering policies under minimal repair", *IEEE Transactions on Reliability* 35, 82-84, 1986


[PM94] Pham, Malone, "Optimal design of a system with competing failure modes", *IEEE Transactions on Reliability* 43, 251-2xx, 1994


[SLT92] Sheu SH, Liou CT, Tseng BC, "Optimal ordering policies and optimal number of minimal repairs before replacement", *Microelectronics and Reliability* 32, 995-1002, Jul 1992


[Sm77] Smith SA, "Optimal inventories for an (s-1,s) system with no backorders", *Management Science* 23, 522-528, 1977

[Sr92] Sridharan V, "Optimum planned policies with minimal repair and random lead times", *Microelectronics and Reliability* 32, 905-909, Jul 1992


Appendix A

This appendix describes the specific KKT conditions that apply to the optimization of equation \( JOD \) in Chapter 2 and outlines how one might use a customized algorithm to search for an optimal solution \((t_o, t_r)\).

An interior solution (both derivatives zero) will occur at a point where (2-13) is true, so we could direct our search amongst such points. To start with, the particular KKT conditions for the joint optimality of \( JOD \) are shown below.

**General KKT Conditions:**

\[
\frac{\partial JOD(t_o, t_r)}{\partial t_o} - u + v = 0 \\
\frac{\partial JOD(t_o, t_r)}{\partial t_r} - v = 0 \\
u(t_o) = 0 \\
v(t_r - t_o - L) = 0 \\
u \geq 0, \quad v \geq 0 \\
t_o \geq 0, \quad t_r \geq t_o + L
\]

We can write these conditions out more specifically for each of the 4 categories of optimal solutions, using our earlier results to simplify the derivative terms:

(a) interior solutions, i.e. where \( t_o \geq 0 \) and \( t_r \geq t_o + L \)

\[
v'(t_r) + aq(t_r) + h + bz(t_r) = k - h \left( \frac{\bar{F}(t_o + L)}{F(t_o + L) - F(t_o)} \right) = JOD(t_o, t_r)
\]

(b) \( t_o \) boundary solutions, i.e. where \( t_o = 0 \) and \( t_r \geq t_o + L \)

\[
v'(t_r) + aq(t_r) + h + bz(t_r) = JOD(t_o, t_r)
\]

\[
k - h \left( \frac{\bar{F}(t_o + L)}{F(t_o + L) - F(t_o)} \right) \geq JOD(t_o, t_r)
\]

(c) \( t_r \) boundary solutions, i.e. where \( t_o \geq 0 \) and \( t_r = t_o + L \)

\[
v'(t_r) + aq(t_r) + h + bz(t_r) \geq JOD(t_o, t_r)
\]

\[
\frac{\partial JOD}{\partial t_o} = -\frac{\partial JOD}{\partial t_r}
\]
(d) \((0, L)\) boundary solutions, i.e. where \(t_o = 0\) and \(t_r = L\)

\[
   c v'(t_r) + a q(t_r) + h + b z(t_r) \geq JOD(t_o, t_r)
\]

\[
   \partial JOD / \partial t_o \geq -\partial JOD / \partial t_r
\]  

We could substitute these optimality conditions into an existing NLP package to simplify the evaluation of candidate solutions. Alternatively, they could be used directly in the following procedure, which acts as a front end to a standard NLP package; it attempts to find the "easy" solutions using the above conditions, but defaults to the NLP code if it runs into "difficulties". It proceeds in basically 4 steps as follows:

Step 1: Start at \((0, L)\) and check to see if (A-6) and (A-7) hold as per case (d) above.

Step 2: Look for a point on the \(t_o = 0\) boundary such that (A-2) is true. Check this point to see if the unconstrained optimum has \(t_o < 0\), and if so the optimality conditions under case (b) will be met.

Step 3: Search along the (increasing) curve defined by (2-14) to try to find a feasible point where (A-1) holds and so the optimality conditions under case (a) will be met.

Step 4: If difficulties are encountered with any of the above steps, or the \(t_r = t_o + L\) boundary is run into during the search steps, then switch over to the standard NLP code to finish solving the problem.

Of course, if we somehow knew in advance that the optimum was an interior solution of case (a), we could greatly simplify our search procedure to cover only the portion in Step 3 above and not refer to an NLP package at all.

An outline of an algorithm follows below. Note that for convenience of notation, we define:

\[
   Gr(t_r) = c v'(t_r) + a q(t_r) + h + b z(t_r)
\]  

\[
   Go(t_o) = k - h \left[ \frac{\overline{F}(t_o + L)}{F(t_o + L) - F(t_o)} \right]
\]  

\[\text{[A-8]}\]  

\[\text{[A-9]}\]
Algorithm 2-1:

1. Set $t_o = 0$, $t_r = L$, $Z = Go(0)$
2. If $Gr(L) \geq JOD(0,L)$ and $\frac{\partial JOD(t_o, t_r)}{\partial t_o} \geq -\frac{\partial JOD(t_o, t_r)}{\partial t_r}$, then Goto 15
3. If $Go(0) \geq Gr(L)$ then
   4. Find by Line Search $t_r : Gr(t_r) = Z$
   5. If $Gr(t_r) \geq JOD(0, t_r)$ then
      6. Find by Line Search $t_r : Gr(t_r) = JOD(0, t_r)$
      7. Goto 15
4. Set $Z = Z + 0.5 \left[ JOD(t_o, t_r) - Z \right]$
5. Find by Line Search $t_o : Go(t_o) = Z$
6. Find by Line Search $t_r : Gr(t_r) = Z$
7. If $t_r < t_o + L$ then set $t_r = t_o + L$
8. If $JOD(t_o, t_r) = Z$ then Goto 15
9. If $t_r > t_o + L$ then Goto 8
10. Call NLP Routine with $(t_o, t_r)$ as initial feasible point
11. Stop with $(t_o, t_r)$ optimal

The "0.5" in line 8 determines the step size for the search; any value in the interval $(0,1)$ could be substituted instead. Also note that we should skip immediately to line 14 if any of the line searches fail.

$Go$ and $Gr$ are known to be monotonic and do not involve integration in their calculation, so their line searches proceed fairly quickly. The evaluation of $JOD$ and its derivatives can be time consuming if $F(t)$ is not analytically integrable (as with a Weibull distribution, for example) but instead must be evaluated numerically. Thus the line search in step 6 would likely be the slowest part of the algorithm (though it is executed only once), while step 12 would be the slowest step that is repeated each pass through the search loop.

We did not use this algorithm in our numerical study, but instead used a simple coordinate search algorithm with line searches done using a bisection procedure (refer to [BSS93] chapters 8 and 9 for explanations of these methods).
Appendix B

For convenience, in what follows the numerator and denominator of an equation are sometimes represented by $N$ and $D$ respectively. We also use the results from chapter 3 of [BP75] that

$$F(t) \text{ is IFR} \iff \frac{\partial}{\partial t} \left[ \frac{f(t)}{F(t)} \right] \geq 0 \iff \frac{\partial}{\partial t} \left[ \frac{\bar{F}(t)}{\bar{F}(t + x)} \right] \geq 0$$

The inequalities would be strict for a strictly IFR distribution, while they would be reversed for a DFR distribution.

Most of the proofs in this appendix use the following line of reasoning. Begin by taking the first derivative of a cost function with respect to a decision variable. When this derivative equals zero, the function is at an extreme point (either a maximum or a minimum). Next take the second derivative of the cost function, evaluated at such an extreme point. If that second derivative is always positive, that indicates that the extreme point is always a minimum, and hence must be the global minimum of the cost function with respect to that decision variable. Refer to chapter 3 of [BSS93] for more information on non-linear optimality properties.

**Proof of Theorem 2-1:** For this theorem and the next two, we actually consider the more general function $JOD$ rather than $JCF$. Take $\partial JOD / \partial t$, and rearrange terms to get

$$\frac{\partial}{\partial t} JOD = \frac{\bar{F}(t_r)}{D} \left\{ cv'(t_r) + h + bz(t_r) + aq(t_r) - JOD(t_o, t_r) \right\} \tag{B-1}$$

When this is zero, take derivatives ($\partial / \partial t$) again to get

$$\frac{\partial^2}{\partial t^2} JOD = \frac{\bar{F}(t_r)}{D} \left\{ cv''(t_r) + bz'(t_r) + aq'(t_r) \right\} \geq 0 \tag{B-2}$$

where the inequality is strict if $F(t)$ is strictly IFR, if $q(t)$ is strictly increasing, or if $v(t)$ is strictly convex. This means that any zero of $\partial JOD / \partial t_r$ will be a minimum, and hence there can be only one minimum. So $JOD$ is unimodal and pseudo-convex in $t_r$. Note that for (B-1) to equal zero we must have the bracketed term equal to zero, which occurs when

$$[ cv'(t_r) + h + bz(t_r) + aq(t_r) ] = JOD(t_o, t_r) \tag{B-3}$$
Proof of Theorem 2-2: Take $\frac{\partial JOD}{\partial t_o}$ and rearrange terms to get

$$\frac{\partial}{\partial t_o} JOD = \left[ \bar{F}(t_o) - \bar{F}(t_o + L) \right] \left\{ k - h \frac{\bar{F}(t_o + L)}{\bar{F}(t_o) - \bar{F}(t_o + L)} - JOD(t_o, t_r) \right\} \tag{B-4}$$

When this is zero, take derivatives ($\frac{\partial}{\partial t_o}$) again to get

$$\frac{\partial^2}{\partial t_o^2} JOD = \left[ \bar{F}(t_o) - \bar{F}(t_o + L) \right] \left\{ h \frac{\partial}{\partial t_o} \left( \frac{-\bar{F}(t_o + L)}{\bar{F}(t_o) - \bar{F}(t_o + L)} \right) \right\} \geq 0 \tag{B-5}$$

This means that any zero of $\frac{\partial JOD}{\partial t_o}$ will be a minimum, and hence there can be only one minimum. So $JOD$ is unimodal and pseudo-convex in $t_o$. Note that for (B-4) to equal zero we must have the bracketed term equal to zero, which occurs when

$$\left[ k - h \frac{\bar{F}(t_o + L)}{\bar{F}(t_o) - \bar{F}(t_o + L)} \right] = JOD(t_o, t_r). \tag{B-6}$$

