Markovian queue, controllable service system, random shocks, optimum repair rate
Abstract
Motivated by the need to study traffic incident management, we consider a Markovian infinite server queue that is subjected to randomly occurring shocks. These shocks affect the service of all servers to deteriorate, i.e., increase the service time of all servers, and they might also cause other shocks, thus cause further service deterioration. There are a finite number of service levels, zero being the normal service with the highest service rate and last being the worst with the slowest service rate which could even be equal to zero, implying the complete service breakdown. The repair process is performed only at the last level. These types of queues also represent an approximation to multi-server callcenters with deteriorating service. We derive the mean and variance of the stationary number in the system, and show that the mean is convex with respect to the repair rate. Furthermore, we study the optimal repair rate that minimizes the expected long-run average cost incurred due to delay and repairs. We show that the expected total cost per unit time as a function of repair rate is unimodal. We derive conditions under which the cost function is in one of three simple forms, so that the optimum repair rate can easily be obtained. Numerical examples are also provided.

1. Introduction
We consider the maintenance of a Markovian infinite server queue with deteriorating service. The literature on maintenance of degradable systems is vast. Kazaz and Sloan (2013) are concerned with a single-stage production system that deteriorates with production activities and improves through maintenance. The authors consider a four-state deteriorating system with two products and two maintenance types. Then, they find
the best policy (produce or do maintenance, in the former case which product to pro-
duce and in the latter case which type of maintenance to carry out) for each state in
order to maximize average reward per unit time. Le and Tan (2013) consider an oper-
ating system that suffers from degradation, whose level can only be known through a
costly inspection procedure. The aim of the paper is to find an inspection-maintenance
scheme which reduces unnecessary inspection without deteriorating the system reliabil-
ity in such a way that the mean long-run cost of the system is minimized. Kumar et al.
(2013) study system availability of a system in which the degradation rate increases with
the aging process. They develop a semi-Markov model of which its steady-state solution
provides system availability. Xiang et al. (2012) approximate time to failure distribu-
tion of a single component system as a Weibull distribution through simulation. Then,
they use this distribution to find the optimal schedule based maintenance policy and
compare it with their alternative condition-based policy using a simulation optimization
approach. Golmakani and Moakedi (2012) investigate a repairable system with several
components, each of which is subjected to soft failures. Soft failures cause system degra-
dation, in turn increase the system operating costs and are detected only if inspection
is performed. Thus, it becomes important to devise an optimal inspection scheme that
will minimize the systems expected total cost including inspection costs, repair costs,
and the penalty costs due to fault detection delay. The authors propose an approach,
in which the expected cost is first formulated heuristically in terms of the inspection
scheme. Then, a search algorithm is used to identify the optimal inspection scheme.

On the other hand, study of queueing systems with randomly occurring systemwide
service failure has found great interest from variety of application areas such as produc-
tion and manufacturing systems, inventory/ supply chains, and other service systems
such as traffic, and call centers. Researchers develop several models concentrated on
such systems with different variations including single-server queue with random ser-
vice breakdown (White and Christie (1958); Gaver (1962); Keilson (1962); Avi-Itzhak
and Naor (1963); Halfin (1972); Fischer (1977); Federgruen and Green (1986, 1988);
Sengupta (1990); Takine and Sengupta (1997); Economou and Kapodistria (2010)) and
single-server queue with random deteriorating service (Boxma and Kurkova (2000, 2001); Bekker and Boxma (2007); Boxma et al. (2008)); multi-server queue with random service breakdown (Mitran and Avi-Itzhak (1968); Neuts (1978, 1981); O’Cinneide and Purdue (1986); Keilson and Servi (1993); Jayawardene and Kella (1996); Baykal-Gürsoy and Xiao (2004); Baykal-Gürsoy and Duan (2006, 2013); D’Auria (2005, 2007); Falin (2008); Pang and Whitt (2009a); Yechiali (2007); Laghaie et al. (2012)) and multi-server queue with random deteriorating service (Purdue (1973); Yechiali (1973); Neuts (1978, 1981); O’Cinneide and Purdue (1986); Keilson and Servi (1993); Baykal-Gürsoy and Xiao (2004); Baykal-Gürsoy and Duan (2006, 2013); D’Auria (2005, 2007); Falin (2008); Pang and Whitt (2009a); Yechiali (2007)).

Most of the articles on deteriorating service is focused on the two server state case. Contrary to the vacation type queues, single-server and multi server queues in random environment (even in two-state) do not exhibit the decomposition property (Fuhrmann and Cooper (1985)). However, decomposition property has been shown to be valid for the infinite server queues, with server breakdowns (Jayawardene and Kella (1996)), and with server deterioration (Baykal-Gürsoy and Xiao (2004); D’Auria (2007)). The validity of the decomposition property for multiple levels of server deterioration has not been proven yet. In addition, there are only a few articles on the performance analysis of many server queues with deteriorating service. O’Cinneide and Purdue (1986) study the steady state behavior of an M/M/∞ queue in multi-state random environment and present a matrix equality to obtain the factorial moments. D’Auria (2008) proposes a recursion to calculate the factorial moments of the stationary number of customers in an M/M/∞ system in random environment.

