IISE Transactions

ISSN: 2472-5854 (Print) 2472-5862 (Online) Journal homepage: http://www.tandfonline.com/loi/uiie21

Condition-based joint maintenance optimization for a large-scale system with homogeneous units

Young Myoung Ko & Eunshin Byon

To cite this article: Young Myoung Ko & Eunshin Byon (2017) Condition-based joint maintenance optimization for a large-scale system with homogeneous units, IISE Transactions, 49:5, 493-504, DOI: 10.1080/0740817X.2016.1241457

To link to this article: http://dx.doi.org/10.1080/0740817X.2016.1241457

Accepted author version posted online: 04 Oct 2016.
Published online: 04 Oct 2016.

IISE TRANSACTIONS
2017, VOL. 49, NO. 5, 493–504
http://dx.doi.org/10.1080/0740817X.2016.1241457

Taylor & Francis
Taylor & Francis Group

Condition-based joint maintenance optimization for a large-scale system with homogeneous units

Young Myoung Ko[a] and Eunshin Byon[b]

[a]Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Korea; [b]Department of Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI, USA

ABSTRACT
A joint maintenance policy that simultaneously repairs multiple units is useful for large-scale systems where the setup cost to initiate the maintenance is generally higher than the repair costs. This study proposes a new method for scheduling maintenance activities in a large-scale system with homogeneous units that degrade over time. Specifically, we consider the maintenance type that renews all units at each maintenance activity, which is practically applicable for systems where the units need to be regularly maintained. To make the analysis computationally tractable, we discretize the health condition of each unit into a finite number of states. The proposed optimization formulation triggers the maintenance activity based on the fraction of units at each degradation state. Based on relevant asymptotic theories, we analytically obtain the optimal threshold in the fraction of units at each state that minimizes the long-run average maintenance cost. Our implementation results with a wide range of parameter settings show that the proposed maintenance strategy is more cost-effective than alternative strategies.

ARTICLE HISTORY
Received 17 December 2015
Accepted 6 September 2016

KEYWORDS
Asymptotic theories; multi-unit systems; reliability; renewal process

1. Introduction

Modeling and analyzing multi-unit systems can be challenging, due to inter-dependency among units. In general, there are three different inter-dependency types. Statistical (or stochastic) dependence refers to the health condition of units affecting the condition (or lifetime) of other units or when random degradation and shock stresses cause damage in several components (Bian and Gebraeel, 2014). Structural dependence implies that failure of a unit may cause multiple units to be repaired or replaced (Nicolai and Dekker, 2008). Lastly, economic dependence refers to the case where repair costs include both the variable costs that depend on the number of units to be repaired and their degradation severity, and the fixed costs—i.e., setup cost—to initiate a repair.

Our study concerns the economic dependence for a large-scale system where the units independently degrade over time. When each maintenance activity requires dispatching skilled crew with specialized equipment, a high setup cost is incurred and thus repairing multiple units together, when possible, would be more cost-effective than repairing a single unit individually.

This study aims to provide a cost-effective maintenance strategy that minimizes the long-run average cost of multi-unit repairs. In particular, we consider a system with homogeneous units that follow the same stochastic degradation process (Wang and Wei, 2011; Ko and Byon, 2015) and analyze a maintenance scheduling problem where all of the units in the system are repaired and renewed simultaneously.

This type of joint maintenance is practical in several applications where the fixed cost is high, the units in the system require regular maintenance, and the repair time is negligible. For example, gearbox oil needs to be changed regularly in wind power systems, and operators prefer changing oil in all turbines in a single visit due to logistical difficulties and the high cost of site visits. The common practice in the wind industry is to regularly maintain the gearbox oil quality based on a manufacturer-recommended fixed time interval. However, thanks to the development of condition monitoring systems, more proactive condition-based oil changes are becoming recognized in the wind industry (Dvorak, 2014).

Joint maintenance of multiple units is also effective in applications such as rechargeable underwater batteries used in hydrologic and water quality sensors (Horsburgh *et al.*, 2015) and manufacturing systems where all production lines have to be stopped during the repair; e.g., due to a required electricity shutdown (Wang and Wei, 2011). Another example is an electricity power distribution system that needs vegetation management such as tree trimming to avoid tree-caused power outages and maintain power distribution reliability (Guikema *et al.*, 2006).

Suppose that a system has 100 units, and each unit's degradation condition can be categorized into normal, alert, alarm, and failure states (Byon *et al.*, 2010). The operator must determine whether to perform maintenance; e.g., when 50% of units are in the alert state, 10% of units are in the alarm state, or in some combination of both states. Clearly, there are many combinations of possible thresholds. In this study, we propose a new maintenance strategy to trigger repair activities when the number of units at each degradation state reaches a certain threshold and find the optimal thresholds to minimize the overall maintenance cost, based on the renewal theories combined with

CONTACT Eunshin Byon ebyon@umich.edu
Color versions of one or more of the figures in this article can be found online at www.tandfonline.com/uiie.

asymptotic theories. Our results suggest that the optimal thresholds vary, depending on the cost structure and units' degradation processes.

We want to emphasize the major differences of our proposed approach from existing studies in the literature. Unlike previously published renewal-based studies, which consider a fixed renewal cycle (Wang and Pham, 2011; Jiang *et al*., 2012; Liu *et al*., 2013), our analysis starts with a random renewal cycle caused by multiple units' stochastic degradation processes. We show that the random maintenance cycle, determined from the stochastic evolution of a system's health condition, asymptotically converges to a certain fixed value. We also link the maintenance strategy that triggers maintenance based on the optimal thresholds with the scheduled maintenance that performs maintenance with a fixed cycle. Moreover, many analytical approaches available in the literature have been extensively applied to a single-unit system (Wang and Pham, 2011; Jiang *et al*., 2012; Byon, 2013; Liu *et al*., 2013), while some studies considering the maintenance optimization of multi-unit systems have been limited to a small number (e.g., two or three) of units (Tian and Liao, 2011).

To the best of our knowledge, this article is the first that analytically derives the optimal maintenance policy for a large-scale system to minimize the maintenance cost. We believe that the proposed formulation and solution approach represents a breakthrough in analyzing a large-scale system maintenance problem and overcomes the limitations of existing studies that have been applied to either single-unit systems or small-scale systems, due to their limited scalability.

The remainder of this study is organized as follows. Section 2 reviews the relevant literature. Section 3 formulates the problem and derives the optimal solution. Section 4 implements the proposed approach. Section 5 concludes and suggests future research directions.

2. Literature review

Due to the complexity of analyzing systems with multiple units, most maintenance optimization studies consider a single-unit system. Several stochastic approaches study a unit's probabilistic degradation process and estimate a lifetime distribution, assuming that the data are generated from a stochastic process such as gamma, Brownian motion, or other processes (Kharoufeh *et al*., 2010; Chen and Tsui, 2013; Ye *et al*., 2013). Renewal processes are employed either to minimize the maintenance cost or to maximize the system availability (Jiang *et al*., 2012; Liu *et al*., 2013). Jiang *et al*. (2012) study the reliability of a unit subject to multiple competing failure precesses and provide the maintenance optimization model to minimize the long-run average maintenance cost. Liu *et al*. (2013) propose a condition-based maintenance model that considers both progressive degradation and multiple sudden failures in a unit.

Some recent studies attempt to model the reliability of systems with multiple units. Song *et al*. (2014) analyze a system of multiple dependent units that are simultaneously exposed to external loads causing degradation or sudden failures. Bian and Gebraeel (2014) propose a stochastic model to characterize the degradation processes of multiple interacting components, where the components' degradation rates are inter-dependent. Mixed-effects models are employed for degradation analysis for a population of independent units considering unit-to-unit variations (Hong *et al*., 2015). However, the focus of these studies is on estimating the remaining lifetime (or reliability) of the system, rather than providing cost-effective maintenance strategies.