Proof of Proposition 2-1: Proven simply by examining the signs of the derivatives of $\frac{\partial}{\partial t_o} JOD$ and $\frac{\partial JOD}{\partial t_o}$ by each of the cost coefficients. If the second derivative is positive, this implies that the first derivative is increasing in that cost coefficient and hence that the minimum of the function is decreasing; the opposite results follows if the second derivative is negative.

$$\frac{\partial^2}{\partial t_o \partial t_r} JOD = \frac{\bar{F}(t_r)}{D^2} \left\{ q(t_r) \left[ \int_0^{t_r} \bar{F}(x)dx + \int_{t_r}^{t_r+L} F(x)dx + R \right] - \int_0^{t_r} q(x)\bar{F}(x)dx \right\} > 0 \tag{B-7}$$

$$\frac{\partial^2}{\partial t_o \partial \bar{t}} JOD = \frac{\bar{F}(t_r)}{D^2} \left\{ z(t_r) \left[ \int_0^{t_r} \bar{F}(x)dx + \int_{t_r}^{t_r+L} F(x)dx + R \right] - F(t_r) \right\}$$

$$> \frac{\bar{F}(t_r)}{D^2} \left\{ \int_0^{t_r} \frac{f(t_r)}{F(t_r)} dx - F(t_r) \right\} = \frac{\bar{F}(t_r)}{D^2} \left\{ \int_0^{t_r} \frac{f(x)}{\bar{F}(x)} dx - F(t_r) \right\} = 0$$
\[
\frac{\partial^2}{\partial t_0 \partial \delta} JOD = \frac{\bar{F}(t_0) + F(t_0)}{D^2} \left\{ - \int_0^{t_0} q(x) \bar{F}(x) \, dx \right\} < 0 \quad (B-12)
\]

\[
\frac{\partial^2}{\partial t_0 \partial \delta} JOD = \frac{F(t_0) + L}{D^2} \left\{ -F(t_0) \right\} < 0 \quad (B-13)
\]

\[
\frac{\partial^2}{\partial t_0 \partial \delta} JOD = \frac{F(t_0) + L}{D^2} \left\{ \int_0^{t_0} v(x) f(x) \, dx - v(t_0) \bar{F}(t_0) \right\} < 0 \quad (B-14)
\]

\[
\frac{\partial^2}{\partial t_0 \partial \delta} JOD = \frac{F(t_0) + L}{D^2} \left\{ - \frac{\bar{F}(t_0 + L)}{[F(t_0 + L) - F(t_0)]} D - \int_0^{t_0} \bar{F}(x) \, dx \right\} < 0 \quad (B-15)
\]

\[
\frac{\partial^2}{\partial t_0 \partial \delta} JOD = \frac{F(t_0) + L}{D^2} \left\{ D - R - \int_0^{t_0 + L} \bar{F}(x) \, dx \right\} > 0 \quad (B-16)
\]

**Proof of Theorem 2-3:** It is helpful to refer to Figure B-1 while reading through this rather long proof, which proceeds in three main steps. Step 1 shows that there can be at most one local minimum point in the interior or on the lower boundary of the feasible region, and that it will always be viable; step 2 shows that there can be at most one viable local minimum point on the diagonal boundary; and step 3 shows that these two types of minima are mutually exclusive, so that a viable local minimum must be a global minimum.
Step 1: consider points \((t_o, t_r)\) such that \(t_r > t_o + L\). From Theorems 2-1 and 2-2 it follows that any KKT point with \(t_r > t_o + L\) is a local minimum. Also by Theorem 2-1, there cannot exist more than one local minimum on the lower boundary (i.e. with \(t_o = 0\)).

**Figure B-1:** Interior KKT point

Suppose \((t_o', t_r')\) is an interior minimum, like the point indicated in Figure B-1. From (B-4), we see that

\[
\left[ \frac{\partial}{\partial t_o} JOD(t_o, t_r) \right] \text{ has the same sign as } \left\{ k - h \frac{\overline{F}(t_o + L)}{[F(t_o + L) - F(t_o)]} \right\} - JOD(t_o, t_r) \tag{B-17}
\]

and by Theorem 2-1, we know that \(JOD(t_o', t_r') > JOD(t_o', t_r') \forall t_r \neq t_r'\). Combining these facts shows that

\[
\left[ k - h \frac{\overline{F}(t_o' + L)}{[F(t_o' + L) - F(t_o')] - JOD(t_o', t_r)} \right] < 0 \quad \forall t_r \neq t_r' \tag{B-18}
\]

and so

\[
\left[ \frac{\partial}{\partial t_o} JOD(t_o', t_r) \right] < 0 \quad \forall t_r \neq t_r' \tag{B-19}
\]

Next make use of the above together with Theorem 2-2 to show that
\[ \left[ \frac{\partial}{\partial o} JOD(t_o, t_r) \right] < 0 \quad \forall \ t_o < t_o', \forall \ t_r. \]  
(B-20)

So there cannot exist a local minimum \((t_o, t_r)\) with \(0 < t_o < t_o'\); or in terms of Figure B-1, there can not be another minimum "below" the given interior minimum, including on the lower boundary where \(t_o = 0\). Conversely, there cannot exist another interior local minimum "above" the given one, as we could then use our argument to rule out the existence of the first one. Thus we conclude that there can be at most one local minimum with \(t_r > t_o + L\).

Next we show that any KKT point with \(t_r > t_o + L\) must be viable. Since we are at a KKT point, either we are at an interior minimum with \(t_o > 0\) so that \(\partial JOD(t_o, t_r) / \partial t_o = 0\); or else we are at a lower boundary minimum with \(t_o = 0\) so that \(\partial JOD(t_o, t_r) / \partial t_o \geq 0\); either way, from equation (B-4) we have that

\[ \left\{ \left[ k - h \frac{F(t_o + L)}{F(t_o) - F(t_o + L)} \right] - JOD(t_o, t_r) \right\} \geq 0. \]  
(B-21)

Since the fraction inside the square brackets is non-negative for any \(t_o\), it must be that

\[ k \geq JOD(t_o, t_r) \]  
(B-22)

where the inequality is strict when \(F(l)\) is strictly IFR. This completes step 1 of the proof.

Step 2: consider points such that \(t_r = t_o + L\) (i.e. a KKT point on the diagonal boundary). This step proceeds in the following way. First note that a KKT point with \(t_r = t_o + L\) is either a local minimum or a saddle point. We show that if such a point is viable then it must be a local minimum, whereas if it is a saddle point then it cannot be viable. Next show that there cannot be more than one viable diagonal minimum.

Consider \(JOD\) restricted to points on the diagonal boundary, i.e. let \(J(t_o) = JOD(t_o, t_o + L)\), so as to create a univariate function in \(t_o\). In this representation a KKT point for \(JOD\) corresponds to either a univariate minimum (if the KKT point is a local minimum) or else a univariate maximum (if the KKT point is a saddle point).

(B-23)

\[ J(t_o) = \frac{cC(t_o + L) + bF(t_o + L) + a \int_{0}^{t_o + L} q(x) F(x) dx + k \int_{t_o}^{t_o + L} F(x) dx + kR \int_{0}^{t_o} F(x) dx + L + R}{kF(x) dx + L + R} \]
Take $\frac{\partial J(t_o)}{\partial t_o}$ and rearrange terms to get (B-24)

$$\frac{\partial}{\partial t_o} J = \frac{1}{D} \left\{ cv' (t_o + L) \bar{F}(t_o + L) + bf (t_o + L) \\
=k \left[ F(t_o + L) - F(t_o) \right] + aq(t_o + L) \bar{F}(t_o + L) - \bar{F}(t_o) J(t_o) \right\}$$

Note that when this is zero, it must be that

$$J(t_o) = \frac{1}{\bar{F}(t_o)} \left\{ cv' (t_o + L) \bar{F}(t_o + L) + bf (t_o + L) \\
=k \left[ F(t_o + L) - F(t_o) \right] + aq(t_o + L) \bar{F}(t_o + L) \right\}$$

When (B-24) is zero, take derivatives $(\frac{\partial J(t)}{\partial t_o})$ again to get (B-25)

$$\frac{\partial^2}{\partial t_o^2} J = \frac{1}{D} \left\{ cv'' (t_o + L) \bar{F}(t_o + L) - cv' (t_o + L) f(t_o + L) \\
+b f' (t_o + L) + k \left[ f(t_o + L) - f(t_o) \right] \\
+a q' (t_o + L) \bar{F}(t_o + L) - a q(t_o + L) f(t_o + L) \\
+f (t_o) J(t_o) \right\}$$

Replace the $J(t_o)$ term in (B-26) by its equivalent from (B-25) to get (B-27)

$$\frac{\partial^2}{\partial t_o^2} J = \frac{1}{D} \left\{ cv'' (t_o + L) \bar{F}(t_o + L) - cv' (t_o + L) f(t_o + L) \\
+b f' (t_o + L) + k \left[ f(t_o + L) - f(t_o) \right] \\
+a q' (t_o + L) \bar{F}(t_o + L) - a q(t_o + L) f(t_o + L) \\
+z(t_o) \left\{ cv' (t_o + L) \bar{F}(t_o + L) + f(t_o + L) \\
+k \left[ F(t_o + L) - F(t_o) \right] + a q(t_o + L) \bar{F}(t_o + L) \right\} \right\}$$

and then rearrange terms to get
\[
\frac{\partial^2 J}{\partial \alpha_0^2} = \frac{1}{F(t_o) D} \left\{ c \left[ v''(t_o + L)F(t_o + L)F(t_o) - v'(t_o + L)f(t_o + L)F(t_o) \right] \\
+ v'(t_o + L)F(t_o + L)f(t_o) \\
+ b \left[ f'(t_o + L)F(t_o) + f(t_o + L)f(t_o) \right] \\
+ a \left[ q'(t_o + L)F(t_o + L) - q(t_o + L)f(t_o + L)F(t_o) \right] \\
+ k \left[ f(t_o + L) - f(t_o) \right] F(t_o) + \left( F(t_o + L) - F(t_o) \right) f(t_o) \right\}
\]

and then

\[
\frac{\partial^2 J}{\partial \alpha_0^2} = \frac{1}{F(t_o + L) D} \left\{ c \left[ v''(t_o + L) - v'(t_o + L)z(t_o + L) + v'(t_o + L)z(t_o) \right] \\
+ b \left[ f'(t_o + L)F(t_o + L) + z(t_o + L)z(t_o) \right] \\
+ a \left[ q'(t_o + L) - q(t_o + L)z(t_o + L) + q(t_o + L)z(t_o) \right] \\
+ k \left[ z(t_o + L) - z(t_o) \right] \right\}
\]