Moreover, the maintenance of queues with random service failures has received even much less attention (Laghaie et al. (2012)), except in the single-server case (Federgruen and So (1989); Kaufman and Lewis (2007); Yang et al. (2009, 2011); Chakravarthy (2012)). Federgruen and So (1989, 1990) consider a single-server queue that is subject breakdowns and model the system as a semi-Markov decision process. Federgruen and So (1989, 1990) show that under some conditions the optimal repair policy is a threshold
type (bang-bang) policy that repairs the server only if there are more customers than the threshold level, otherwise the repair is not performed. Kaufman and Lewis (2007), on the other hand study the single-server queue with deteriorating service rates and failures. It is assumed that the deterioration states are finitely many, the highest state represents the normal operating conditions and the lowest state is the breakdown state. The authors show that the optimal maintenance policies for the replacement modal are threshold type depending on the deterioration state, i.e., replace the machine only if the server state is below a threshold level otherwise do not replace. However, the authors give a counter example that demonstrates that this may not always be true for the repair model. Yang et al. (2009) investigate the same system and impose a (deterioration level) threshold type repair policy. The authors use matrix-geometric approach to derive the long-run average cost including holding and repair costs, of the system for each threshold level. They give numerical examples to show some properties of the cost function. Chakravarthy (2012) consider a single deteriorating server system with Markovian arrivals. Repairs can be performed both as a preventive repair at some service level or after the server breaks down. The authors analyze the system in steady-state to obtain first order performance measures using matrix-geometric approach and study an optimization problem giving numerical examples. Laghaie et al. (2012) analyze a multi-machine deterministic production system with random machine breakdowns. Repairs are performed at a multi-server repair facility. The authors use search methods to find the optimum number of servers to assign to the repair facility in order to minimize the production, delay, lost sale and repair costs.

In this article, we investigate the maintenance policy for many server service systems subjected to random shocks that deteriorate the service time, thus cause time delay and congestion. Although the system in mind is coming from the transportation literature, such queues also arise in other service systems, for example, call-centers (Pang and Whitt (2009b)), communication systems or production systems. Traffic flow interrupted by roadway incidents is modeled as a many server queue operating in a multi-state Markovian environment that affects the service process (Baykal-Gürsoy and Xiao (2004);
Baykal-Gürsoy and Duan (2006); Baykal-Gürsoy et al. (2009b,a)). An “incident is defined here as any occurrence that affects capacity of the roadway (Skabardonis et al. (1998)), thus affecting the service to deteriorate, and it may be due to spilled loads, hazardous materials, work zones, traffic accidents, disabled vehicles, or natural causes (e.g., adverse weather conditions).

The system we consider in this study may be best explained through the following example. Suppose that we have two failure levels in the system explained above. Service state 0 corresponds to the state in which the system works normally. As soon as a shock occurs caused by adverse weather conditions, a disabled vehicle or spilled load, the state of service changes to state 1 and all servers slow down. If a further shock happens such as a fatal accident, the state changes to state 2 and servers slow down even more and accident clearance (repair) process starts. We obtain the first two factorial moments of the stationary number of customers in the system. A critical assumption in our model is that we cannot directly control arrival, service and failure rates but we can determine the repair rate. Note that since vehicles slow down when failures occur, the number of vehicles in the system increases, thus the overall delay cost increases as well. By increasing the repair rate, we can limit the number of vehicles in the system. The increase in repair rate is accomplished by decreasing the clearance time of an incident, in turn, this is achieved by allocating more first responders, locating first response centers close to incident sites, etc. However, since repair costs increase with repair rate, the optimal repair rate should be able to obtain a good tradeoff between these two cost components. We show that the expected number of customers in steady-state is convex with respect to the repair rate. Furthermore, we study the optimal repair rate that minimizes the expected long-run average cost incurred due to delay and repairs. We show that the expected total cost per unit time as a function of repair rate is unimodal. We derive conditions under which the cost function is in one of three simple forms, so that the optimum repair rate can easily be obtained. Numerical examples are also provided.

The rest of the paper is organized as follows: The system description and its math-
emathematical model are presented in Section 2 together with some performance measures such as the expected number and variance of the stationary number of customers in the system. Section 3 is devoted to determining the optimal repair rate. Finally, conclusions and future research directions are discussed in Section 4.

2. Mathematical Model and System Analysis

The underlying stochastic process can be modeled as \( \{ X(t), U(t) \} \) where \( t \) is the system time, \( X(t) \) is the number of customers in the system at time \( t \) and \( U(t) \) is the service state of the system at time \( t \). The service states are divided into \( K + 1 \) levels where \( K \) is a positive integer. State 0 represents the system when it functions normally. When a shock occurs, system transits to state 1. In general, if a shock occurs in state \( k \), \( k = 0, \ldots, K - 1 \), then state becomes \( k + 1 \). \( K \) represents the highest failure level. Once the system enters this state, repair is deployed and system returns to state 0.

The number of servers are assumed to be infinite. Arrival of customers is a Poisson process with rate \( \lambda \) and service interruptions are distributed exponentially with rate \( f_k \) for states \( k = 0, \ldots, K - 1 \). Arrival and interruption processes are independent of each other. Repair process is also distributed exponentially with rate \( f_K \). The service times are i.i.d. exponentially distributed with rate \( \mu_k \) at state \( k \), \( k = 0, \ldots, K \) and when an interruption happens, service rate becomes worse with respect to the previous state. Hence, the relationship between service rates is \( \mu_{k+1} < \mu_k \), for \( k = 0, \ldots, K - 1 \). Repair and service processes are independent of each other. All the rates mentioned are assumed to be positive. For convenience and ease of demonstration, also let index \(-1\) correspond to index \( K \) and index \( K + 1 + k \) correspond to index \( k \), for \( k = 0, \ldots, K \).