Tian and Liao (2011) investigate maintenance optimization considering economic dependence, where the system consists of multiple identical and independent units. They consider predefined rules such as "when one component's failure risk is over a certain threshold, perform preventive maintenance for components whose failure risk is over another threshold" and numerically determine the two threshold levels. A similar rule-based approach is used by Castanier *et al*. (2005) and by Shafiee and Finkelstein (2015) for a two-unit system and a system with non-identical units, respectively. Perez *et al*. (2015) use a discrete-event simulation to study the effects of different rules on the operations and maintenance cost for multi-component wind turbines.

Generally, the rule-based approaches for the repair of multiple units that are presented in the literature are based on each individual unit's degradation condition. They can be useful for finding the optimal policy in a small-scale system. The existing approaches, however, can undermine the opportunity to further reduce maintenance costs in large-scale systems. For example, in the studies by Castanier *et al*. (2005), Tian and Liao (2011), and Perez *et al*. (2015), maintenance for the other units is triggered when one unit's condition reaches a certain threshold. When the number of units is large and the fixed cost is high, it could be more cost-effective to wait until a certain number of multiple units (not one unit) reach a certain degraded condition. The challenge is that enumerating all combinations of possible rules and finding an optimal policy among them are computationally intractable for large-scale problems.

3. Methodology

This section formulates the problem and discusses the proposed solution procedure.

3.1. Problem description and formulation

We consider a system with N units, where a system implies a group of units. It is assumed that the degradation processes of the units are homogeneous, implying that the units randomly degrade over time following the same probability law. The degradation process of each unit is assumed to be independent of other units.

In general, a unit's degradation process is a continuous stochastic process. Handling a continuous degradation state, however, imposes some difficulties in analyzing the dynamics of a large-scale system. First, when a massive number of units are in different health conditions (because units degrade stochastically), it is not straightforward to interpret individual units' conditions to get the system-level health information. Continuous states also require considering an infinite number of possible rules for triggering maintenance. Therefore, we categorize the degradation levels into a finite number of states, $1, 2, \ldots, M$, where State 1 denotes a unit's best condition and State M denotes the failure state.

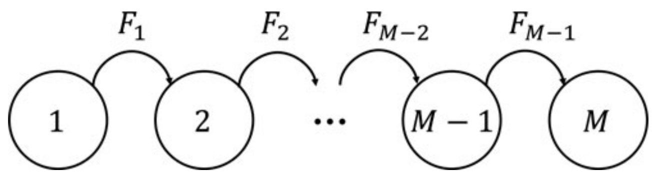


Figure 1. Degradation process of each unit. (Note: each unit's degradation condition is classified into M discrete states, and we assume that the transition time from one state to the next degradation state can follow any general distribution; e.g., Weibull, lognormal, or gamma distribution.)

Each unit degrades from State i to State $i+1$ with a general inter-transition distribution $F_i(\cdot)$ (see Fig. 1). For a wide applicability of the proposed approach, we do not impose any explicit assumptions on $F_i(\cdot)$ but allow any distribution. For example, $F_i(\cdot)$ can be Weibull, lognormal, or gamma distributed. Alternatively, the transition can follow the Markovian degradation process. We assume that the distribution parameters can be estimated from historical data.

Analyzing the multi-unit system's dynamics and minimizing the long-run maintenance cost begins by asking two questions: (i) How to translate individual units' degradation processes to the system-level health evolution over time? and (ii) Based on the extracted system-level heath information, when to perform maintenance activities? The first question is answered in Ko and Byon (2015), which analytically derives the asymptotic multivariate distribution of the system-level health condition. This article answers the second question, building upon the results in Ko and Byon (2015).

Let $\mathbf{S}_0(t) = (X^1(t), \ldots, X^M(t))'$ denote the system's heath status at time t, where $X^i(t)$ is the number of units at State i at time t with $\sum_{i=1}^{M} X^i(t) = N$. We assume $X^1(0) = N$; i.e., all units are new in the beginning. As the system operates, the units in the system degrade over time; i.e., transition from one degradation state to another, and some units may reach State M, the failure state. Each unit at State M incurs a revenue loss of L per unit time.

The system operator continually monitors the system to decide when to visit the system site and perform maintenance operations. Once maintenance operations are completed, all of the units in the system become "as good as new"; i.e., $X^1(\tau) = N$, where τ is the time when the maintenance operations are completed. We assume that the duration for the maintenance operations is negligible, compared with the time-to-maintenance interval. Each maintenance activity incurs the fixed setup cost P per visit and a repair cost C_i for each unit at State i.

With the definition of the system's health status $\mathbf{S}_0(t)$, the objective is to identify the optimal thresholds triggering maintenance operations, so that the operator can minimize the long-run average cost for maintenance. Specifically, the operator needs to know the optimal fraction of units at each degradation state; e.g., what percentage of units should be in the alarm condition to initiate the maintenance. Let τ denote the first passage time to the thresholds. We define τ as a stopping time:

$$\tau = \inf\Big\{t : (X^1(t)/N - \gamma_1)(\gamma_2 - X^2(t)/N)\cdots (\gamma_M - X^M(t)/N) \le 0\Big\}, \tag{1}$$

where τ is the minimum time that any of the fraction of units at State i, $X^i(t)/N$, reaches its corresponding threshold γ_i. Therefore, τ can be regarded as a hitting time in the system's stochastic degradation process. Note that $X^1(t)$, the number of units at State 1, is non-increasing from N as units transition to more deteriorated states over time. The small fraction at State 1 implies that the majority of units have degraded to more deteriorated conditions. Therefore, we "trigger" the maintenance when the non-increasing fraction, $X^1(t)/N$, reaches the threshold γ_1. On the other hand, the number of units in other states are initially non-decreasing from zero. As such, we use $\gamma_i - X^i(t)/N$ for $i = 2, 3, \ldots, M$ to define a valid stopping time, so that a maintenance action is invoked, when $X^i(t)/N$ reaches its threshold γ_i. Under the proposed formulation, when the number of units at each state hits any of γ_i, we perform maintenance.

As all of the units' conditions are renewed after each maintenance operation, we formulate the problem as a renewal process (Kulkarni, 1995) and minimize the long-run average maintenance cost as

$$\underset{\boldsymbol{\gamma}}{\text{minimize}} \quad f_0(\boldsymbol{\gamma}) = \frac{\mathrm{E}\Big[P + \sum_{i=1}^{M} C_i X^i(\tau) + \int_0^\tau L X^M(t)dt\Big]}{\mathrm{E}[\tau]},$$

where

$$\tau = \inf\Big\{t : (X^1(t)/N - \gamma_1)(\gamma_2 - X^2(t)/N)\cdots (\gamma_M - X^M(t)/N) \le 0\Big\}, \tag{2}$$

and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_M)$ is the threshold vector. In the objective function, the numerator implies the expected repair costs during τ periods. The first and second terms in the numerator correspond to the fixed and variable costs, respectively. The last term $\int_0^\tau L X^M(t)dt$ is the total revenue losses due to failed units. The denominator denotes the expected renewal cycle.

With this formulation, our question now becomes, "What is the optimal threshold vector $\boldsymbol{\gamma}$ that minimizes the long-run average cost in Equation (2)?" Solving the optimization problem in Equation (2) is challenging, as the stochastic process $\{\mathbf{S}_0(t), t \ge 0\}$ is multi-dimensional and non-Markovian (note that a Markovian degradation process is a special case). Additionally, the objective function in Equation (2) includes $E[\tau]$ and $E[X^i(\tau)]$, $i = 1, 2, \ldots, M$, and the probability law required for calculating them is unknown. It should also be noted that the stopping time τ is a random variable, due to units stochastically deteriorating and the system's health status $\mathbf{S}_0(t) = (X^1(t), \ldots, X^M(t))'$ being a random vector. As mentioned in Section 1, this aspect differs from other renewal-based maintenance optimization studies that consider a fixed time-to-maintenance cycle. The analysis becomes more complicated as the number of units in the system increases.