Note that because \( z(t) \) is increasing in \( t \), the \( k \) term is always positive; furthermore, in many cases the entire second derivative will be positive and hence a zero of the first derivative will be a minimum. Unfortunately, while this is often true it is not always the case, so next we show a sufficient condition for when it will be true. In what follows we assume that we are at a \( t_o : \partial J(t_o) / \partial t_o = 0 \) (one of the points indicated on Figure B-2) which corresponds to a KKT point of \( JOD \).
Figure B-2: Diagonal Boundary KKT Points

Suppose that $J(t_o) < k$, i.e. it is viable. Substitute $k$ in place of $J(t_o)$ in (B25) to get

$$\begin{align*}
\left\{ cv'(t_o + L)\overline{F}(t_o + L) + bf(t_o + L) \\
+ k\left[F(t_o + L) - F(t_o)\right] + aq(t_o + L)\overline{F}(t_o + L) \right\} &< \overline{F}(t_o)k \\
\end{align*}$$

(B-30)

Divide both sides by $\overline{F}(t_o + L)$ and rearrange terms to see that this is equivalent to

$$cv'(t_o + L) + bz(t_o + L) + aq(t_o + L) < k$$

(B-31)

We know that the hazard rate is increasing, so its derivative must be positive, i.e.

$$\frac{\partial}{\partial t_o} z(t_o + L) = \frac{\overline{F}(t_o + L)f'(t_o + L) + f(t_o + L)f(t_o + L)}{\overline{F}(t_o + L)^2} > 0$$

(B-32)

Work with the numerator (which must also be positive) as follows.

$$f(t_o + L)f(t_o + L) > -\overline{F}(t_o + L)f'(t_o + L)$$

(B-33)

$$f(t_o + L)f(t_o + L)\overline{F}(t_o) - f(t_o + L)f(t_o)\overline{F}(t_o + L)$$

$$> -f'(t_o + L)\overline{F}(t_o + L)\overline{F}(t_o) - f(t_o + L)f(t_o)\overline{F}(t_o + L)$$

(B-34)

$$\frac{f(t_o + L)}{\overline{F}(t_o + L)} \left[f(t_o + L)\overline{F}(t_o) - f(t_o)\overline{F}(t_o + L)\right]$$

$$> -f'(t_o + L)\overline{F}(t_o) - f(t_o + L)f(t_o)$$

(B-35)
\[
f(t_o + L) > \frac{-f'(t_o + L)\bar{F}(t_o) - f(t_o + L)f(t_o)}{f(t_o + L)\bar{F}(t_o) - f(t_o)\bar{F}(t_o + L)} \quad \text{(B-36)}
\]

\[
z(t_o + L) > -\frac{f'(t_o + L) + z(t_o + L)z(t_o)}{z(t_o + L) - z(t_o)} \quad \text{(B-37)}
\]

We also know that \( q(t) \) is increasing, so its derivative must be positive, i.e.

\[
\frac{\partial}{\partial t_o} q(t_o + L) > 0 \quad \text{(B-38)}
\]

Work with this as follows.

\[
q'(t_o + L) - q(t_o + L)z(t_o + L) + q(t_o + L)z(t_o) > -q(t_o + L)z(t_o + L) + q(t_o + L)z(t_o) \quad \text{(B-39)}
\]

\[
-\frac{q'(t_o + L) - q(t_o + L)z(t_o + L) + q(t_o + L)z(t_o)}{z(t_o + L) - z(t_o)} < q(t_o + L) \quad \text{(B-40)}
\]

Substitute (B-37) and (B-40) into (B-31) to get

\[
\begin{bmatrix}
-cv'(t_o + L) - b \frac{f'(t_o + L)}{\bar{F}(t_o + L)} + z(t_o + L)z(t_o) \\
-a q'(t_o + L) - q(t_o + L)z(t_o + L) + q(t_o + L)z(t_o)
\end{bmatrix} < k \quad \text{(B-41)}
\]

Re-arranging terms yields

\[
\begin{bmatrix}
-cv'(t_o + L)[z(t_o + L) - z(t_o)] + k[z(t_o + L) - z(t_o)] \\
b \left[ \frac{f'(t_o + L)}{\bar{F}(t_o + L)} + z(t_o + L)z(t_o) \right] \\
+a[q'(t_o + L) - q(t_o + L)z(t_o + L) + q(t_o + L)z(t_o)]
\end{bmatrix} > 0 \quad \text{(B-42)}
\]
Referring back to (B-29), we see that this implies that the second derivative must be positive, and therefore $t_o$ must be a minimum of $J(t_o)$ and in turn $(t_o, t_o + L)$ must be a minimum of $JOD$. This gives the result that if a KKT point is viable, it must be a minimum.

By reversing the preceding line of reasoning, we can derive the result that if a KKT point is not a minimum (i.e. is a saddle point) of $JOD$, then it cannot be viable.

Next suppose that we are at a KKT point $(t_o, t_o + L)$ that is not viable, so that $J(t_o) > k$ and $\partial J(t_o) / \partial t_o = 0$. Then substitution of $k$ into (B-31) shows that

$$cv'(t_o + L) + bz(t_o + L) + aq(t_o + L) > k \quad \text{(B-43)}$$

Since $z(t), q(t),$ and $v'(t)$ are all increasing in $t$, we know that the left hand side of this expression is increasing $\forall t > t_o$, and so this relation will remain true $\forall t > t_o$. Hence if there exists a $t_o$ that is viable, it must be earlier i.e. at some $t < t_o$ (refer again to Figure B-2).

Now suppose we find a $t_o$ that is viable and hence is a local minimum of $J(t_o)$. For there to be some other local minimum on the diagonal, we must have local maximum in between them. But a local maximum of $J(t_o)$ is a saddle point of $JOD$, which we already know is never viable. By the preceding argument, any point beyond a saddle point can not be viable either. Thus there can be no other viable local minimum on the diagonal. This result completes step 2 of the proof.

Step 3: if an interior or lower boundary minima exists then there cannot be any viable diagonal minimum, and vice versa. First, suppose $(t_o', t_o')$ is an interior minimum, like the point indicated in Figure B-3.

**Figure B-3: Interior KKT Point Again**

From (B-1), we see that
\[
\left[ \frac{\partial}{\partial r} JOD(t_o, t_r) \right] \text{ has the same sign as }
\left\{ \left[ h + cv'(t_r) + bz(t_r) + aq(t_r) \right] - JOD(t_o, t_r) \right\}
\]  
(B-44)

and by Theorem 2-2, we know that \( JOD(t_o, t_r') > JOD(t_o', t_r') \ \forall \ t_o \neq t_o' \). Combining these facts shows that
\[
\left[ h + cv'(t_r) + bz(t_r) + aq(t_r) - JOD(t_o, t_r) \right] < 0 \ \forall t_o \neq t_o' \quad (B-45)
\]

and so
\[
\left[ \frac{\partial}{\partial r} JOD(t_o, t_r') \right] < 0 \ \forall t_o \neq t_o' \quad . 
\]  
(B-46)

Next make use of the above together with Theorem 2-1 to show that
\[
\left[ \frac{\partial}{\partial r} JOD(t_o, t_r) \right] < 0 \ \forall t_r < t_r', \forall t_o \quad . 
\]  
(B-47)

So there cannot exist another local minimum \( (t_o, t_r) \) with \( 0 \leq t_r \leq t_r' \); or in terms of Figure B-3, there cannot be another minimum "to the left of" the given interior minimum, including on the diagonal boundary where \( t_r = t_o + L \). Conversely, there cannot exist an interior local minimum \( (t_o', t_r') \) "to the right of" a diagonal boundary minimum \( (t_o - L, t_r) \) with \( t_r \leq t_r' \), as we could then use our argument to rule out the existence of the latter. Thus there cannot exist both an interior minimum \( (t_o', t_r') \) and a diagonal boundary minimum with \( t_r \leq t_r' \).

Continue to suppose that an interior minimum \( (t_o', t_r') \) exists, and let \( (t_o'', t_r'') \) be a feasible point such that \( t_o'' > t_o' \) and \( t_r'' > t_r' \), i.e. to the upper right of the interior minimum, as shown in Figure B-4.
Consider the curve where $\frac{\partial J(t^*_0, t^*_r)}{\partial t^*_0} = \frac{\partial J(t^*_0, t^*_r)}{\partial t^*_r}$:

(a) if $(t^*_0, t^*_r)$ lies to the upper left of this curve, i.e. $t^*_0 > t^*_r$ for some $t^*_r$, then Theorem 2-2 shows that $\frac{\partial J(t^*_0, t^*_r)}{\partial t^*_0} > 0$ and so $(t^*_0, t^*_r)$ cannot be a KKT point; or,

(b) if $(t^*_0, t^*_r)$ lies to the lower right of this curve, i.e. $t^*_0 > t^*_r$ for some $t^*_r$, then Theorem 2-1 shows that $\frac{\partial J(t^*_0, t^*_r)}{\partial t^*_0} > 0$ and so $(t^*_0, t^*_r)$ cannot be a KKT point.

These two cases together show that there cannot exist both an interior minimum $(t^*_0, t^*_r)$ and a diagonal boundary minimum with $t^*_r < t^*_0$. So we have the result that there cannot exist both an interior minimum and a diagonal boundary minimum.