2.1. Formulation

Let \( P_{ik} \) denote the probability of the system having \( i \) customers \( i = 0, 1, \ldots \) and being in state \( k \), \( k = 0, \ldots, K \). Boundary equations for the mathematical model described above are as follows:
\[(\lambda + f_k)P_{0k} = \mu_k P_{1k} + f_{k-1}P_{0,k-1} \quad k = 0, \ldots, K \quad (2.1)\]

Also, for \(i \geq 1\), we have the following global balance equations:

\[(\lambda + i\mu_k + f_k)P_{ik} = \lambda P_{i-1,k} + f_{k-1}P_{i,k-1} + (i + 1)\mu_k P_{i+1,k} \quad k = 0, \ldots, K \quad (2.2)\]

Let \(G(z)\) be the probability generating function of this system defined as

\[G(z) = \sum_{k=0}^{K} G_k(z) \quad (2.3)\]

where \(G_k(z)\) represents the partial generating functions for each service state \(k = 0, \ldots, K\), as \(G_k(z) = \sum_{i=0}^{\infty} P_{ik} z^i\). Multiplying each side of (2.2) by \(z^i\) and summing over \(i\) gives

\[(\lambda + f_k) \sum_{i=1}^{\infty} P_{ik} z^i + \mu_k \sum_{i=1}^{\infty} iP_{ik} z^i = \lambda \sum_{i=1}^{\infty} P_{i-1,k} z^i + f_{k-1} \sum_{i=1}^{\infty} P_{i,k-1} z^i + \mu_k \sum_{i=1}^{\infty} (i+1)P_{i+1,k} z^i \quad (2.4)\]

for \(k = 0, \ldots, K\).

Using the definitions in (2.3), we get

\[(\lambda + f_k)(G_k(z) - P_{0k}) + \mu_k zG_k'(z) = \lambda zG_k(z) + f_{k-1}(G_{k-1}(z) - P_{0,k-1}) + \mu_k (G_k'(z) - P_{1k}) \quad (2.5)\]

for \(k = 0, \ldots, K\).

Finally, we obtain the following differential equations by simplifying the above equalities and using (2.1):

\[\mu_k (z-1)G_k'(z) = (\lambda z - \lambda - f_k)G_k(z) + f_{k-1}G_{k-1}(z) \quad (2.6)\]

for \(k = 0, \ldots, K\).
2.2. Analysis

Renewal arguments provide the following fact about the stationary distribution of each service level \( k \), for \( k = 0, \ldots, K \). Note that the expected cycle time, \( E[C] \) can be calculated as follows:

\[
E[C] = \sum_{k=0}^{K} \frac{1}{f_k}
\]  

(2.7)

**Theorem 2.1.** The probability of being in failure level \( k \), \( k = 0, \ldots, K \) is

\[
G_k(1) = \frac{H}{f_k}
\]

(2.8)

where \( H^{-1} = E[C] \).

The following lemma gives the expected number of customers in the system for each failure level \( k \), \( k = 0, \ldots, K \).

**Lemma 2.1.** The expected number of customers in failure level \( k \), \( k = 0, \ldots, K \) is

\[
G'_k(1) = \frac{H \lambda}{P f_k} \sum_{l=0}^{K} \left( \prod_{j=k+1}^{K+k-l} (\mu_j + f_j) \right) \left( \prod_{j=K+k+2-l}^{K+k+1} f_j \right),
\]

(2.9)

where \( P = \prod_{k=0}^{K} (\mu_k + f_k) - \prod_{k=0}^{K} f_k \).

**Proof.** First of all, take the derivative of (2.6) to obtain

\[
\mu_k G'_k(z) + \mu_k(z-1)G''_k(z) = \lambda G_k(z) + (\lambda z - \lambda - f_k)G'_k(z) + f_{k-1}G'_{k-1}(z)
\]

(2.10)

for \( k = 0, \ldots, K \).

Putting \( z = 1 \) into (2.10) yields

\[
(\mu_k + f_k)G'_k(1) - f_{k-1}G'_{k-1}(1) = \lambda G_k(1) \quad k = 0, \ldots, K.
\]

(2.11)

Note that the right hand side of above equations are known by Fact 2.1. Therefore, we have \( K \) unknowns in \( K \) equations, which can be solved to obtain equation (2.9).

The following theorem gives the expected number of customers in the system:
**Theorem 2.2.** The expected number of customers in the system is

\[
E[X] = H \frac{\lambda}{P} \sum_{k=0}^{K} \frac{1}{f_k} \sum_{l=0}^{K} \left( \prod_{j=k+1}^{K+k-l} (\mu_j + f_j) \right) \left( \prod_{j=K+k+2-l}^{K+k+1} f_j \right).
\]  

(2.12)

**Proof.** The statement directly follows from the identity \( E[X] = G'(1) = \sum_{k=0}^{K} G'_k(1) \) and Lemma 2.1.

Note that since \( \lambda, \mu_k \) and \( f_k, k = 0, \ldots, K \) are assumed to take positive real values, \( E[X] \) is also a positive real number.