To facilitate the solution process for the problem, we approximate the formulation in Equation (2) based on the asymptotic properties of the system's health status $\mathbf{S}_0(t)$. The next section describes the details of our approach.

3.2. Derivation of asymptotic optimization model

To tackle the issues described in the previous section and obtain the optimal policy that can be easily scaled up for large-size systems, we analyze the dynamics of the system using relevant asymptotic theories. We first use the results in Ko and Byon (2015) where the multivariate distribution of $\mathbf{S}_0(t)$ is analytically derived. For the sake of readability, we briefly summarize their results below.

We approximate the inter-transition time distributions $F_i(\cdot)$s, $i = 1, 2, \ldots, M-1$, using phase-type distributions, as it has been shown that phase-type distributions are dense in any distributions with positive support (Asmussen *et al.*, 1996). Therefore, in the subsequent discussions, we regard $F_i(\cdot)$ as the phase-type distribution that approximates the transition distribution from State i to State $i+1$. With the phase-type distribution, the number of units at State i (except the failure state M), $i = 1, \ldots, M-1$, is further categorized into n_i phases. Let $X_j^i(t)$ denote the number of units at Phase j, $j = 1, 2, \ldots, n_i$, of State i, $i = 1, 2, \cdots, M$, at time t. For the failure state M, we use $X^M(t) = X_1^M(t)$ with $n_M = 1$. Then, the system's health status $\mathbf{S}_0(t)$ at time t can be re-expressed as

$$\mathbf{S}(t) = \left(X_1^1(t), \ldots, X_{n_1}^1(t), \ldots, X_1^{M-1}(t), \ldots X_{n_{M-1}}^{M-1}(t), X_1^M(t)\right)', \tag{3}$$

and the number of units at State i is obtained by

$$X^i(t) = \sum_{j=1}^{n_i} X_j^i(t), \text{ for } i = 1, 2, \ldots, M-1, \tag{4}$$

for $t \in [0, T]$, where $T > 0$ is the lifetime of the system.

Ko and Byon (2015) apply the fluid and diffusion limits combined with the uniform acceleration technique to obtain the asymptotic distribution of $\mathbf{S}(t)$. To analyze the system's asymptotic behavior, they parameterize the number of units in the system and let η denote the scale of the system (i.e., number of units in the system). Then, the system's health status $\mathbf{S}(t)$ in Equation (3), can be represented as

$$\mathbf{S}^\eta(t) = \Big(X_1^{1,\eta}(t), \ldots, X_{n_1}^{1,\eta}(t), \ldots, X_1^{M-1,\eta}(t), \ldots X_{n_{M-1}}^{M-1,\eta}(t), X_1^{M,\eta}(t)\Big)', \tag{5}$$

for $t \in [0, T]$ where $\{X_j^{i,\eta}(t)\}_{\eta \geq 1}$, $i = 1, \ldots, M$, $j = 1, \ldots, n_i$, is a new stochastic process by accelerating the number of units. Then, it follows that

$$\lim_{\eta\to\infty} \frac{\mathbf{S}^\eta(t)}{\eta} = \bar{\mathbf{s}}(t) \text{ almost surely (a.s.)}, \tag{6}$$

where $\bar{\mathbf{s}}(t) = (\bar{x}_1^1(t), \ldots, \bar{x}_{n_1}^1(t), \ldots, \bar{x}_1^{M-1}(t), \ldots, \bar{x}_{n_{M-1}}^{M-1}(t), \bar{x}_1^M(t))'$ is the fluid limit that can be obtained by solving the ordinary differential equations (ODEs; see Ko and Byon (2015) for more details). The element of $\bar{\mathbf{s}}(t)$, $\bar{x}_j^i(t)$, represents the asymptotic fraction of units at Phase j, $j = 1, 2, \ldots, n_i$, of State i, $i = 1, 2, \ldots, M$. For a large η, we can approximate $E[\mathbf{S}^\eta(t)]$ as

$$E[\mathbf{S}^\eta(t)] \approx \eta\bar{\mathbf{s}}(t). \tag{7}$$

Now, based on the above result from Ko and Byon (2015), to find the optimal thresholds, we rewrite the optimization problem in Equation (2) as

$$\underset{\boldsymbol{\gamma}}{\text{minimize}}\ f_0^\eta(\boldsymbol{\gamma}) = \frac{\mathrm{E}\Big[P + \sum_{i=1}^M C_i \sum_{j=1}^{n_i} X_j^{i,\eta}(\tau^\eta) + \int_0^{\tau^\eta} LX_1^{M,\eta}(t)dt\Big]}{\mathrm{E}[\tau^\eta]}, \tag{8}$$

where

$$\tau^\eta = \inf\left\{t : \left(\sum_{j=1}^{n_1} \frac{X_j^{1,\eta}(t)}{\eta} - \gamma_1\right)\left(\gamma_2 - \sum_{j=1}^{n_2} \frac{X_j^{2,\eta}(t)}{\eta}\right)\cdots \left(\gamma_{M-1} - \sum_{j=1}^{n_{M-1}} \frac{X_j^{M-1,\eta}(t)}{\eta}\right)\left(\gamma_M - \frac{X_1^{M,\eta}(t)}{\eta}\right) \leq 0\right\}. \tag{9}$$

Using the linearity of expectation and Fubini theorem (Folland, 1999), the long-run average cost, $f_0^\eta(\boldsymbol{\gamma})$ in Equation (8), becomes

$$f_0^\eta(\boldsymbol{\gamma}) = \frac{P + \sum_{i=1}^M C_i \sum_{j=1}^{n_i} \mathrm{E}\Big[X_j^{i,\eta}(\tau^\eta)\Big] + \mathrm{E}\Big[\int_0^{\tau^\eta} LX_1^{M,\eta}(t)dt\Big]}{\mathrm{E}[\tau^\eta]}. \tag{10}$$

In order to obtain the long-run average cost, we should know $\mathrm{E}[\tau^\eta]$, $\mathrm{E}[X_j^{i,\eta}(\tau^\eta)]$, and $\mathrm{E}[\int_0^{\tau^\eta} X^{M,\eta}(t)dt]$. Their exact values, however, cannot be directly obtained, as the distributions of τ^η and $\mathbf{S}^\eta(\tau^\eta)$ are unknown. Therefore, we derive their asymptotic values and prove the convergence of our objective function (see Appendix for the detailed proofs).

First, Lemma 1 shows that, given the threshold values $\gamma_1, \gamma_2, \ldots, \gamma_M$, the random hitting time τ^η in Equation (9) converges to a deterministic value almost surely (a.s.).

Lemma 1. *Define*

$$\bar{\tau} = \inf\left\{t : \left(\sum_{j=1}^{n_1} \bar{x}_j^1(t) - \gamma_1\right)\left(\gamma_2 - \sum_{j=1}^{n_2} \bar{x}_j^2(t)\right)\cdots \left(\gamma_{M-1} - \sum_{j=1}^{n_{M-1}} \bar{x}_j^{M-1}(t)\right)\left(\gamma_M - \bar{x}_1^M(t)\right) \leq 0\right\}. \tag{11}$$

Then, $\tau^\eta \to \bar{\tau}$ *a.s.*

In Lemma 1, $\bar{x}_j^i(t)$, $i = 1, \ldots, M$, $j = 1, \ldots, n_i$, the fluid limit at time t in $\bar{\mathbf{s}}(t)$ in Equation (6), is obtained from the state transition distribution $F_i(\cdot)$, $i = 1, 2, \ldots, M-1$. Therefore, given the transition distributions and the threshold vector, the asymptotic hitting time $\bar{\tau}$ is uniquely determined. The result of Lemma 1 has an important implication. The original hitting time τ in Equation (1) (or τ^η in Equation (9)) is a random variable. However, when a system consists of a large number of independent units, this random quantity asymptotically converges to the deterministic $\bar{\tau}$ in Equation (11). This property makes problem solving more tractable (to be detailed in Section 3.4).