Next suppose that $(0, t^*_r)$ is a lower boundary minimum, like the point indicated in Figure B-5; and further suppose that there also exists a minimum on the diagonal boundary at $(t^*_r - I, t^*_r)$ The argument built on (B-44) to (B-47) can be used to show that there can be no diagonal boundary minimum to the left, i.e. with $t^*_r < t^*_r$, so we need only worry about the existence of a diagonal minimum to the right, i.e. with $t^*_r > t^*_r$.
By Theorem 2-2, \( \partial JOD(t_r' - L, t_r') / \partial t_o > 0 \). Since \((t_r'' - L, t_r'') \) is a KKT point, \( \partial JOD(t_r'' - L, t_r'') / \partial t_o < 0 \). Hence there exists some \( t_r^* \): \( \partial JOD(t_r^* - L, t_r^*) / \partial t_o = 0 \), \( t_r' < t_r^* < t_r'' \). And then again by Theorem 2-2 it must be that \( \partial JOD(0, t_r^*) / \partial t_o < 0 \). Now again use the argument built on (B-44) to (B-47), this time to show that \( \partial JOD(0, t_r^*) / \partial t_o < 0 \). But since \( \partial JOD(0, t_r') / \partial t_o = 0 \) then by Theorem 2-1 it must be that \( \partial JOD(0, t_r^*) / \partial t_o > 0 \). This contradiction shows that there cannot exist a diagonal minimum to the right, i.e. with \( t_r'' > t_r' \). Thus we have the result that there cannot exist both a lower boundary minimum and a diagonal boundary minimum. This result completes step 3 of the proof.

These three steps allow us to conclude that a viable KKT point is a global minimum. In effect it seems that while \( JOD \) is not in general pseudo-convex, it is quasi-connected or arc-wise quasi-convex (see chapter 9 of [ADSZ88]). Note that if we have found a non-viable KKT point with \( t_r = t_o + L \), then it could be a global minimum, a local minimum, or a saddle point; if there also exists a KKT point which is viable, then it lies at some \( t_r^* < t_r \).

Also note that if \( h > k \) then there is no KKT point with \( t_o + L < t_r \), as \( h > k \) implies that it is cheaper to go short than it is to ever hold stock.

Proof of Proposition 2-2: We show this with an example. Let \( F(t) \) be the uniform distribution on \([0.5,1.5] \); \( L = 0.4 \); \( c = 5 \); \( b = 5 \); \( k = 100 \); \( R = 0 \); \( a = 0 \); \( v(t) = 1 \). Arbitrarily choose \( t_o = 0.1 \) and \( t_r = 0.5 \) to get \( JCF = 10 \) regardless of \( h \).

\[
JCF(0, 0.5) = \frac{5 + 100h \int_{0.5}^{0.1} F(x) dx + h \int_{0.5}^{0.1} F(x) dx + 5F(0.5)}{5 + 0 + 0 + 0} = \frac{5 + 0 + 0 + 0}{0.5 + 0} = 10
\]

Thus regardless of \( h \), when using the joint approach the optimum \( JCF \leq 10 \). For the sequential approach, first find the sequential \( t_r \) at \( \partial RCF / \partial t_r = 0 \). This gives a sequential optimum \( t_r = 0.9 \); substitute this value into \( JCF \) to get the following

\[
JCF: RCF = \frac{5 + 100h \int_{t_o}^{t_r+L} F(x) dx + h \int_{t_o}^{t_r+L} F(x) dx + 5F(0.9)}{5 + 0 + 0 + 0} \geq \frac{7 + 0.4h}{0.9}
\]
Letting \( h \to \infty \) causes the sequential optimal \( JCF:RCF \to \infty \), while the joint optimal \( JCF \leq 10 \). So the ratio of \( JCF:RCF / JCF \to \infty \) also.

**Proof of Proposition 2-3:** Note that \( OCF \) is equivalent to \( JCF \) with \( c = b = a = 0 \). By Proposition 2-1 the optimal \( t_o \) is increasing in \( c, b, \) and \( a \), so a decrease in one cost will lead to a lower (or at least no greater) new optimal \( t_o \). Setting \( c = b = a = 0 \) makes \( JCF \) equivalent to \( OCF \) and lowers the new optimal \( t_o \); hence

\[
(t_o \text{ optimal for } OCF:RCF) \leq (t_o \text{ optimal for } JCF:RCF).
\]

**Proof of Proposition 2-4:** We show this with an example. Let \( F(t) \) be the uniform distribution on \([0.5, 1.5]\), so \( \mu = 1; L = 0.4; c = b = 5, k = 2, h = 0.5, R = 0, a = 0, v(t) = 1 \). From the proof of Theorem 2-4, we know that the sequential optimum has \( t_r = 0.9 \). Use this in evaluating \( OCF \) at both endpoints to find that

\[
OCF(0) = \frac{0 + 0.41}{0.82 + 0} = 0.500 \quad OCF(0.5) = \frac{0.16 + 0}{0.82 + 0.08} = 0.178 \quad (B-50), (B-51)
\]

Clearly \( t_o = 0.5 \) is the best endpoint to choose. Now evaluate \( SOF \):

\[
SOF = kL - h\mu = 2(0.4) - 0.5(1) = 0.3 \quad (B-52)
\]

\( SOF \)'s positive value indicates that the other endpoint \( t_o = 0 \) should be chosen.

**Proof of Theorem 2-4:** For this proof instead of using \( JED(t_o) \) which corresponds to \( JCF(t_o, t_r) \) we use a more general form corresponding to \( JOD(t_o, t_r) \), except that the time required to perform the replacement is required to be a deterministic value (rather than a random variable) equal to \( R \).

Take \( \partial JED / \partial t_o \) and rearrange terms to get

\[
\frac{\partial}{\partial t_o} JED = \frac{F(t_o + L)}{D} \left\{ cv'(t_o + L) + bz(t_o + L) + aq(t_o + L) \right. \right.
\]

\[
\left. \left\{ + \frac{kF(t_o + L) - JED(t_o)}{F(t_o + L)} \right. \right. \right.
\]

\[
\left. \left\{ + \frac{[k - JED(t_o)]f(t_o + L)}{[F(t_o + L)]^2} \right. \right. \right.
\]

When this is zero, take derivatives (\( \partial / \partial t_o \)) again to get

\[
\frac{\partial^2}{\partial t_o^2} JED = \frac{F(t_o + L)}{D} \left\{ cv''(t_o + L) + bz'(t_o + L) + aq'(t_o + L) \right. \right.
\]

\[
\left. \left\{ + \frac{[k - JED(t_o)]f'(t_o + L)}{[F(t_o + L)]^2} \right. \right. \right.
\]

\[
\left. \left\{ + \frac{[k - JED(t_o)]f(t_o + L)}{[F(t_o + L)]^2} \right. \right. \right.
\]
which is non-negative when \( F(t) \) is strictly IFR and \( t_o \) is viable. This means that any viable zero of \( \partial JED(t_o) / \partial t_o \) will be a minimum, and hence there can be only one minimum (refer to the proof of Theorem 2-3 to see more details of a similar argument). So \( JED \) is unimodal and pseudo-convex in \( t_o \).

**Proof of Theorem 2-5:** Take the derivative of ORS with respect to \( t_o \) and rearrange terms to get

\[
\frac{\partial}{\partial t_o} ORS = \frac{F(t_o + L) - F(t_o)}{D} \left\{ k - h \frac{\bar{F}(t_o + L)}{F(t_o + L) - F(t_o)} - ORS(t_o, t_s) \right\}
\]

When this equals zero, the second derivative will be

\[
\frac{\partial^2}{\partial t_o^2} ORS = \frac{F(t_o + L) - F(t_o)}{D} \left\{ \frac{h}{\bar{F}(t_o)} \right\}^2 \frac{\partial}{\partial t_o} \left( \frac{\bar{F}(t_o)}{F(t_o + L)} \right) \leq 0
\]

This means that any zero of \( \partial ORS / \partial t_o \) will be a maximum, and hence there can be only one maximum. The desired global minimum then must occur at one of the endpoints, either at \( t_o = 0 \) or \( t_o = \infty \). Of course if \( F(t) \) were IFR, then the second derivative would be non-negative, indicating the opposite result that any zero of \( \partial ORS / \partial t_o \) would be a minimum.

**Proof of Theorem 2-6:** Take the derivative of ORS with respect to \( t_s \) to get

\[
\frac{\partial}{\partial t_s} ORS(t_o, t_s) = \frac{L f(t_s)}{D} \left\{ -k + \frac{h}{L z(t_s)} - \frac{r}{L} + ORS(t_o, t_s) \right\}
\]

When this equals zero, the second derivative will be

\[
\frac{\partial^2}{\partial t_s^2} ORS(t_o, t_s) = \frac{L f(t_s)}{D} \left\{ \frac{\partial}{\partial t_s} \frac{h}{L z(t_s)} \right\} \geq 0
\]

This means that any zero of \( \partial ORS / \partial t_s \) will be a minimum, and hence there can be only one minimum.