The following lemma gives the second moments of the partial generating functions \( G_k(z), k = 0, \ldots, K \) evaluated at \( z = 1 \):

**Lemma 2.2.** The second moments of the partial generating functions \( G_k(z), k = 0, \ldots, K \) evaluated at \( z = 1 \) are

\[
G''_k(1) = \frac{2\lambda}{Q} \sum_{l=0}^{K} \left( \prod_{j=k+1}^{K+k-l} (2\mu_j + f_j) \right) \left( \prod_{j=K+k+2-l}^{K+k+1} f_j \right) G'_{K+k-l+1}(1),
\]

where \( Q = \prod_{k=0}^{K} (2\mu_k + f_k) - \prod_{k=0}^{K} f_k \).

**Proof.** First of all, take the derivative of (2.10) to obtain

\[
2\mu_k G''_k(z) + \mu_k (z-1) G'''_k(z) = 2\lambda G'_k(z) + (\lambda z - \lambda - f_k) G''_k(z) + f_{k-1} G''_{k-1}(z) k = 0, \ldots, K.
\]

(2.14)

Putting \( z = 1 \) into (2.14) yields

\[
(2\mu_k + f_k) G''_k(1) - f_{k-1} G''_{k-1}(1) = 2\lambda G'_k(1) k = 0, \ldots, K.
\]

(2.15)

Again, the right hand side of above equations are known by Lemma 2.1 and the solution gives the equation (2.13).

The following theorem gives the second moment of the generating function \( G(z) \), evaluated at \( z = 1 \):
Theorem 2.3. The second moment of the generating function $G(z)$, evaluated at $z = 1$ is

$$G''(1) = \frac{2\lambda}{Q} \sum_{k=0}^{K} \sum_{l=0}^{K} \left( \prod_{j = k+1}^{K+k-l} (2\mu_j + f_j) \right) \left( \prod_{j = K+k-l+1}^{K+k} f_j \right) G'_{K+k-l+1}(1).$$  \tag{2.16}$$

Proof. The statement directly follows from the identity $G''(1) = \sum_{k=0}^{K} G''_k(1)$ and Lemma 2.2.

The variance of number of customers in the queue can be obtained from

$$Var(X) = G''(1) + G'(1) - (G'(1))^2.$$  \tag{2.17}$$

Note that eq. 2.11 and eq. 2.15 match with the matrix equation given in (O’Cinneide and Purdue (1986)), since this is a special case of the system considered in (O’Cinneide and Purdue (1986); Falin (2008)), in which the detailed proofs of these facts can be found.

3. Optimal Repair Rate

In this section, our aim is to find a way of determining the optimal repair rate such that total cost incurred by delay and repair is minimized under the assumption that arrival, service and failure rates are exogenous. Note that since service slows down when failures occur, the number of customers in the system increases and consequently, the delay cost increases. By increasing the repair rate, we can limit the number of customers in the system. However, since repair cost increases with repair rate, there is a trade-off between delay and repair costs. Therefore, the optimal repair rate which minimizes the sum of these two cost components should be determined such that the repair rate is between a lower bound $L$ and an upper bound $U$.

Suppose that the holding cost of a customer is $c_1$ units per unit time. Also, let $c_2$ be the one time repair cost. Let $E[X(f_K)]$ and $E[C(f_K)]$ denote, respectively, the expected number of customers in the system and the expected cycle time when repair rate is $f_K$. 

11
Finally, let \( c(f_K) \) be the total cost per unit time when repair rate is \( f_K \) and given by

\[
c(f_K) = c_1 E[X(f_K)] + \frac{c_2}{E[C(f_K)]}
\]

\[= c_1 E[X(f_K)] + c_2 H(f_K). \quad (3.1)
\]

Through a series of lemmas and theorems, we prove that \( c(f_K) \) is a unimodal function that may be convex, concave or neither, although \( E[X(f_K)] \) is a convex function of \( f_K \).

The first lemma is as follows:

**Lemma 3.1.** \( E[X(f_K)] \) can be written as the ratio of two second order polynomials of \( f_K \) with appropriate positive coefficients.

**Proof.** For ease of notation, let \( H(f_K) = \frac{f_K}{1 + (\sum_{k=0}^{K-1} \frac{1}{f_k})} \) and \( P(f_K) = p\mu_K + (p-h)f_K \), where \( p = \prod_{k=0}^{K-1} (\mu_k + f_k) \) and \( h = \prod_{k=0}^{K-1} f_k \). Using these definitions, first moments found in section 2.2 can be rewritten as

\[
G'_k(1) = \frac{\lambda f_K}{(1 + (\sum_{k=0}^{K-1} \frac{1}{f_k}) f_K)(p\mu_K + (p-h)f_K)f_k} (\alpha_k + \beta_k f_K), \quad (3.2)
\]

where

\[
\alpha_K = \prod_{j=0}^{K-1} (\mu_j + f_j) = p,
\]

\[
\alpha_k = \prod_{j=k+1}^{K-1} (\mu_j + f_j) \left( \mu_K \prod_{l=0}^{k-1} \prod_{j=0}^{k-1-l} (\mu_j + f_j) \prod_{j=k+1-l}^{k} f_j + \prod_{j=0}^{k} f_j \right), \quad (3.3)
\]

for \( k = 0, \ldots, K - 1 \) and

\[
\beta_K = \beta_{K-1} = \sum_{l=0}^{K-1} \prod_{j=0}^{K-1-l} (\mu_j + f_j) \prod_{j=K-l}^{K-1} f_j,
\]