Note that we need to know the expected value of τ^η—i.e., $\mathrm{E}[\tau^\eta]$—to obtain the objective value in Equation (8). The almost sure convergence in Lemma 1 does not guarantee the convergence of the expectation, but Lemma 2 shows that $E[\tau^\eta]$ also converges to $\bar{\tau}$, as the number of units increases.

Lemma 2. $\lim_{\eta\to\infty} E[\tau^\eta] = \bar{\tau}$.

Next, using the fact that the fraction of units at each phase of each state is finite, Lemma 3 states that the expected fraction of units at Phase j of State i converges to its asymptotic fraction $\bar{x}^i_j(\bar{\tau})$, as the number of units increases.

Lemma 3. $\lim_{\eta\to\infty} E[X^{i,\eta}_j(\tau^\eta)/\eta] = \bar{x}^i_j(\bar{\tau})$ *for* $i = 1, \dots, I$, $j = 1, \dots, n_i$.

Finally, in Lemma 4 we obtain the value to which the expected revenue losses $E[\int_0^{\tau^\eta} X^{M,\eta}_1(t)/\eta dt]$ converge, as the number of units increases.

Lemma 4. $\lim_{\eta\to\infty} E[\int_0^{\tau^\eta} X^{M,\eta}_1(t)/\eta dt] = \int_0^{\tau} \bar{x}^M_1(t)dt$.

From the results in Lemmas 1 to 4, we obtain Theorem 1 that describes the asymptotic behavior of the long-run average cost excluding the setup cost P in Equation (10).

Theorem 1.

$$\lim_{\eta\to\infty} \frac{1}{\eta} \frac{E\Big[\sum_{i=1}^M C_i \sum_{j=1}^{n_i} X^{i,\eta}_j(\tau^\eta) + L\int_0^{\tau^\eta} X^{M,\eta}_1(t)dt\Big]}{E[\tau^\eta]} = \frac{\sum_{i=1}^M C_i \sum_{j=1}^{n_i} \bar{x}^i_j(\bar{\tau}) + \int_0^{\bar{\tau}} L\bar{x}^M_1(t)dt}{\bar{\tau}}. \quad (12)$$

With the result in Theorem 1, the long-run average cost $f_0(\boldsymbol{\gamma})$ in Equation (2) for a large-scale system with N units can be approximated by

$$\tilde{f}_0(\boldsymbol{\gamma}) = \frac{P + \sum_{i=1}^M C_i N\bar{x}^i(\bar{\tau}) + L\int_0^{\bar{\tau}} N\bar{x}^M(t)dt}{\bar{\tau}}, \quad (13)$$

where $\bar{x}^i(\bar{\tau}) = \sum_{j=1}^{n_i} \bar{x}^i_j(\bar{\tau})$ is the asymptotic fraction of units at State i, $i = 1, 2, \dots, M$, at time $\bar{\tau}$. Also, the original random hitting time τ in Equation (1) converges to $\bar{\tau}$ due to Lemma 1. With these asymptotic properties, the original optimization formulation in Equation (2) can be approximated as

$$\underset{\boldsymbol{\gamma}}{\text{minimize}}\ \tilde{f}_0(\boldsymbol{\gamma}) = \frac{P + \sum_{i=1}^M C_i N\bar{x}^i(\bar{\tau}) + L\int_0^{\bar{\tau}} N\bar{x}^M(t)dt}{\bar{\tau}}, \quad (14)$$

where

$$\bar{\tau} = \inf\big\{t : \big(\bar{x}^1(t) - \gamma_1\big)\big(\gamma_2 - \bar{x}^2(t)\big)\cdots \big(\gamma_{M-1} - \bar{x}^{M-1}(t)\big)\big(\gamma_M - \bar{x}^M(t)\big) \le 0\big\}. \quad (15)$$

Note that $\bar{x}^i(t)$, $i = 1, 2, \dots, M$, at time t is deterministic. Therefore, the optimization formulation in Equations (14) and (15) greatly simplifies the original formulation in Equation (2) by estimating the unknown quantities, $E[X^i(\tau)]$ and $E[\int_0^\tau X^M(t)]$, and the random quantity, τ, with the asymptotic values to which they converge.

Regarding the convergence rate, from the fluid limit (Kurtz, 1978), we note that

$$\frac{\mathbf{S}^\eta(t)}{\eta} = \bar{\mathbf{s}}(t) + O\Big(\frac{1}{\sqrt{\eta}}\Big).$$

Therefore, with N units in a system, the convergence rate of the fraction vector of the system's health state to the fluid limit $\bar{s}(t)$ becomes $O(1/\sqrt{N})$. As our asymptotic results in expectation are based on the fluid limits, they have the same convergence rate; i.e., $O(1/\sqrt{N})$ for a system with N units.

The value of the objective function in Equation (14) depends on $\bar{\tau}$, which is determined from the threshold vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_M)$ in Equation (15). That is, as the threshold vector changes, $\bar{\tau}$ and the objective value change accordingly. Therefore, finding the optimal maintenance policy implies obtaining the optimal threshold vector $\boldsymbol{\gamma}^*$ that minimizes the asymptotic long-run average cost (Section 3.4 explains how to find the optimal solution).

3.3. Relationship with scheduled maintenance

A scheduled maintenance that repairs units with a fixed cycle is one of the most convenient maintenance practices for large-scale systems. Even though the threshold-based maintenance strategy proposed in this study is devised to invoke repairs based on the units' health conditions, the results provide useful insights into determining a fixed "time-to-maintenance" schedule for large-scale systems. Note that $\bar{\tau}$ is linked to the threshold vector $\boldsymbol{\gamma}$ in Equation (15). That is, given γ, the maintenance cycle $\bar{\tau}$ is determined and *vice versa*. Consequently, finding the optimal threshold vector $\boldsymbol{\gamma}^*$ that minimizes the long-run average cost can be translated into finding the optimal maintenance cycle $\bar{\tau}^*$.

Again, it is worthwhile to highlight three reasons why our approach for finding the optimal renewal cycle $\bar{\tau}^*$ differs from the existing renewal-based literature. First, contrary to the existing studies that consider a fixed renewal cycle, the renewal cycle τ in our problem is originally a random variable, due to τ being defined as the stopping time when any fraction of the units at each degradation state hits the corresponding threshold value. Therefore, the derivation of the optimal renewal cycle $\bar{\tau}^*$ in this study is more mathematically sophisticated. Second, our threshold vector $\boldsymbol{\gamma}^*$, which determines the optimal maintenance cycle $\bar{\tau}^*$, identifies the system-level health status that needs maintenance. Therefore, we can view $\bar{\tau}^*$ as a condition-based renewal cycle. Third, the resulting renewal cycle produces a cost-effective maintenance strategy for multi-unit systems, unlike existing studies that focus on single-unit systems.

3.4. Solution procedure

Although we simplify the optimization model using the asymptotic properties, solving the problem in Equations (14) and (15) using standard nonlinear optimization techniques is still difficult for a couple of reasons. First, the objective function in Equation (14) is non-convex, and the closed forms of the $\bar{x}^i(\bar{\tau})$s and $\int_0^{\bar{\tau}} \bar{x}^M_1(t)dt$ take very complicated forms. Therefore, the derivative of the objective function is difficult to obtain. Second, there are numerous many combinations of γ_is to be considered.