**Proof of Proposition 2-5:** Proven simply by examining the signs of the derivatives of \( \partial ORS / \partial t_s \) and \( \partial ORS / \partial t_o \), by each of the four cost coefficients.

\[
\frac{\partial^2}{\partial t_s \partial t_o} ORS = -\frac{L f(t_s)}{D^2} \{ -1 \} > 0
\]
\[
\frac{\partial^2}{\partial \tau_0 \partial \tau} \text{ORS} = \frac{-L f(t_s)}{D^2} \left\{ \frac{D}{L z(t_s)} - \int_{t_o + L}^{t_s} \bar{F}(x) dx \right\} > 0 \quad (B-60)
\]

\[
\frac{\partial^2}{\partial \tau_0 \partial \tau} \text{ORS} = \frac{-L f(t_s)}{D^2} \left\{ \frac{1}{L} \left[ \mu + \int_{t_o}^{t_o + L} F(x) dx + L \bar{F}(t_s) + R \right] - \bar{F}(t_s) \right\} < 0 \quad (B-61)
\]

\[
\frac{\partial^2}{\partial \tau_0 \partial \tau} \text{ORS} = \frac{-L f(t_s)}{D^2} \left\{ \left[ \mu + \int_{t_o}^{t_o + L} F(x) dx + L \bar{F}(t_s) + R \right] \right\} < 0 \quad (B-62)
\]

\[
\frac{\partial^2}{\partial \tau_0 \partial \tau} \text{ORS} = \frac{[F(t_o + L) - F(t_o)]}{D^2} \left\{ -1 \right\} < 0 \quad (B-63)
\]

\[
\frac{\partial^2}{\partial \tau_0 \partial \tau} \text{ORS} = \frac{[F(t_o + L) - F(t_o)]}{D^2} \left\{ -\bar{F}(t_s) \right\} < 0 \quad (B-64)
\]

\[
\frac{\partial^2}{\partial \tau_0 \partial \tau} \text{ORS} = \frac{[F(t_o + L) - F(t_o)]}{D^2} \left\{ -\bar{F}(t_o + L) D - \int_{t_o + L}^{t_s} \bar{F}(x) dx \right\} < 0 \quad (B-65)
\]

\[
\frac{\partial^2}{\partial \tau_0 \partial \tau} \text{ORS} = \frac{[F(t_o + L) - F(t_o)]}{D^2} \left\{ \left[ \mu + \int_{t_o}^{t_o + L} F(x) dx + L \bar{F}(t_s) + R \right] \right\} > 0
\]
Proof of Theorem 2-7: Take the derivative of \( RS \) with respect to \( t_s \) to get

\[
\frac{\partial}{\partial t_s} RS(t_s) = \frac{Lf(t_s)}{D} \left\{ -k + \frac{h}{Lz(t_s)} - \frac{r}{L} + RS(t_s) \right\}.
\]  
(B-67)

When this equals zero, the second derivative will be

\[
\frac{\partial^2}{\partial t_s^2} RS(t_s) = \frac{Lf(t_s)}{D} \left\{ \frac{\partial}{\partial t_s} \frac{h}{Lz(t_s)} \right\} \geq 0.
\]  
(B-68)

This means that any zero of \( \partial RS / \partial t_s \) will be a minimum, and hence there can be only one minimum.

Proof of Proposition 2-6: Proven simply by examining the signs of the derivatives of \( \partial RS / \partial t_s \) by each of the four cost coefficients.

\[
\frac{\partial^2}{\partial t_s^2} RS = -\frac{Lf(t_s)}{D^2} \left\{ -1 \right\} > 0
\]  
(B-69)

\[
\frac{\partial^2}{\partial t_s \partial c} RS = -\frac{Lf(t_s)}{D^2} \left\{ \frac{D}{Lz(t_s)} - \int_0^t F(x) \, dx \right\} > 0
\]  
(B-70)

\[
\frac{\partial^2}{\partial t_s \partial r} RS = -\frac{Lf(t_s)}{D^2} \left\{ \frac{1}{L} \left[ \mu + L \overline{F}(t_s) + R \right] - \overline{F}(t_s) \right\} < 0
\]  
(B-71)

\[
\frac{\partial^2}{\partial t_s \partial k} RS = -\frac{Lf(t_s)}{D^2} \left\{ \left[ \mu + L \overline{F}(t_s) + R \right] - L \overline{F}(t_s) - R \right\} < 0
\]  
(B-72)

Proof of Theorem 2-8: First consider \( t_r \). Take \( \partial / \partial t_r \) and rearrange terms to get

\[
\frac{\partial}{\partial t_r} JTD = \frac{\overline{F}(t_r)}{D} \left\{ h + cv'(t_r) + bz(t_r) + aq(t_r) - JTD(t_x, t_o, t_r) \right\}
\]  
(B-73)

When this is zero, take derivatives ( \( \partial / \partial t_r \) ) again to get

\[
\frac{\partial^2}{\partial t_r^2} JTD = \frac{\overline{F}(t_r)}{D} \left\{ cv''(t_r) + bz'(t_r) + aq'(t_r) \right\} \geq 0
\]  
(B-74)

where the inequality is strict when \( F(t) \) is strictly IFR. This means that any zero of the first derivative will be a minimum, and hence there can be only one minimum; hence \( JTD(t_x, t_o, t_r) \) is unimodal and pseudo-convex in \( t_r \).
Take $\partial / \partial t_o$ and rearrange terms to get

$$\frac{\partial}{\partial t_o} JTD = \frac{F(t_x) - \bar{F}(t_o + L)}{D} \left\{ k - h \frac{\bar{F}(t_o + L)}{\bar{F}(t_x) - \bar{F}(t_o + L)} - JTD(t_x, t_o, t_r) \right\}$$  \hspace{1cm} (B-75)

When this is zero, take derivatives ($\partial / \partial t_o$) again to get

$$\frac{\partial^2}{\partial t_o^2} JTD = \left[ \frac{\bar{F}(t_x) - \bar{F}(t_o + L)}{D} \right] \left\{ h \frac{f(t_o + L)\bar{F}(t_x)}{\left[ \bar{F}(t_x) - \bar{F}(t_o + L) \right]^2} \right\} \geq 0$$  \hspace{1cm} (B-76)

where the inequality is strict when $F(t)$ is strictly IFR. This means that any zero of the first derivative will be a minimum, and hence there can be only one minimum; hence $JTD(t_x, t_o, t_r)$ is unimodal and pseudo-convex in $t_o$.

Take $\partial / \partial t_x$ and rearrange terms to get

$$\frac{\partial}{\partial t_x} JTD = \frac{t_o + L - t_x - L_x}{D} f(t_x) \left\{ -k + \frac{c_x u(t_x)}{t_o + L - t_x - L_x} + JTD(t_x, t_o, t_r) \right\}$$  \hspace{1cm} (B-77)

When this is zero, take derivatives ($\partial / \partial t_x$) again to get

$$\frac{\partial^2}{\partial t_x^2} JTD = \frac{f(t_x)}{D} \left\{ \frac{c_x u(t_x)[t_o + L - t_x - L_x] + c_x u(t_x)}{[t_o + L - t_x - L_x]} \right\} \geq 0$$  \hspace{1cm} (B-78)

where the inequality is strict when $F(t)$ is strictly IFR. This means that any zero of the first derivative will be a minimum, and hence there can be only one minimum; hence $JTD(t_x, t_o, t_r)$ is unimodal and pseudo-convex in $t_x$.

**Proof of Theorem 2-9:** First consider $t_w$. Take $\partial / \partial t_w$ and rearrange terms to get

$$\frac{\partial}{\partial t_w} JTR = \frac{F(t_w)G(t_w - t_o)}{D} \left\{ cv'(t_w) + bz(t_w) + aq(t_w) - k \right\}$$  \hspace{1cm} (B-79)

When this is zero, take derivatives ($\partial / \partial t_w$) again to get

$$\frac{\partial^2}{\partial t_w^2} JTR = \frac{F(t_w)G(t_w - t_o)}{D} \left\{ cv''(t_w) + bz'(t_w) + aq'(t_w) \right\} \geq 0$$  \hspace{1cm} (B-80)
where the inequality is strict when $F(t)$ is strictly IFR. This means that any zero of the first derivative will be a minimum, and hence there can be only one minimum, hence $JTR(t_x, t_o, t_r, t_w)$ is unimodal and pseudo-convex in $t_w$.

Next consider $t_r$. Take $\partial / \partial t_r$ and rearrange terms to get

$$
\frac{\partial}{\partial t_r} JTR = \frac{\overline{F}(t_r) G(t_r - t_o)}{D} \left\{ cv'(t_r) + bz(t_r) + aq(t_r) + h - JTR \right\} \quad (B-81)
$$

When this is zero, take derivatives ($\partial / \partial t_r$) again to get

$$
\frac{\partial^2}{\partial t_r^2} JTR = \frac{\overline{F}(t_r) G(t_r - t_o)}{D} \left\{ cv''(t_r) + bz'(t_r) + aq'(t_r) \right\} \geq 0 \quad (B-82)
$$

where the inequality is strict when $F(t)$ is strictly IFR. This means that any zero of the first derivative will be a minimum, and hence there can be only one minimum, hence $JTR(t_x, t_o, t_r, t_w)$ is unimodal and pseudo-convex in $t_r$.

Take $\partial / \partial t_x$ and rearrange terms to get

$$
\frac{\partial}{\partial t_x} JTR = \left[ t_o + L - t_x - L_x \right] f(t_x) \left\{ \frac{c_x u_x(t_x)}{[t_o + L - t_x - L_x]} - k + JTR \right\} \quad (B-83)
$$

When this is zero, take derivatives ($\partial / \partial t_x$) again to get

$$
\frac{\partial^2}{\partial t_x^2} JTR = \frac{f(t_x)}{D} \left\{ \frac{c_x u'(t_x)[t_o + L - t_x - L_x] + c_x u(t_x)}{[t_o + L - t_x - L_x]} \right\} \quad (B-84)
$$

where the inequality is strict when $F(t)$ is strictly IFR. This means that any zero of the first derivative will be a minimum, and hence there can be only one minimum, hence $JTR(t_x, t_o, t_r, t_w)$ is unimodal and pseudo-convex in $t_x$.

**Proof of Proposition 2-7:**
Part (a): proven by examining the signs of the derivatives of $\partial JTR / \partial t_w$ (shown in the proof of the previous theorem) by each of the cost coefficients. In this particular case, simple inspection of (B-79) gives the desired results.