\[
\beta_k = \sum_{l=0}^{K-k} \prod_{j=k+1}^{K-k-l+1} (\mu_j + f_j) \prod_{j=K-k-l+2}^{K-k-l} f_j + \sum_{l=k+1}^{K-1} \prod_{j=k+1}^{K-k-l+1} (\mu_j + f_j) \prod_{j=K-k-l+2}^{K-k-l+1} f_j, \quad (3.4)
\]

for \( k = 0, \ldots, K - 2 \).
Then, the expected stationary number in the system is

\[ E[\mathbb{X}(f_K)] = \lambda f_K \frac{\sum_{k=0}^{K-1} \frac{1}{f_k}}{1 + \left( \sum_{k=0}^{K-1} \frac{1}{f_k} \right) f_K} \left( \rho f_K + (p-h)f_K \right) \frac{1}{f_k} \]

\[ = \lambda \frac{\left( \sum_{k=0}^{K-1} \frac{\beta_k}{f_k} \right) f_K^2 + \left( \beta_K + \sum_{k=0}^{K-1} \frac{\alpha_k}{f_k} \right) f_K + \alpha_K}{(p-h) \left( \sum_{k=0}^{K-1} \frac{1}{f_k} \right) f_K^2 + \left( \rho f_K (\sum_{k=0}^{K-1} \frac{1}{f_k}) + (p-h) \right) f_K + \rho \mu K} \]  

\[ = \lambda \frac{a f_K^2 + b f_K + c}{u f_K + v f_K + w}, \]

where \( a, b, c, u, v, w \) are the corresponding coefficients. Specifically,

\[ a = \left( \sum_{k=0}^{K-1} \frac{\beta_k}{f_k} \right), \quad b = \left( \beta_K + \sum_{k=0}^{K-1} \frac{\alpha_k}{f_k} \right), \quad c = \alpha_K, \]

\[ u = (p-h) \left( \sum_{k=0}^{K-1} \frac{1}{f_k} \right), \quad v = \left( \rho \mu K (\sum_{k=0}^{K-1} \frac{1}{f_k}) + (p-h) \right), \quad w = \rho \mu K. \]

Observe that \( E[\mathbb{X}(0)] = \lambda \frac{c}{w} = \lambda \frac{\alpha K}{\rho \mu K} = \frac{\lambda}{\mu K} \), which is an intuitive result since if the repair rate is 0, i.e., repair is not allowed, the state of the system will eventually become \( K \) and the system will behave as an ordinary M/M/\( \infty \) queue with arrival rate \( \lambda \) and service rate \( \mu_K \).

On the other extreme, if we let the repair rate to take arbitrarily large values, then

\[ \lim_{f_K \to \infty} E[\mathbb{X}(f_K)] = \lim_{f_K \to \infty} \lambda \frac{a f_K^2 + b f_K + c}{u f_K + v f_K + w} = \lambda \frac{a}{u} = \lambda \frac{\left( \sum_{k=0}^{K-1} \frac{\beta_k}{f_k} \right)}{(p-h) \left( \sum_{k=0}^{K-1} \frac{1}{f_k} \right)} > 0. \]  

Hence, the expected number in the system is always a positive real number even if the repair rate is extremely large.

In the following lemma, we show that the expected number in the system is a decreasing function of the repair rate. Due to its length, a proof of this lemma is deferred to the appendix.

**Lemma 3.2.** \( E[\mathbb{X}(f_K)] \) is a decreasing function of \( f_K \) in the interval \([0, \infty)\).

The following theorem is an important result of this section:
Theorem 3.1. \( E[X(f_K)] \) is a convex function of \( f_K \) in the interval \([0, \infty)\).

Proof. We know that \( E[X(f_K)] \) is decreasing by Lemma 3.2. Since \( av - bu < 0 \), \( aw - cu < 0 \) and \( bw - cv < 0 \), we conclude that the sum of the roots of \( \frac{dE[X(f_K)]}{df_K} \) is negative whereas the product of the roots is positive. Hence, \( \frac{dE[X(f_K)]}{df_K} \) has no roots in the interval \([0, \infty)\), which implies that \( \frac{d^2E[X(f_K)]}{df_K^2} \) does not change its sign in that interval. Therefore, we conclude that \( E[X(f_K)] \) is either convex or concave. However, since \( E[X(f_K)] \) is decreasing and bounded below by 0, it cannot be concave. Hence, we prove that \( E[X(f_K)] \) is a convex function of \( f_K \) in the specified interval.

In the following theorem, we show that \( H(f_K) \) is an increasing, concave function.

Theorem 3.2. \( \frac{1}{E[C(f_K)]} = H(f_K) \) is an increasing, concave function of \( f_K \) in the interval \([0, \infty)\).

Proof. Recall that \( H(f_K) \) can be written as follows:

\[
H(f_K) = \frac{1}{\sum_{k=0}^{K} \frac{1}{f_k}} = \frac{1}{\sum_{k=0}^{K-1} \frac{1}{f_k} + \frac{1}{f_K}} = \frac{f_K}{1 + \left(\sum_{k=0}^{K-1} \frac{1}{f_k}\right) f_K}.
\] (3.7)

Taking the first derivative of \( H(f_K) \) in equation (3.7) gives

\[
\frac{dH(f_K)}{df_K} = \frac{1}{\left[1 + \left(\sum_{k=0}^{K-1} \frac{1}{f_k}\right) f_K\right]^2} > 0.
\] (3.8)

Hence, we conclude that \( H(f_K) \) is decreasing. Now, let us find the second derivative of \( H(f_K) \) using equation (3.8):

\[
\frac{d^2H(f_K)}{df_K^2} = -\frac{2 \sum_{k=0}^{K-1} \frac{1}{f_k}}{\left[1 + \left(\sum_{k=0}^{K-1} \frac{1}{f_k}\right) f_K\right]^3} < 0.
\] (3.9)

Therefore, we prove that \( H(f_K) \) is concave in the specified interval.