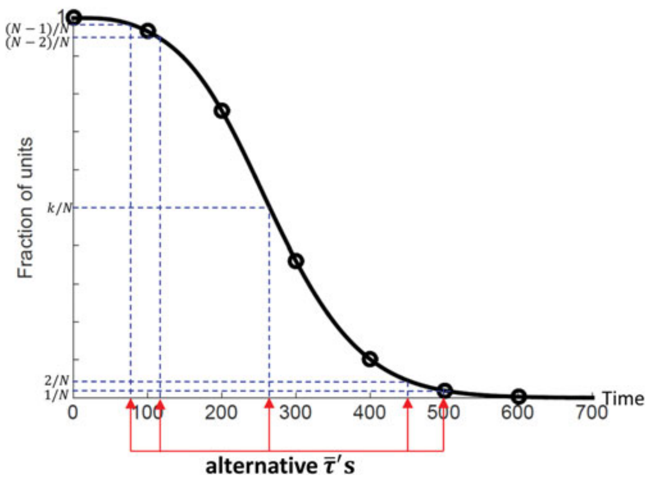


Figure 2. Asymptotic fraction of units at State 1, $\bar{x}^1(t)$, over time.

In this section, we present a computationally efficient method to find the optimal threshold vector $\boldsymbol{\gamma}^*$. The basic idea is to reduce the dimension of the optimization problem. Recall that the objective function depends on $\bar{\tau}$, which is a function of the thresholds $\boldsymbol{\gamma}_i$, $i = 1, 2, \ldots, M$, as in Equation (15). Therefore, instead of solving the problem with respect to the M-dimensional threshold vector, we solve the problem with respect to $\bar{\tau}$ (one-dimensional problem) and obtain the $\boldsymbol{\gamma}^*$ that corresponds to the optimal $\bar{\tau}^*$ value.

To find the optimal renewal cycle $\bar{\tau}^*$ in the asymptotic formulation, we solve the one-dimensional optimization problem as

$$\bar{\tau}^* = \operatorname*{argmin}_{\bar{\tau}} \frac{P + \sum_{i=1}^{M} C_i \bar{x}^i(\bar{\tau}) + \int_0^{\bar{\tau}} L\bar{x}^M(t)dt}{\bar{\tau}}. \quad (16)$$

One possible approach to find $\bar{\tau}^*$ is to approximately solve the optimization problem in Equation (16) by discretizing $\bar{\tau}$ with a sufficient granularity and evaluating the objective function at the discretized time points. This approach, however, requires us to define the granularity level and know the upper bound of $\bar{\tau}$. We overcome the challenges by linking the asymptotic renewal cycle $\bar{\tau}$ with the threshold vector $\boldsymbol{\gamma}$. We note that the asymptotic fraction of units at State 1, $\bar{x}^1(t)$, is non-increasing over time on (0, 1]. For example, Fig. 2 illustrates that $\bar{x}^1(t)$, obtained from the distribution parameters described in Section 4, is a monotone function of time t. Considering that with N number of units, the fraction of units at State 1 decreases with a decrement of $1/N$ as the units transit to the next state over time, we only explore the $\bar{\tau}$s at which the asymptotic fraction of units at State 1 reaches k/N for $k = 1, \ldots, N-1$. Then, we find the optimal $\bar{\tau}^*$ that minimizes Equation (14) among the alternative $\bar{\tau}$s (see Fig. 2). The optimal solution always exists, as the proposed solution approach exhaustively searches all possible solutions and chooses the best solution. The solution uniqueness is, however, not guaranteed, as among all the possible $\bar{\tau}$s, more than one $\bar{\tau}$ can give the same minimum cost.

With the resulting $\bar{\tau}^*$, we obtain the optimal threshold γ_i^*, $i = 1, \ldots, M$, as, for $i = 2, \ldots, M-1$ and $\epsilon > 0$:

$$\gamma_i^* = \begin{cases} \bar{x}^i(\bar{\tau}^*) \text{ if } \max\limits_{0\le t\le\bar{\tau}^*} \bar{x}^i(t) \le \bar{x}^i(\bar{\tau}^*), \\ 1+\epsilon \quad \text{otherwise,} \end{cases} \quad (17)$$

and for $i = 1, M$

$$\gamma_i^* = \bar{x}_i(\bar{\tau}^*). \quad (18)$$

In Equations (17) and (18), the optimal threshold γ_i^* is the asymptotic fraction of units at each state at time $\bar{\tau}^*$ (see Fig. 3(a)). However, there is one exception we need to consider. For the intermediate states, $\bar{x}^i(t)$, $i = 2, \ldots, M-1$, is a unimodal function of time t, having one maximum point. Recall that τ is the first passage time to hit any of the thresholds at M states. If $\max_{0\le t\le\bar{\tau}^*} \bar{x}^i(t) > \bar{x}^i(\bar{\tau}^*)$, there exists a $\bar{\tau}^+ < \bar{\tau}^*$ such that $\bar{x}^i(\bar{\tau}^+) = \bar{x}^i(\bar{\tau}^*)$ (see Fig. 3(b)). In this case, if we set $\gamma_i^* = \bar{x}^i(\bar{\tau}^*)$, γ_i^* actually yields a hitting time $\bar{\tau}^+$ instead of $\bar{\tau}^*$. Therefore, we set $\gamma_i^* = 1+\epsilon$ for some $\epsilon > 0$ to avoid allowing $\bar{x}^i(t)$ to hit the threshold. Note that this adjustment does not affect the optimality because, as the objective function in Equation (16) is minimized at $\bar{\tau}^*$ among the possible alternative $\bar{\tau}$s.

Our implementation results with a wide range of parameter settings, described in Section 4, suggest that $\max_{0\le t\le\bar{\tau}^*} \bar{x}^i(t) \le \bar{x}^i(\bar{\tau}^*)$ holds in general, so we do not need the adjustment of any thresholds, but we include the relevant discussions in the previous paragraphs for the method's general applicability. For the first and the last states (i.e., $i = 1, M$), we always set $\gamma_i^* = \bar{x}^i(\bar{\tau}^*)$, due to $\bar{x}^i(t)$ being a monotone function of time t.

We briefly summarize the procedure:

Step 1: Obtain the asymptotic fraction of units at each state, $\bar{x}^i(t)$, $i = 1, 2, \ldots, M$, over time by numerically solving the system of ODEs (see Section IV of Ko and Byon (2015) for more details).

Step 2: Obtain $\bar{\tau}$s when the asymptotic fraction of units at State 1, $\bar{x}^1(\bar{\tau})$, reaches k/N for $k = 1, \ldots, N-1$.

Step 3: Among the alternatives of $\bar{\tau}$s obtained in Step 2, find the optimal $\bar{\tau}^*$ that minimizes the objective function in Equation (16).

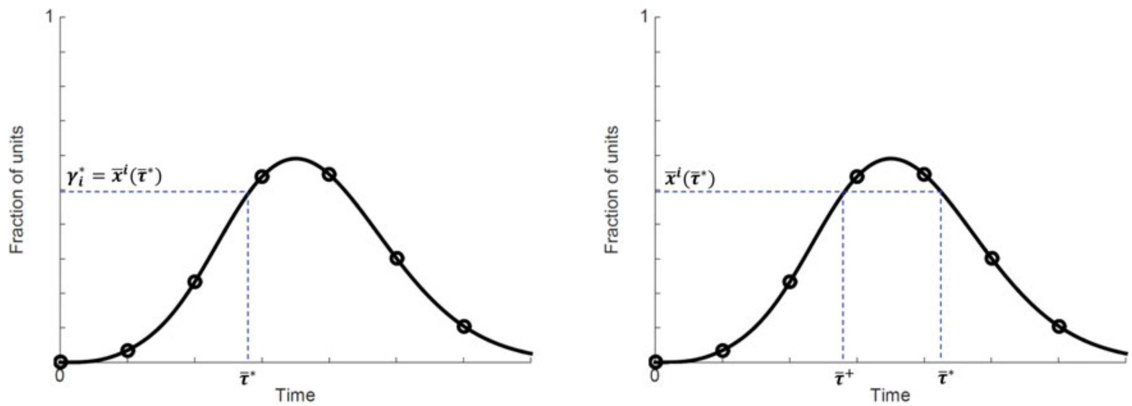
Step 4: Find $\boldsymbol{\gamma}^*$ from Equations (17) and (18).