Part (b): examine the signs of the derivatives of $\partial JTR / \partial t_r$ by each of the cost coefficients as follows.
\[
\frac{\partial^2}{\partial t_x \partial \bar{\alpha}} JTR = \frac{\overline{F}(t_r)G(t_r - t_o)}{D} \left\{ \frac{h}{D} - \frac{H(t_o, t_r)}{D} \right\} \geq 0 \tag{B-85}
\]

\[
\frac{\partial^2}{\partial t_x \partial \bar{\kappa}} JTR = \frac{\overline{F}(t_r)G(t_r - t_o)}{D} \left\{ -\frac{K(t_o, t_r)}{D} \right\} \leq 0 \tag{B-86}
\]

\[
\frac{\partial^2}{\partial t_x \partial \alpha_x} JTR = \frac{\overline{F}(t_r)G(t_r - t_o)}{D} \left\{ -\frac{1}{D} \int_0^t u(t) f(t) dt \right\} \leq 0 \tag{B-87}
\]

Part (c): examine the signs of the derivatives of \( \partial JTR/\partial t_x \) by each of the cost coefficients as follows.

\[
\frac{\partial^2}{\partial t_x \partial t_r} JTR = \frac{[t_o + L - t_x - L_x]}{D} f(t_x) \left\{ -1 + \frac{K(t_x, t_o, t_w)}{D} \right\} \leq 0 \tag{B-88}
\]

\[
\frac{\partial^2}{\partial t_x \partial \bar{\alpha}} JTR = \frac{[t_o + L - t_x - L_x]}{D} f(t_x) \left\{ \frac{u(t_x)}{[t_o + L - t_x - L_x]} + \frac{1}{D} \int_0^t u(t) f(t) dt \right\} \geq 0 \tag{B-89}
\]

\[
\frac{\partial^2}{\partial t_x \partial \bar{\alpha}_x} JTR = \frac{[t_o + L - t_x - L_x]}{D} f(t_x) \left\{ \frac{C(t_o, t_x, t_w)}{D} \right\} \geq 0 \tag{B-90}
\]

\[
\frac{\partial^2}{\partial t_x \partial \bar{\bar{\alpha}}} JTR = \frac{[t_o + L - t_x - L_x]}{D} f(t_x) \left\{ \frac{B(t_o, t_x, t_w)}{D} \right\} \geq 0 \tag{B-91}
\]

\[
\frac{\partial^2}{\partial t_x \partial \bar{\alpha}_x} JTR = \frac{[t_o + L - t_x - L_x]}{D} f(t_x) \left\{ \frac{A(t_o, t_x, t_w)}{D} \right\} \geq 0 \tag{B-92}
\]

\[
\frac{\partial^2}{\partial t_x \partial \bar{\alpha}} JTR = \frac{[t_o + L - t_x - L_x]}{D} f(t_x) \left\{ \frac{H(t_o, t_r)}{D} \right\} \geq 0 \tag{B-93}
\]

**Proof of Lemma 2-1:** In \( CB(t_o, s, t_r) \), \( t_o \) and \( s \) appear only in the shortage and holding cost terms; the other terms are constant with respect to these two decision variables. Thus for a given \( t_r \), we can optimize the expected cost rate with respect to \( t_o \) and \( s \) by optimizing the expected inventory cost per spare. We can do this using the results of [AK95], which is effectively what we re-state in this proposition.
Appendix C

Proof of Lemma 3-1: Take the derivative of \( C_j(t) \) with respect to \( t \) and rearrange terms to get

\[
\frac{\partial}{\partial t} C_j(t) = \frac{\bar{F}(t)}{D} \left\{ b z(t) + a q(t) - C_j(t) \right\}
\]

(C-1)

When this equals zero (i.e. when the term in braces equals zero), we find that the second derivative of \( C_j(t) \) with respect to \( t \) is

\[
\frac{\partial^2}{\partial t^2} C_j(t) = \frac{\bar{F}(t)}{D} \left\{ b z'(t) + a q'(t) \right\} \geq 0
\]

(C-2)

where the inequality is strict if \( F(t) \) is strictly IFR. Thus any interior extreme point \( t \) of \( C_j(t) \) is a unique minimum, i.e. at \( t_j : 0 < t_j < \infty \). If the term in braces in (C-1) never equals zero, then its sign indicates the optimal decision: if it is negative at \( t = \infty \), then repair only after failure; if it is positive at \( t = 0 \), then do not run the system (it is not viable).

Proof of Proposition 3-1:

Take the derivative of (3-10) by \( t \) to get

\[
\frac{\partial}{\partial t} C_m(t) = \frac{D_m N'_m - D'_m N_m}{D_m D'_m} = \frac{D'_m}{D_m} \left\{ \frac{N'_m}{D'_m} - C_m(t) \right\}
\]

(C-3)

which will equal zero when the term in braces equals zero. First work with the term \( D'_m \).

\[
\frac{\partial}{\partial t} D_m = \begin{cases} \sum_{i=0}^{n} \frac{m!}{[m-i]!i!} R^i \bar{F}(t)[m-i\left( \int_{0}^{t} \bar{F}(x)dx \right)^{m-i-1} \end{cases} + \sum_{i=n+1}^{m-1} \frac{m!}{[m-i]!n!n^{-n}} R^i \bar{F}(t)[m-i\left( \int_{0}^{t} \bar{F}(x)dx \right)^{m-i-1} \end{cases} \]

(C-4)
\[ \frac{\partial}{\partial t} D_m = m \bar{F}(t) \left\{ \sum_{i=0}^{n} \frac{[m-1]!}{[m-1-i]! i!} R^i \left( \int_0^t \bar{F}(x) \, dx \right)^{m-i-1} \right\} \] (C-5)

\[ \frac{\partial}{\partial t} D_m = m \bar{F}(t) D_{m-1} \] (C-6)

Similarly transform the \( N'_m \) term.

\[ \frac{\partial}{\partial t} N_m = \sum_{i=0}^{n} \frac{m!}{[m-i]! i!} R^i \bar{F}(t)[m-i] \left( \int_0^t \bar{F}(x) \, dx \right)^{m-i-1} [ik + (m-i)\text{rate}(t)] \]

\[ + \sum_{i=n+1}^{m-1} \frac{m!}{[m-i]! n! n^{i-n}} R^i \bar{F}(t)[m-i] \left( \int_0^t \bar{F}(x) \, dx \right)^{m-i-1} [ik + (m-i)\text{rate}(t)] + 0 \]

\[ + \sum_{i=0}^{n} \frac{m!}{[m-i]! i!} R^i \left( \int_0^t \bar{F}(x) \, dx \right)^{m-i} \left[ (m-i)\text{rate}'(t) \right] \]

\[ + \sum_{i=n+1}^{m-1} \frac{m!}{[m-i]! n! n^{i-n}} R^i \left( \int_0^t \bar{F}(x) \, dx \right)^{m-i} \left[ (m-i)\text{rate}'(t) \right] + 0 \] (C-7)

\[ \frac{\partial}{\partial t} N_m = m \bar{F}(t) \sum_{i=0}^{n} \frac{[m-1]!}{[m-1-i]! i!} R^i \left( \int_0^t \bar{F}(x) \, dx \right)^{m-i-1} [ik + (m-i)\text{rate}(t)] \]

\[ + m \bar{F}(t) \sum_{i=n+1}^{m-1} \frac{[m-1]!}{[m-1-i]! n! n^{i-n}} R^i \left( \int_0^t \bar{F}(x) \, dx \right)^{m-i-1} [ik + (m-i)\text{rate}(t)] \]

\[ + m \bar{F}(t) \sum_{i=0}^{n} \frac{[m-1]!}{[m-i]! i!} R^i \left( \int_0^t \bar{F}(x) \, dx \right)^{m-i-1} [-\text{rate}(t)(m-i) + (bz(t) + aq(t))(m-i)] \]

\[ + m \bar{F}(t) \sum_{i=n+1}^{m-1} \frac{[m-1]!}{[m-i]! n! n^{i-n}} R^i \left( \int_0^t \bar{F}(x) \, dx \right)^{m-i-1} (m-i)[-\text{rate}(t) + (bz(t) + aq(t))] \] (C-8)
\[
\frac{\partial}{\partial t} N_m = m \overline{F}(t) \sum_{i=0}^{n} \frac{[m - 1]!}{[m - 1 - i]! i!} \left[ ik + (m - 1 - i) \text{rate}(t) \right] \\
+ m \overline{F}(t) \sum_{i=n+1}^{n} \frac{[m - 1]!}{[m - 1 - i]! n! n^{i-n} \left[ i \int_0^t \overline{F}(x)dx \right]^{m-i-1}} \left[ ik + (m - 1 - i) \text{rate}(t) \right] \\
+ m \overline{F}(t) \sum_{i=0}^{n} \frac{[m - 1]!}{[m - i]! i!} \left[ i \int_0^t \overline{F}(x)dx \right]^{m-i-1} \left\{ b \mathbf{z}(t) + a \mathbf{q}(t) \right\} \\
+ m \overline{F}(t) \sum_{i=n+1}^{n} \frac{[m - 1]!}{[m - i]! n! n^{i-n} \left[ i \int_0^t \overline{F}(x)dx \right]^{m-i-1}} \left\{ b \mathbf{z}(t) + a \mathbf{q}(t) \right\}
\]

\[
\frac{\partial}{\partial t} N_m = m \overline{F}(t) N_{m-1} + m \overline{F}(t) \left[ b \mathbf{z}(t) + a \mathbf{q}(t) \right] D_{m-1} \tag{C-10}
\]

Substitution of (C-6) and (C-10) back into (C-3) gives

\[
\frac{\partial}{\partial t} C_m(t) = \frac{m \overline{F}(t) D_{m-1}}{D_m} \left\{ C_{m-1}(t) + b \mathbf{z}(t) + a \mathbf{q}(t) - C_m(t) \right\} \tag{C-11}
\]

which leads to the result of (3-13).