Now, observe that since \( E[X(f_K)] \) is convex by Theorem 3.1 and \( H(f_K) \) is concave by Theorem 3.2, the total cost function \( c(f_K) \) is, in general, nonconvex. At this point, one can think that this nonconvex problem cannot be solved analytically and turn to numerical methods. Figure 1 gives typical cost functions related to an instance of this...
Figure 1: Comparison of cost components as functions of the repair rate.

Figure 1: Comparison of cost components as functions of the repair rate.

Observe that the shape of the total cost function changes depending on parameters $c_1$ and $c_2$. For $c_1 = 1$ and $c_3 = 6$, $c(f_3)$ becomes a convex, decreasing function and $f_3^* = U = 2.25$. For $c_1 = 1$ and $c_2 = 12$, $c(f_3)$ becomes a quasiconvex function and a simple bisection algorithm finds the optimal repair rate $f_3^* = 0.621$. For $c_1 = 1$ and $c_2 = 24$, $c(f_3)$ becomes a concave, increasing function and $f_3^* = L = 0.25$. It turns out that this behavior is not by coincidence, that is, given the parameters of the problem, we can always characterize the total cost function by the values of $c_1$ and $c_2$. In the remaining part of this section, we show that $c(f_K)$ is, in fact, unimodal on $[0, \infty)$ and give conditions under which $c(f_K)$ behaves as seen above.

First, observe that using equations (3.5) and (3.7), the total cost function can be
written as
\[
c(f_K) = c_1 \lambda - \frac{af_K^2 + bf_K + c}{1 + \left(\sum_{k=0}^{K-1} \frac{1}{k} f_K\right) (p\mu_K + (p-h)f_K)} + c_2 \frac{f_K}{1 + \left(\sum_{k=0}^{K-1} \frac{1}{k}\right) f_K},
\] (3.10)
and calculating the derivative yields
\[
c'(f_K) = \frac{Af_K^2 + 2Bf_K + C}{(u f_K^2 + v f_K + w)^2},
\] (3.11)
where
\[
A = (av - bu)\lambda c_1 + [(p-h)v - uw]c_2,
\]
\[
B = (aw - cu)\lambda c_1 + w(p-h)c_2,
\] (3.12)
\[
C = (bw - cv)\lambda c_1 + w^2 c_2.
\]
We see that \(A, B\) and \(C\) are linear functions of \(c_1\) and \(c_2\). Therefore, they can be treated as lines in 2-dimensional \((c_1, c_2)\) space.

**Lemma 3.3.** Slopes of \(A, B\) and \(C\) are positive and in increasing order, i.e.,
\[
0 < -\frac{(av - bu)\lambda}{(p-h)v - uw} < -\frac{(aw - cu)\lambda}{w(p-h)} < -\frac{(bw - cv)\lambda}{w^2}.
\] (3.13)

**Proof.** By definition of \(u, v\) and \(w\) (refer to equation (3.5)), one can easily show that \((p-h)v - uw = (p-h)^2\). Hence, the coefficients of \(c_2\) are positive for \(A, B\) and \(C\) whereas the coefficients of \(c_1\) are negative by Lemma 3.2. Therefore, all the slopes are positive.

In order to show the correct ordering, let us first consider \(T = (p-h)(bw-cv)-w(aw-cu)\). If this quantity is negative, then the second inequality follows.
\[ T = (p - h)p \left[ \left( \frac{\beta_K}{f_k} + \sum_{k=0}^{K-1} \frac{\alpha_k}{f_k} \right) \mu_K - \left( (p - h) + p\mu_K \sum_{k=0}^{K-1} \frac{1}{f_k} \right) \right] \]

\[-p^2 \mu_K \left( \mu_K \sum_{k=0}^{K-1} \frac{\beta_k}{f_k} - (p - h) \sum_{k=0}^{K-1} \frac{1}{f_k} \right)\]

\[= p \left[ \left( p - h \right) \left( \mu_K \beta_K - (p - h) \right) \right] - \mu_K \sum_{k=0}^{K-1} \frac{p\mu_K \beta_k - (p - h) \alpha_k}{f_k} \tag{4.3} \]

\[= p \sum_{l=0}^{K-1} (\mu_K - \mu_{K-1-l}) \prod_{j=0}^{K-2-l} (\mu_j + f_j) \prod_{j=K-l}^{K-1} f_j \sum_{k=0}^{K-1} (\mu_k - \mu_K) \prod_{j=k+1}^{K-1} (\mu_j + f_j) \prod_{j=0}^{k-1} f_j \]

< 0, \tag{3.14}

since \( \mu_K < \mu_{K-1-l} \) for \( l = 0, \ldots, K - 1 \) and \( \mu_k > \mu_K \) for \( k = 0, \ldots, K - 1 \).

Now, consider the first inequality in (3.13). Since \( T < 0 \), we have

\[ [(p - h)v - uw](aw - cu) - (p - h)w(aw - bu) = u[p - h](bw - cv) - w(aw - cu)] = uT < 0. \]

Hence, the result follows. \( \square \)

The following theorem shows why this property of slopes is crucial for unimodality of \( c(f_K) \).