The proposed procedure is computationally efficient. For example, using a standard desktop computer solves the maintenance optimization problem for a system with 100 units in a few seconds. In fact, increasing the number of units in a system does not increase the computational burden, as the proposed procedure is based on asymptotic properties.

3.5. Extension to a system with multiple heterogeneous groups

We can extend our formulation to maintain a system with multiple heterogeneous groups, where the units across groups are heterogeneous and the units in the same group are homogeneous. For example, in a solar power system, a utility company operates multiple solar parks, and each park has a large number of solar panels with the same specification from the same manufacturer, but different solar parks may use different solar panels.

Suppose there are G heterogeneous groups of units and that Group g has N_g units. As our approach relies on the asymptotic results, we assume that N_g is sufficiently large. The degradation states in the units in Group g can be divided into M_g number of states $\{1, 2, \ldots, M_g\}$. Group g has a setup cost P_g per visit, a revenue loss L_g per period for a failed unit,



(a) When $\max_{0\leq t\leq \bar{\tau}^*} \bar{x}^i(t) \leq \bar{x}^i(\bar{\tau}^*)$

(b) When $\max_{0\leq t\leq \bar{\tau}^*} \bar{x}^i(t) > \bar{x}^i(\bar{\tau}^*)$

Figure 3. Obtaining optimal thresholds for intermediate states.

and a repair cost $C_{i,g}$ per unit at State i. The fluid limit for State i in Group g is denoted by $\bar{x}^{i,g}$. We define the threshold vector $\boldsymbol{\gamma} = (\gamma_{1,1}, \ldots, \gamma_{M_1,1}, \ldots, \gamma_{1,G}, \ldots, \gamma_{M_G,G})$. For the maintenance policy that triggers the repair activity when a certain number of units in a group exceeds a threshold, the optimization problem can be formulated as

$$\underset{\boldsymbol{\gamma}}{\text{minimize}}\ \tilde{f}_0^{\text{ext}}(\boldsymbol{\gamma}) = \sum_{g=1}^{G} \frac{P_g + \sum_{i=1}^{M_g} C_{i,g} N_g \bar{x}^{i,g}(\bar{\tau}) + L_g \int_0^{\bar{\tau}} N_g \bar{x}^{M_g,g}(t)dt}{\bar{\tau}}, \quad (19)$$

where

$$\begin{aligned}\bar{\tau} = \inf\Big\{t : &\big(\bar{x}^{1,1}(t) - \gamma_{1,1}\big)\big(\gamma_{2,1} - \bar{x}^{2,1}(t)\big)\cdots\\ &\big(\gamma_{M_1-1,1} - \bar{x}^{M_1-1,1}(t)\big)\big(\gamma_{M_1,1} - \bar{x}^{M_1,1}(t)\big)\\ &\big(\bar{x}^{1,2}(t) - \gamma_{1,2}\big)\big(\gamma_{2,2} - \bar{x}^{2,2}(t)\big)\cdots\\ &\big(\gamma_{M_2-1,2} - \bar{x}^{M_2-1,2}(t)\big)\big(\gamma_{M_2,2} - \bar{x}^{M_2,2}(t)\big)\\ &\cdots\\ &\big(\bar{x}^{1,G}(t) - \gamma_{1,G}\big)\big(\gamma_{2,G} - \bar{x}^{2,G}(t)\big)\cdots\\ &\big(\gamma_{M_G-1,G} - \bar{x}^{M_G-1,G}(t)\big)\big(\gamma_{M_G,G} - \bar{x}^{M_G,G}(t)\big) \leq 0\Big\}.\end{aligned} \quad (20)$$

The procedure to solve the problem in Equations (19) and (20) is similar to that in Section 3.4, so we omit the procedure to save space.

The optimization formulation can be flexibly modified to fit the application context. For example, the triggering rule can be redefined; e.g., "when the total number of units in a specific condition in all groups exceeds a certain threshold."

4. Numerical results

This section demonstrates the proposed maintenance approach for a large-scale system using numerical examples.

4.1. Implementation setting

We consider a system similar to Tian and Liao (2011), who analyze the degradation patterns of bearings using vibration measurements collected from accelerometers. In Tian and Liao (2011), the reliability of bearings is modeled using the proportional hazard model, with the Weibull distribution being used for the baseline hazard function. Although our problem setting differs from that in Tian and Liao (2011) in many aspects, we borrow some of their distribution and cost parameters.

In determining the number of degradation states, a too small M will not differentiate the evolving health condition of units, whereas a too large M will require estimations of large numbers of parameters. In our implementation, we consider four degradation states, namely, normal, alert, alarm, and failure states (i.e., $M = 4$), as we feel that it is an appropriate number of states in actual implementations and it has been used in the literature (Maillart, 2006; Maillart and Zheltova, 2007; Byon, 2013; Ko and Byon, 2015; Perez *et al.*, 2015). We use the Weibull distribution to represent the transition distribution F_i from State i to State $i+1$ for $i = 1, 2, 3$. We use the shape parameter of 3.05 in F_i, $\forall i$, as in Tian and Liao (2011). We use 300, 200, and 168 for F_i, $i = 1, 2, 3$, respectively, for the scale parameters in our transition distributions.

In approximating each transition distribution with a phase-type distribution, the approximation quality generally increases as n_i increases, as the phase-type distribution is known to be dense in all positive-support distributions (Asmussen *et al.*, 1996; Ko and Byon, 2015). However, a large n_i requires a large number of parameter estimations in fitting the phase-type distributions. In this study, we select an n_i large enough to provide a good approximation quality based on the study by Ko and Byon (2015).

At each maintenance visit, different repair costs for individual units are required, depending on the degradation states. We use $C_1 = \$300$, $C_2 = \$600$, $C_3 = \$1800$, and $C_4 = \$16\,300$ for each state. We use the profit loss of \$10 per day during each unit's failure. We consider different setup costs and study how the optimal policy changes accordingly. We let λ denote the ratio of setup cost P to the major preventive repair cost C_3 and call λ the setup cost ratio. In our implementation, we consider $\lambda = 2, 5, 10, 20, 30$, and 50 corresponding to the setup

Table 1. Comparison of the results from the proposed analytical approach with simulation results: Average maintenance costs of 100 units per day.

Setup cost ratio	*Analytical results (unit: $)*	*Simulation results (unit: $)*
2	210	211 (1.3)
5	234	234 (1.4)
10	272	272 (1.6)
20	344	343 (2.0)
30	412	411 (2.6)
50	538	537 (3.1)

Note. Numbers in parentheses in the third column are the sample standard deviations from 1000 simulation instances.

costs of $3600, $9000, $18 000, $36 000, $54 000, and $90 000, respectively.

4.2. *Implementation results*

First, to validate the proposed analytical approach, we compare the analytical results with the simulation results. In the simulation study, we simulate the operations of a system with 100 units for a sufficiently long time horizon (60 000 periods) using the resulting policy from the analytical approach. Table 1 summarizes the long-run average costs of 100 units per day in a range of setup cost ratios. The second column lists the optimal costs, $\tilde{f}_0(\boldsymbol{\gamma})$ in Equation (14), from the analytical procedure. In each case, the CPU time to analytically obtain the optimal solution is about 0.3 seconds under a regular PC environment. The third column presents the average costs obtained from 1000 independent simulations. Note that the costs from the analytical procedure and simulations are similar in all setup cost ratios, demonstrating the accuracy of the proposed analytical approach.