To get (3-14), begin with (3-9) and take the derivative by \( t \) to get

\[
\frac{\partial}{\partial t} C_m(t) = \text{rate}'(t) \left[ m - \sum_{i=0}^{m} i \ \text{prob}_i(t) \right] - \text{rate}(t) \sum_{i=0}^{m} i \ \text{prob}_i(t) + k \sum_{i=0}^{m} i \ \text{prob}_i(t) \tag{C-12}
\]

where

\[
\text{rate}'(t) = \left\{ \left[ \int_0^t \overline{F}(x)dx \right] \left[ b \mathbf{f}(t) + a \mathbf{q}(t) \overline{F}(t) \right] - \overline{F}(t) \left[ c + b \mathbf{F}(t) + a \int_0^t q(x) \overline{F}(x)dx \right] \right\} \\
\left[ \int_0^t \overline{F}(x)dx \right]^2
\]

\[
\left[ \int_0^t \overline{F}(x)dx \right]^{m-i-1}
\]
\[ \text{rate'}(t) = \frac{[bz(t) + aq(t) - \text{rate}(t)] \bar{F}(t)}{\int_0^t \bar{F}(x) \, dx} \]  

(C-14)

and

\[ \text{prob'}(t) = \frac{\text{prob}(t) \left[ -i + \sum_{j=0}^{m} j \, \text{prob}_j(t) \right] \bar{F}(t)}{\int_0^t \bar{F}(x) \, dx} \]  

(C-15)

Substitute (C-14) and (C-15) into (C-12) and rearrange terms to get

(C-16)

\[ \frac{\partial}{\partial t} \mathcal{C}_m(t) = \frac{\left[ \begin{array}{c}
[bz(t) + aq(t) - \text{rate}(t)] [m - \sum_{i=0}^{m} i \, \text{prob}_i(t)] \\
- [\text{rate}(t) - k] \sum_{i=0}^{m} i \, \text{prob}_i(t) [\sum_{j=0}^{m} j \, \text{prob}_j(t)]
\end{array} \right]}{\int_0^t \bar{F}(x) \, dx} \]

where the double summation in the numerator equals \( \text{Var}[\# \text{ machines up}] \). Thus we have the result that (C-12) equals zero if and only if (3-14) is true.

**Proof of Lemma 3-2:** Take the derivative of \( C_2(t) \) with respect to \( t \) as shown in (C-11).

\[ \frac{\partial}{\partial t} C_2(t) = \frac{2 \bar{F}(t) D_1}{D_2} \left\{ C_1(t) + bz(t) + aq(t) - C_2(t) \right\} \]  

(C-19)

This equals zero only when the term in braces equals zero. We know that \([bz(t) + aq(t)]\) is increasing \( \forall t \), and from Lemma 3-1 we know that \( C_2(t) \) is increasing \( \forall t > t_j \); so together \([C_2(t) + bz(t) + aq(t)]\) is increasing \( \forall t > t_j \). Thus for any potential minimum point \( t_2 > t_j \).
\[ \frac{\partial^2}{\partial t^2} C_2(t) = \frac{2F(t)D_1}{D_2} \left\{ \frac{\partial}{\partial t} \left[ C_1(t) + bz(t) + aq(t) \right] - 0 \right\} \geq 0 \quad \forall t_2 > t_1 \]

which means there can be at most one minimum point \( t_2 > t_1 \). We also know from part (a) that \( C_1(t) < \left[ bz(t) + aq(t) \right] \forall t < t_1 \), and clearly we must have \( C_2(t) > 2C_1(t) \); together these imply that

\[ \frac{\partial}{\partial t} C_2(t) = \frac{2F(t)D_1}{D_2} \left\{ C_1(t) + bz(t) + aq(t) - C_2(t) \right\} \]

\[ < \frac{2F(t)D_1}{D_2} \left\{ C_1(t) + C_1(t) - C_2(t) \right\} < 0 \quad \forall t < t_1 \]

which means there cannot be a minimum \( t_2 < t_1 \). Thus any interior extreme point of \( C_2(t) \) must be a unique minimum, i.e. at \( t_2 > t_1 \). Note that if the term in braces in (C-19) never equals zero, then its sign indicates the optimal decision: if it is negative at \( t = \infty \), then repair only after failure; if it is positive at \( t = 0 \), then do not run the system (it is not viable).

**Proof of Proposition 3-2:** Begin by letting \( m = n \), so that there is one repair shop for each machine. Clearly there will be no interference or delay waiting for an available shop, so each machine will be down a proportion of time \( q_m \).

Next add one extra machine to the system, for a total of \( m+1 \). Since the machines are identical and the repair times are Markovian, let us give the original \( m \) machines priority for entering the repair shop, i.e. they can preempt the repair of machine \( m+1 \). This prioritization does not affect the system performance overall, though it does affect the performance of individual machines. The original \( m \) machines will still be down a proportion of time \( q_m \), but machine \( m+1 \) must occasionally wait for a shop and therefore will be down a proportion of time \( q_{m+1} > q_m \).

In a similar fashion add another machine to the system (total of \( m+2 \)), giving it the lowest priority for repair (so that it can be preempted by any of the other \( m+1 \)). Then machine \( m+2 \) will face a larger expected wait for repairs and will be down a proportion of time \( q_{m+2} > q_{m+1} > q_m \).

Now note that

\[
E[MD(t)_m] = q_1 + q_2 + \ldots + q_m \\
E[MD(t)_{m+1}] = q_1 + q_2 + \ldots + q_m + q_{m+1} \\
E[MD(t)_{m+2}] = q_1 + q_2 + \ldots + q_m + q_{m+1} + q_{m+2}.
\]
Subtraction gives the desired results that the expected number of machines down is increasing in $m$

$$E[MD(t)_{m+1}] - E[MD(t)_m] = q_{m+1} \quad (C-22)$$

at a non-decreasing rate

$$\{ E[MD(t)_{m+2}] - E[MD(t)_{m+1}] \} - \{ E[MD(t)_{m+1}] - E[MD(t)_m] \} = q_{m+2} - q_{m+1} \geq 0 \quad (C-23)$$

**Proof of Theorem 3-1:** Take the derivative of $C_{m+2}(t)$ with respect to $t$ and rearrange terms to get

$$\frac{\partial}{\partial t} C_{m+2}(t) = \frac{(m + 2) \bar{F}(t) D_{m+1}^{m+2}}{D_{m+2}} \left\{ C_{m+1}(t) + bz(t) + aq(t) - C_{m+2}(t) \right\} \quad (C-24)$$

When this equals zero (i.e. when the term in braces equals zero), we find that the second derivative of $C_{m+2}(t)$ with respect to $t$ is

$$\frac{\partial^2}{\partial t^2} C_{m+2}(t) = \frac{(m + 2) \bar{F}(t) D_{m+1}^{m+2}}{D_{m+2}} \left\{ \frac{(m + 1) \bar{F}(t) D_m}{D_{m+1}} \left[ C_m(t) + bz(t) + aq(t) \right] - C_{m+1}(t) \right\} \quad (C-25)$$

which by substitution of terms from the braces in the first derivative (C-24) becomes

$$\frac{\partial^2}{\partial t^2} C_{m+2}(t) = \frac{(m + 2) \bar{F}(t) D_{m+1}^{m+2}}{D_{m+2}} \left\{ \frac{(m + 1) \bar{F}(t) D_m}{D_{m+1}} \left[ C_m(t) - C_{m+1}(t) + C_{m+2}(t) \right] \right\} + bz'(t) + aq'(t) \quad (C-26)$$

We need this to be positive; we already know that $z'(t)$ and $q'(t)$ are both positive, so we need only worry about the term in square brackets. Proposition 3-1.1 shows that

$$\{ E[MD(t)_{m+2}] - E[MD(t)_{m+1}] \} - \{ E[MD(t)_{m+1}] - E[MD(t)_m] \} \geq 0 \quad (C-27)$$

Combine this with (3-9) to find that for any viable $t$
\[ C_m(t) - 2C_{m+1}(t) + C_{m+2}(t) = \text{rate}(t) \left[ m - \sum_{i=0}^{m} i \, \text{prob}_{i,m}(t) \right] + k \sum_{i=0}^{m} i \, \text{prob}_{i,m}(t) \]

\[ - 2\text{rate}(t) \left[ m + 1 - \sum_{i=0}^{m+1} i \, \text{prob}_{i,m+1}(t) \right] - 2k \sum_{i=0}^{m+1} i \, \text{prob}_{i,m+1}(t) \]

\[ + \text{rate}(t) \left[ m + 2 - \sum_{i=0}^{m+2} i \, \text{prob}_{i,m+2}(t) \right] + k \sum_{i=0}^{m+2} i \, \text{prob}_{i,m+2}(t) \]

(C-29)

\[ C_m(t) - 2C_{m+1}(t) + C_{m+2}(t) = [k - \text{rate}(t)] \left[ \sum_{i=0}^{m} i \, \text{prob}_{i,m}(t) \right] + (m)\text{rate}(t) \]

\[ - 2[k - \text{rate}(t)] \left[ \sum_{i=0}^{m+1} i \, \text{prob}_{i,m+1}(t) \right] - 2(m + 1)\text{rate}(t) \]

\[ + [k - \text{rate}(t)] \left[ \sum_{i=0}^{m+2} i \, \text{prob}_{i,m+2}(t) \right] + (m + 2)\text{rate}(t) \]

C_m(t) - 2C_{m+1}(t) + C_{m+2}(t) =

\[ [k - \text{rate}(t)] \left\{ E[MD(t)_m] - 2E[MD(t)_{m+1}] + E[MD(t)_{m+2}] \right\} > 0 \] (C-30)

This result can be combined with (C-26) to show that the second derivative must be positive. Thus any viable interior extreme point \( t \) of \( C_m(t) \) is a unique minimum, i.e. at \( t_m : 0 \leq t_m \leq \infty \). If the term in braces in (C-24) never equals zero, then its sign indicates the optimal decision: if it is negative at \( t = \infty \), then repair only after failure; if it is positive at \( t = 0 \), then do not run the system (it is not viable).

That \( t_{m+1} \leq t_{m+2} \) follows from a similar line of reasoning. At \( t_{m+1} \), the term in braces in (C-24) equals

\[ \{ C_{m+1}(t_{m+1}) + b z(t_{m+1}) + a q(t_{m+1}) - C_{m+2}(t_{m+1}) \} \] (C-31)

or

\[ \{ C_{m+1}(t_{m+1}) + C_{m+1}(t_{m+1}) - C_m(t_{m+1}) - C_{m+2}(t_{m+1}) \} \] (C-32)
which by the same argument of (C-27) to (C-30) must be negative. Since \( C_m(t) \) is pseudo-convex, it must be that \( t_m+1 \leq t_{m+2} \).