**Theorem 3.3.** \( c(f_K) \) is unimodal on \( [0, \infty) \).

**Proof.** Figure 2 shows the signs of \( A, B \) and \( C \) with respect to cost coefficients \( c_1 \) and \( c_2 \).

Here, we have 3 cases:

1. \( A < 0 \) and \( C < 0 \): In this case, we also have \( B < 0 \). So, \( c(f_K) \) is decreasing and \( c'(f_K) \) has no roots in the interval \( [0, \infty) \). Hence, \( c(f_K) \) is convex and \( f_K^* = U \).

2. \( A > 0 \) and \( C < 0 \): In this case, independent of the sign of \( B \), \( c(f_K) \) is first decreasing and then, increasing since \( c'(f_K) \) has a positive root at \( f^* = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \). So, \( c(f_K) \) is unimodal and \( f_K^* = \max\{L, \min\{f^*, U\}\} \).

17
Figure 2: Signs of $A$, $B$ and $C$ with respect to $c_1$ and $c_2$.

3. $A > 0$ and $C > 0$: In this case, we also have $B > 0$. So, $c(f_K)$ is increasing and $c'(f_K)$ has no roots in the interval $[0, \infty)$. Hence, $c(f_K)$ is concave and $f_K^* = L$.

Let us close this section with some numerical results. Here, we focus on a system with only one failure level and fix the arrival rate as $\lambda = 10$. We experiment with two different values, 5 and 2.5, of the initial service rate, $\mu_0$. Also, for each of these levels, we select 3 different values for $\mu_1$ as $0.8\mu_0$, $0.7\mu_0$ and $0.6\mu_0$, which corresponds to 20%, 30% and 40% decrease from the initial service rate. Finally, for different values of failure rate, $f_0 \in [0.1, 1.6]$, we calculate the “critical ratio” of $c_2/c_1$, i.e., the ratio for which the cases introduced previously occur. Results can be seen in Figure 3.

First of all, we observe that when $\mu_0$ is halved, critical ratios are doubled whenever $\frac{\mu_0}{\mu_1}$ ratios remain the same. In fact, the graphs on the right column are exactly the same as the ones on the left except the values on the vertical axes that are doubled. Secondly, for fixed $\mu_0$, as the decrease in service rate increases, critical ratios increase as well. Thirdly, there is an inverse relation between the failure rate and critical ratios given that the other parameters are kept the same. Finally, although Case 3 region is always the largest, as $\frac{\mu_0}{\mu_1}$ ratio decreases Case 1 and Case 2 regions get larger.
4. Conclusions and Future Work

We consider the maintenance of an M/M/∞ system with $K$ levels of service deterioration. The objective function is the expected total cost of delay and repair per unit time as a function of the repair rate, that is applied at the last service state. After deriving the first two moments of the stationary number of customers in a system, we
show that the expected total cost is unimodal. In fact, we provide conditions on the cost coefficients that will cause the objective function to have simple forms, thus, providing the optimal repair rate immediately as either at the lower bound, the upper bound or in intermediate value obtained by the first order optimality conditions.

In our study, repair is only allowed in the highest failure level. However, allowing the repair at each failure state is a promising future study. In this case, repair process may lead the current state to change to either state 0 or to the previous failure state. Also, a finite server system which is exposed to interruptions, and the implementation of this work into different application areas other than service systems and transportation are two other possible study subjects.

Bibliography


22


Appendix: Proof of Lemma 3.2

Taking the first derivative of $E[X(f_K)]$ in equation (3.5) yields

$$
\frac{dE[X(f_K)]}{df_K} = \lambda \frac{(av - bu)f_K + 2(aw - cu)f_K + (bw - cv)}{(uf_K + vf_K + w)^2}.
$$

(4.1)
In order to prove that \( E[X(f_K)] \) is decreasing in the specified interval, it suffices to show that \( av - bu, aw - cu \) and \( bw - cv \) are negative.

First, consider \( av - bu \).

\[
av - bu = \left( \sum_{k=0}^{K-1} \frac{\beta_k}{f_k} \right) \left( p \mu K \sum_{k=0}^{K-1} \frac{1}{f_k} + (p - h) \right) - \left( \beta_K + \sum_{k=0}^{K-1} \frac{\alpha_k}{f_k} \right) \left( p - h \sum_{k=0}^{K-1} \frac{1}{f_k} \right)
\]

\[
= \left( \sum_{k=0}^{K-1} \frac{1}{f_k} \right) \left( p \mu K \sum_{k=0}^{K-1} \frac{\beta_k}{f_k} - (p - h) \sum_{k=0}^{K-1} \frac{\alpha_k}{f_k} \right) + (p - h) \left( \sum_{k=0}^{K-1} \frac{\beta_k}{f_k} - \beta_K \sum_{k=0}^{K-1} \frac{1}{f_k} \right).
\]

(4.2)

Now, consider these two summands separately. For each \( k = 0, \ldots, K - 1 \), the first part becomes

\[
\mu_K \beta_k - (p - h) \alpha_k = \mu_K \beta_k \prod_{j=k+1}^{K-1} (\mu_k + f_k) - \alpha_k \left( \prod_{j=k+1}^{K-1} (\mu_k + f_k) - \prod_{j=0}^{K-1} f_k \right)
\]

\[
= \prod_{j=k+1}^{K-1} \left( \mu_j + f_j \right) \prod_{j=0}^{K-1} \left( \mu_K - \mu_{K-l-1} \right) \prod_{j=0}^{K-1} \left( \mu_j + f_j \right) \prod_{j=k+1}^{K-1} f_j < 0,
\]

since \( \mu_K < \mu_{K-l-1} \) for \( l = 0, \ldots, K - 1 \). Hence, the first part is negative.