Figure 4 shows the optimal thresholds γ_1, γ_2, γ_3, and γ_4 for the units in the normal, alert, alarm, and failure states, respectively, that trigger the maintenance activity. Recall that the meaning of γ_1 is different from other γ_is for $i > 1$. When the fraction of units in the normal state reaches (or becomes less than) γ_1, we perform the maintenance. For other states, the maintenance is performed when the fraction reaches (or exceeds) the corresponding thresholds. For example, with $\lambda = 2$, maintenance is triggered:

- when the faction of the units in the normal state hits 67.4%;
- when the fraction of the units in the alert state hits 30.5%;
- when the fraction of the units in the alarm state hits 2.0%;

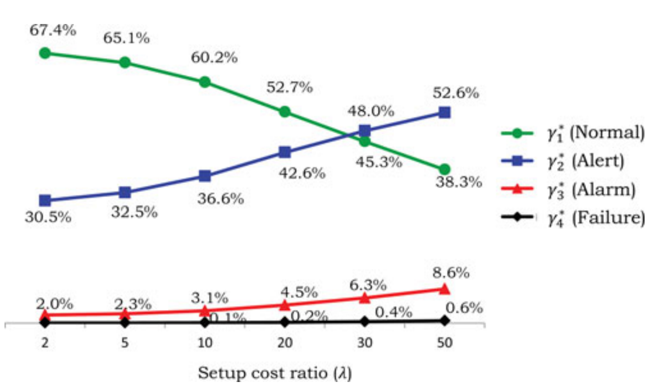


Figure 4. Optimal thresholds at each degradation state with different setup cost ratios.

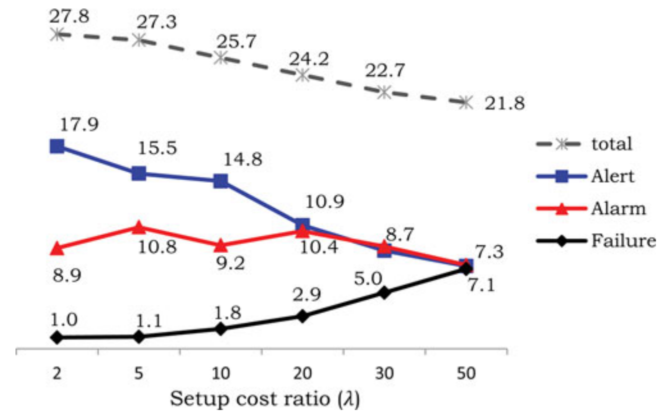


Figure 5. Average number of maintenance visits over 6000 periods when the system's health condition hits the optimal fraction at each degradation state; dotted line denotes the total number of visits during 6000 periods.

- or when the fraction of the units in the failure state hits 0.1%.

As the setup cost increases, the optimal threshold for the normal state, γ_1, decreases, whereas the optimal thresholds for other states increase. These results suggest that with higher setup costs, the system operator can wait until more units deteriorate in order to minimize the long-run average cost. Note that no γ_i^*s are set to be $1+\epsilon$ for some $\epsilon > 0$, implying that no threshold adjustments are needed in all cases, as seen in Fig. 3(a).

Figure 5 shows how often the maintenance action is invoked under the proposed threshold-based strategy. The dotted line shows the average number of maintenance visits during 6000 periods from 1000 simulation instances. As the setup cost ratio increases, the number of visits decreases. Figure 5 also shows that the system's health condition triggering a maintenance visit varies, depending on the setup cost ratios. When $\lambda = 2$, on average, 17.9 out of 27.8 visits are caused when the fraction of units in the alert state reaches the optimal threshold γ_2^* (30.5% in Fig. 4), and another 8.9 and 1.0 visits are invoked when the system's condition hits the optimal fractions, γ_3^* and γ_4^*, in the alarm and failure states (2.0% and 0.1% in Fig. 4), respectively. As the setup cost ratio increases, fewer visits are triggered when the system's health status hits the optimal alert threshold γ_2^*, whereas more visits are performed when the system's health status hits the failure threshold γ_4^*. The number of visits resulting from hitting the alarm threshold γ_3^* slightly varies in a range of setup cost ratios. Figure 5 does not show the number of visits caused by hitting γ_1^*, as the system status hits other thresholds before hitting γ_1^* in all cases considered in our implementation.

4.3. *Comparison with other maintenance strategies*

As discussed earlier, scheduled maintenance is one of the most convenient practices for large-scale system maintenance when units need to be regularly maintained. We compare the proposed approach with a degradation-based scheduled maintenance strategy. In particular, we consider two fixed maintenance cycles; one is the average transition time from normal to alert state—i.e., 268 days—and the other from normal to alarm state; i.e., 447 days—obtained from the transition distributions described in Section 4.1. We call the two strategies $SchM_{268}$ and $SchM_{447}$, respectively.

Table 2. Comparison of the proposed approach with other scheduled maintenance policies.

Setup cost ratio	*Proposed approach* $\bar{\tau}^*$ (unit: days)	*Proposed approach* Cost[a]	$SchM_{268}$ Cost[a,b]	$SchM_{447}$ Cost[a,b]
2	222	210	222 (2.2)	681 (9.8)
5	228	234	243 (2.3)	694 (9.9)
10	240	272	276 (2.2)	714 (10.0)
20	258	344	344 (2.3)	754 (10.1)
30	276	412	410 (2.2)	794 (10.2)
50	294	538	544 (2.3)	875 (10.1)

[a] Long-run average maintenance costs per day for a system with 100 units (unit: $).
[b] Numbers in parentheses are the sample standard deviations from 1000 simulation instances.

Table 3. Sensitivity analysis with different cost parameter settings.

		Proposed approach			$SchM_{268}$	
		$\bar{\tau}^*$	Cost[a]	Number of failures[b]	Cost[a]	Number of failures[b]
C_3	1000	236	202	2.4	204	6.0
	3000	208	220	0.9	245	6.1
	8000	180	244	0.3	347	6.1
C_4	5000	231	207	1.9	209	6.1
	30 000	213	213	1.1	234	6.0
	50 000	208	217	0.9	253	5.9
L	100	221	211	1.5	223	6.1
	500	213	213	1.1	234	6.1
	1000	210	215	1.0	250	6.2

[a] Long-run average maintenance cost per day for a system with 100 units (unit: $).
[b] Average number of failures of 100 units during 6000 periods.

We simulate the system operations with $SchM_{268}$ and $SchM_{447}$ strategies. Table 2 summarizes the results. The second and third columns list the optimal renewal cycles $\bar{\tau}^*$ and optimal average costs per day for 100 units, respectively, both of which are obtained from the proposed analytical approach. The last two columns include the average costs from 1000 instances of $SchM_{268}$ and $SchM_{447}$. We can see that the costs from the proposed approach are generally lower than those in $SchM_{268}$ and $SchM_{447}$.

The differences in the average costs between our approach and $SchM_{268}$ are less significant, compared with the difference between our approach and $SchM_{447}$, as the optimal renewal cycles $\bar{\tau}^*$ in the second column are close to the maintenance cycle of $SchM_{268}$ in these specific cases. As discussed in Section 3.3, asymptotically implementing the threshold-based maintenance strategy with the optimal threshold vector $\boldsymbol{\gamma}^*$ provides a long-run average cost similar to the cost provided by the scheduled maintenance with the optimal renewal cycle $\bar{\tau}^*$. This explains that the result from the proposed approach is similar to that of $SchM_{268}$, especially when the setup cost ratio λ is 20 or 30. We believe that the slightly worse performance of our approach for $\lambda = 30$ is due to the inherent randomness in simulation runs and the possible approximation error in our approach that is based on the asymptotic properties. However, the difference appears to be negligible.