**Proof of Proposition 3-3:**

Part (a): Begin with the derivative shown in (C-24), and substitute in (3-8). We then get the stated results by finding the signs of the respective partial derivatives of \( C'_m(t) \) with respect to the cost coefficients \( k, c, a, \) and \( b \).

\[
\frac{\partial^2}{\partial \theta \partial k} C_m(t) = \frac{m \bar{F}(t) D_{m-1}}{D_m} \left\{ \sum_{i=0}^{m-1} \langle i \rangle \text{prob}_{i,m-1}(t) - \sum_{i=0}^{m} \langle i \rangle \text{prob}_{i,m}(t) \right\} \leq 0 \quad (C-33)
\]

\[
\frac{\partial^2}{\partial \theta \partial c} C_m(t) = \frac{m \bar{F}(t) D_{m-1}}{D_m} \left\{ \sum_{i=0}^{m-1} \langle m - 1 - i \rangle \text{prob}_{i,m-1}(t) \right\} \leq 0 \quad (C-34)
\]

\[
\frac{\partial^2}{\partial \theta \partial a} C_m(t) = \frac{m \bar{F}(t) D_{m-1}}{D_m} \left\{ \sum_{i=0}^{m-1} \langle m - 1 - i \rangle \text{prob}_{i,m-1}(t) \frac{F(t)}{\int_0^t \bar{F}(x) dx} \right\} \quad (C-35)
\]

\[
\frac{\partial^2}{\partial \theta \partial b} C_m(t) = \frac{m \bar{F}(t) D_{m-1}}{D_m} \left\{ \sum_{i=0}^{m-1} \langle m - 1 - i \rangle \text{prob}_{i,m-1}(t) \frac{F(t)}{\int_0^t \bar{F}(x) dx} \right\} \quad (C-36)
\]
Part (b): First show that these expectations are decreasing in \( t \). We use \( y_i \) as defined for (3-10), so as to express the expected number of machines down or idle as

\[
\begin{align*}
E[\text{# machines down}] &= \sum_{i=0}^{m} i y_i \\
E[\text{# machines waiting}] &= \sum_{j=0}^{m} j y_j
\end{align*}
\]  

Take the derivative of the first equation above with respect to \( t \) to get

\[
\frac{\partial}{\partial t} \sum_{i=0}^{m} i y_i = \left[ \sum_{i=0}^{m} y_i \right] \left[ \sum_{j=0}^{m} j y_j' \right] - \left[ \sum_{i=0}^{m} i y_i \right] \left[ \sum_{j=0}^{m} y_j \right] \\
\frac{\partial}{\partial t} \sum_{j=0}^{m} j y_j = \left[ \sum_{j=0}^{m} y_j \right]^2
\]  

(C-37)
Since the denominator is always positive and

$$y_i' = y_i (m - i) \int_0^t \frac{\bar{F}(x)}{F(x)} dx$$  

(C-42)

then the sign of the derivative in (C-41) is the same as

$$\text{sign of } \left( \sum_{i=0}^{m} y_i \right) \left( \sum_{j=0}^{m} (m - j) y_j \right) - \left( \sum_{i=0}^{m} (m - i) y_i \right) \left( \sum_{j=0}^{m} y_j \right)$$

(C-43)

$$= \text{sign of } \sum_{i=0}^{m} \sum_{j=0}^{i} y_i y_j [(m - j) j - (m - i) j] < 0$$

Since the derivative is always negative, we have the result that the expected number of machines down is decreasing in $t$. A similar derivation provides the same conclusion for the number of machines awaiting repair, i.e. for (C-40) instead of (C-39).

For $R$ and $n$, it should be clear that the stated relations hold for any given age $t$ policy; increasing the number of shops $n$ or decreasing the average repair time $R$ could never lead to an increase in the expected number of machines awaiting maintenance.

Part (c): Our conclusion that $\text{ratio}_m = \frac{C_m(t)}{mC_1(t)}$ is increasing in $m$ and decreasing in $n$ is based upon the simple notion that $C_m(t) > C_{m-1}(t) + C_1(t)$ for given $n$. Hence

$$\text{ratio}_m = \frac{C_m(t)}{mC_1(t)} > \frac{C_{m-1}(t) + C_1(t)}{mC_1(t)} > \cdots > \frac{mC_1(t)}{mC_1(t)}$$  

(C-44)

Proof of Theorem 3-2:

Consider a multi-machine system, in which the repair shops are all busy. We wish to know until what age $t_w$ we should wait before we shut down a functioning machine, whose current age is $T$. We can consider each machine individually from now until whenever the first shop becomes available, because until then there is nothing linking them; further, we need only consider a machine's expected costs from now until that shop becomes free, because our actions do not affect costs beyond that point. This myopic expected cost $K(t_w)$ is given by
\[
\overline{F}(T)K(t_w) = \int_0^\Delta \left\{ \int_T^{T+\Delta} \left[ b + k(T + y - x) + a \int_T^{T+\Delta} q(v)dv \right] f(x)dx \right\} \gamma(y)dy
\]

\[
+ \int_0^\Delta \left\{ \int_T^{T+\Delta} q(v)dv \overline{F}(T + y) \right\} \gamma(y)dy
\]

\[
+ \int_\Delta^\infty \left\{ \int_T^{T+\Delta} \left[ b + k(T + y - x) + a \int_T^{T+\Delta} q(v)dv \right] f(x)dx \right\} \gamma(y)dy
\]

\[
+ \int_\Delta^\infty \left\{ \left[ k(y - \Delta) + a \int_T^{T+\Delta} q(v)dv \right] \overline{F}(T + \Delta) \right\} \gamma(y)dy
\]

where \( \Delta = t_w - T \). Take the derivative with respect to \( t_w \) and cancel terms to get

\[
\frac{\partial}{\partial t_w} K(t_w) = \overline{F}(t_w) \overline{\Gamma}(\Delta) \left[ bz(t_w) + aq(t_w) - k \right] \quad \text{(C-46)}
\]

When this is zero take derivatives again to get

\[
\frac{\partial^2}{\partial t_w^2} K(t_w) = \overline{F}(t_w) \overline{\Gamma}(\Delta) \left[ bz'(t_w) + aq'(t_w) \right] \geq 0 \quad \text{(C-47)}
\]

This is our desired result. That \( K(t_w) \) is pseudo-convex follows from \( z(t) \) and \( q(t) \) being increasing. Note that this result means that the optimal \( t_w \) is independent of \( R \) and our choice of maintenance policy; we should actually find \( t_w \) first, and then somehow choose the rest of the maintenance policy so that any other shut-down times are less than \( t_w \).
Appendix D

AMPL Formulation of Discrete Time Model

Following is the AMPL code which describes the linear programming formulation of the discrete time model used by the first numerical study in Chapter 3. The particular formulation shown is for a system of 2 machines and 1 repair shop for which we wish to solve for the optimal policy. This formulation also serves as the basis for the testing of heuristic policies, in that each heuristic involves adding more constraints to force the solution to have the required form.

```AMPL
# failmod1.txt
# MJ Armstrong 8 July 1995
# AMPL LP model for discrete time MDP two machines one shop general solution
# requires data file or manual entry to specify size, fail vector, short, rep
# specifies coefficients by columns rather than rows
#*****************************************************************************

parameter size integer, >0; # number of time slices for failure distribution

set AGES := -l..(size+l); # -1 for machine broken, positive for age
# last one is dummy

parameter fail {0..size} >=0, <= 1; # failure distribution

parameter short >0; # relative shortage cost

parameter rep >0, <=1; # repair probability per period

minimize cost; # objective

subject to Balance {i in AGES, j in AGES : j <= i}: tocome = 0;

subject to Normalization : to_come = 1;

variable M {i in 0..size, j in 0..size : j <= i} >= 0, <=1,
# stationary probability for doing preventive maintenance

objective cost ( short * rep * (1-fail[j])
+ (1+short) * rep * fail[j]
+ short * (1-rep) * (1-fail[j])
+ (1+short) * (1-rep) * fail[j] ),
```

#*****************************************************************************
coefficient Normalization 1,
coefficient Balance[i,j] 1,
coefficient Balance[j+1,0] - rep * (1-fail[j]),
coefficient Balance[0,-1] - rep * fail[j],
coefficient Balance[j+1,-1] - (1-rep) * (1-fail[j]),
coefficient Balance[-1,-1] - (1-rep) * fail[j];

variable N {i in 0..size, j in 0..size : j <= i} >= 0,
   # stationary probability for not doing preventive maintenance

objective cost ( 0
   + fail[i]*(1-fail[j])
   + (1-fail[i])*fail[j]
   + 2* fail[i] * fail[j] ),

coefficient Normalization 1,
coefficient Balance[i,j] 1,
coefficient Balance[i+1,j+1] - (1-fail[i]) * (1-fail[j]),
coefficient Balance[j+1,-1] - fail[i] * (1-fail[j]),
coefficient Balance[i+1,-1] - (1-fail[i]) * fail[j],
coefficient Balance[-1,-1] - fail[i] * fail[j];

variable M1 {i in 0..size} >= 0, <=1,   # stands for M[i,-1]

objective cost ( 2*short),

coefficient Normalization 1,
coefficient Balance[i,-1] 1,
coefficient Balance[0,-1] - rep,
coefficient Balance[-1,-1] - (1-rep);

variable N1 {i in 0..size} >= 0, <=1,   # stands for N[i,-1]

objective cost ( fail[i] + short ),

coefficient Normalization 1,
coefficient Balance[i,-1] 1,
coefficient Balance[i+1,0] - (1-fail[i]) * rep,
coefficient Balance[i+1,-1] - (1-fail[i]) * (1-rep),
coefficient Balance[0,-1] - fail[i] * rep,
coefficient Balance[-1,-1] - fail[i] * (1-rep);
variable N11 >= 0, <=1,  \hspace{1em} \# stands for N[-1,-1] (no M[-1,-1] needed)

objective cost (2*short),

coefficient Normalization 1,
coefficient Balance[-1,-1] 1,
coefficient Balance[0,-1] - rep,
coefficient Balance[-1,-1] - (1-rep);

\# end of model
\# ************************************************************