In order to analyze the second part, first recall that \( \beta_{K-1} = \beta_K \). Therefore, if \( K = 1 \), then the second part is 0. Otherwise, for each \( k = 0, \ldots, K - 2 \), the second part becomes

\[
\beta_k - \beta_K = \sum_{l=0}^{k} (\mu_{K-1} - \mu_{k-l}) \prod_{j=k+1}^{K-1} \left( \mu_j + f_j \right) \prod_{j=k+1}^{K-1} f_j + (\mu_{K+k-l-1} - \mu_0) \sum_{l=k+1}^{K-1} \prod_{j=k+1}^{K-1} \left( \mu_j + f_j \right) \prod_{j=k+1}^{K-1} f_j < 0,
\]

(4.4)

since \( \mu_{K-1} < \mu_{k-l} \) for \( l = 0, \ldots, k \) and \( \mu_{K+k-l-1} < \mu_0 \) for \( l = k+1, \ldots, K - 2 \). Hence, the second part is negative if \( K > 1 \). In either case, the second part is nonpositive. Since the first part is negative, we conclude that \( av - bu < 0 \).
Now, consider $aw - cu$

$$aw - cu = p \mu_K \left( \sum_{k=0}^{K-1} \beta_k \right) - p(p-h) \sum_{k=0}^{K-1} \frac{1}{f_k}$$

$$= p \sum_{k=0}^{K-1} \frac{\mu_K \beta_k - p + h}{f_k}. \quad (4.5)$$

For each $k = 0, \ldots, K-1$, we have

$$\mu_K \beta_k - \prod_{k=0}^{K-1} (\mu_k + f_k) + \prod_{k=0}^{K-1} f_k = \sum_{l=0}^{k} (\mu_K - \mu_{k-l}) \prod_{j=k+1}^{K-l} (\mu_j + f_j) \prod_{j=k+1}^{K-l} f_j$$

$$+ (\mu_K - \mu_{K+k-l}) \sum_{l=k+1}^{K} \prod_{j=k+1}^{K-l} (\mu_j + f_j) \prod_{j=k+1}^{K-l} f_j$$

$$< 0, \quad (4.6)$$

since $\mu_K < \mu_{k-l}$ for $l = 0, \ldots, k$ and $\mu_K < \mu_{K+k-l}$ for $l = k+1, \ldots, K-1$. Hence, we conclude that $aw - cu < 0$.

Finally, consider $bw - cv$

$$bw - cv = \left( \beta_K + \sum_{k=0}^{K-1} \frac{\alpha_k}{f_k} \right) p \mu_K - p \left( p \mu_K \sum_{k=0}^{K-1} \frac{1}{f_k} + (p-h) \right)$$

$$= p \left( \beta_K \mu_K + \mu_K \sum_{k=0}^{K-1} \frac{\alpha_k}{f_k} - p - p \mu_K \sum_{k=0}^{K-1} \frac{1}{f_k} \right) \quad (4.7)$$

$$= p(\beta_K \mu_K - p + h) + p \mu_K \sum_{k=0}^{K-1} \frac{\alpha_k - p}{f_k}.$$ 

Now, consider these two summands separately. The first part becomes

$$\beta_K \mu_K - p + h = \mu_K \left( \sum_{l=0}^{K-1-K-2-l} \prod_{j=0}^{l} (\mu_j + f_j) \prod_{j=K-l}^{K-1} f_j \right) - \prod_{k=0}^{K-1} (\mu_k + f_k) + \prod_{k=0}^{K-1} f_k$$

$$= \sum_{l=0}^{K-1} (\mu_K - \mu_{K-1-l}) \left( \prod_{j=0}^{l} (\mu_j + f_j) \prod_{j=K-l}^{K-1} f_j \right) \quad (4.8)$$

$$< 0,$$
since $\mu_K < \mu_{K-1-l}$ for $l = 0, \ldots, K-1$. Hence, the first part is negative.

For each $k = 0, \ldots, K-1$, the second part becomes

$$\alpha_k - p = \prod_{j=k+1}^{K-1} (\mu_j + f_j) \left( \mu_k \sum_{l=0}^{k-1-l} (\mu_j + f_j) \prod_{j=0}^{k-1-l} f_j + \prod_{j=0}^{k} f_j \right) - \prod_{k=0}^{K-1} (\mu_k + f_k)$$

$$= \prod_{j=k+1}^{K-1} (\mu_j + f_j) \sum_{l=0}^{k} (\mu_K - \mu_{k-l}) \prod_{j=0}^{k-1-l} (\mu_j + f_j) \prod_{j=k+1-l}^{k} f_j$$

$$< 0,$$

(4.9)

since $\mu_K < \mu_{k-l}$ for $l = 0, \ldots, k$. Hence, the second part is also negative. Because both parts are negative, we conclude that $bw - cv < 0$.

Since $av - bu$, $aw - cu$ and $bw - cv$ are all negative, we prove that $\frac{dE[X(f_K)]}{df_K} < 0$ for $f_K \in [0, \infty)$, which consequently implies that $E[X(f_K)]$ is a decreasing function in that interval. \qed