Although our approach can present similar performances with the degradation process-based scheduled maintenance such as $SchM_{268}$ at certain circumstances, the important difference is that our approach finds the optimal renewal cycle by considering both the degradation pattern and cost structures, whereas $SchM_{268}$ only considers the average transition time from normal to alert state regardless of the cost structure. Overall, the results indicate that the proposed approach is more cost-effective than the degradation-based scheduled maintenance strategy.

Remark 1: In Table 1, the standard deviations from multiple runs that simulate the proposed maintenance strategy vary, depending on a range of different setup costs, because the long-run average cost depends on the number of units $X^i(\bar{\tau}^*)$ in each degradation state when the maintenance action is invoked and the distributions of $X^i(\bar{\tau}^*)$s depend on the setup cost ratios. On the other hand, when we use scheduled maintenance such as $SchM_{268}$, the same maintenance cycle is used for all setup cost ratios. As a result, we obtain the same distribution of the number of units in each state. The setup cost does not affect the variance of the long-run average cost, as it is constant. Therefore, we obtain similar standard deviations obtained from simulations of $SchM_{268}$ as shown in Table 2.

4.4. Sensitivity analysis

Although our proposed formulation focuses on minimizing the long-run average cost, the cost structure allows us to consider reliability and availability. For the system that requires a high reliability, we can impose a large repair cost for a failed unit; in this case, the repair cost can include a direct cost and other indirect costs caused by the failure. Similarly, for achieving a high availability, we can impose a large revenue loss L, which will trigger maintenance activity before a failure occurs.

To further investigate the advantage of the proposed approach, we perform a sensitivity analysis in a range of different parameter settings. We use the implementation setting described in Section 4.1 (with the setup cost of $3600) as a baseline setting and vary each parameter to study its impact on the performances of the proposed condition-based maintenance strategy and $SchM_{268}$. In our sensitivity study, we vary the values of C_3, C_4, and L.

Table 3 summarizes the results of the two approaches. The first two columns indicate the cost parameters that are changed from the baseline setting. We also include the average number of failures during 6000 periods to quantify the reliability levels of the system. We omit the availability results because they are over 99.9% in all cases. We summarize our observations as follows:

1. In the proposed approach, when each cost parameter increases, the maintenance cycle $\bar{\tau}^*$ decreases, suggesting a shorter maintenance cycle, and the long-run average cost per period gradually increases. However, as C_3, C_4, or L increases, the number of failures decreases, as the maintenance activity is triggered before units reach the failed state to offset high repair costs or revenue losses.
2. The average costs from $SchM_{268}$ are higher than those given by the proposed approach. Moreover, compared with the proposed approach, the average cost in $SchM_{268}$ sharply increases as C_3, C_4, or L increases. This result indicates that the performance of $SchM_{268}$ in terms of long-run maintenance costs is sensitive to the cost parameters, as the scheduled maintenance strategy does

Table 4. Comparison of the results from the proposed analytical approach with simulation results in different degradation settings: Long-run average maintenance cost per period.

Degradation pattern	*Number of units*	*Analytical results*[a]	*Simulation results*[a,b]
	30	37	37 (0.5)
Slower degradation	50	57	56 (0.7)
	100	105	105 (0.9)
	30	74	74 (0.8)
Baseline degradation	50	113	114 (0.9)
	100	210	211 (1.3)
	30	148	148 (1.1)
Faster degradation	50	226	227 (1.3)
	100	421	422 (1.8)

[a] Average costs per period for the number of units in the second column (unit: $).
[b] Average costs from 1000 simulation instances, each with 60 000 periods. Numbers in parentheses are the sample standard deviations.

not consider the cost structure. The number of failures in SchM_{268} is similar across the different parameter settings, as the same degradation pattern is used in all cases and SchM_{268} performs repairs every 268 periods.

Overall, the proposed approach provides more meaningful results by considering repair costs, effects of failures, and degradation patterns in an integrative manner, compared with the degradation-based scheduled maintenance strategy.

Next, to investigate the convergence quality of our approach with the characteristics of degradation processes and the numerical gap due to our approximation, we compare the analytical results from the proposed approach with the simulation results by using a set of different degradation processes and number of units. For the degradation process, we consider three cases: (i) baseline degradation process with the transition distributions, F_i, $i = 1, 2, 3$, described in Section 4.1; (ii) slower degradation process with scale parameters of 600, 400, and 336 for the transition distributions (note that we multiply the original scale parameters in F_i by two, so the average transition time from one state to the next deteriorated state becomes twice longer); and (iii) faster degradation with scale parameters of 150, 100, and 84 (note that we divide the original scale parameters in F_i by two). In all cases, we use the same shape parameter of 3.05 in the transition distributions. For the number of units, we consider 30, 50, and 100.

Table 4 lists the results. The results from the proposed analytical approximation method in general coincide with the simulation results for sufficiently long periods. In other words, the proposed approach shows a satisfactory convergence quality in mid to large-sized problem settings with different degradation characteristics.

5. Summary

This study proposes an optimal maintenance policy for systems with a large number of homogeneous units. We consider the maintenance type to renew all units simultaneously at each maintenance visit. We categorize the degradation conditions into a finite number of states and formulate the optimization problem such that maintenance is performed when a certain fraction of units at each degradation state hits its corresponding threshold. To make the analysis scalable for large-scale systems, we solve the problem using relevant asymptotic theories and propose a computationally efficient solution procedure. We implement the proposed approach using a diverse set of cost parameters and investigate their effects on the maintenance costs and optimal thresholds. Our implementation results support the cost-effectiveness of the proposed approach, compared with other alternative maintenance strategies for large-scale systems.

We believe that our study is a first step to optimize the maintenance scheduling for large, multiple-unit systems. Based on the results in this study, we plan to extend the approach to more general cases. For example, selective maintenance that repairs high-risk units is likely more cost-effective than simultaneous maintenance when the repair cost of each unit is high and/or repair resources are limited. In general, repairing major components such as blades and gearboxes in wind turbines falls into the selective maintenance category (Lee *et al.*, 2013; Yampikulsakul *et al.*, 2014; Choe *et al.*, 2015). We also plan to develop a new maintenance strategy that considers incomplete maintenance. In another extension of the research presented in this article, we plan to address the statistical/structural dependence and heterogeneous degradation processes among units in optimizing maintenance schedules for large-scale systems. Finally, in the transition distribution approximation with phase-type distributions, determining an appropriate number of phases is beyond the scope of this study, but we will actively look for algorithms that provide an (near) optimal number of phases.

Acknowledgments

We would like to thank the Editor, Department Editor, and anonymous reviewers for their constructive comments on various aspects of this work.

Funding

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1005019), and in part by the U.S. National Science Foundation grants CMMI-1362513 and CMMI-1536924.

Notes on contributors

Young Myoung Ko is an assistant professor of Industrial and Management Engineering at Pohang University of Science and Technology (POSTECH), Korea. He received his B.S. and M.S. degrees in Industrial Engineering from Seoul National University, and his Ph.D. degree in Industrial Engineering from Texas A&M University in 1998, 2000, and 2011, respectively. His research focuses on the analysis and optimization of large-scale stochastic systems and covers the domains of service systems, telecommunication networks, energy-efficient infrastructure, and renewable energy systems.

Eunshin Byon is an assistant professor with the Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor. She received her Ph.D. degree in Industrial and Systems Engineering from Texas A&M University, College Station, in 2010. Her research interests include data analytics, quality and reliability engineering, system infomatics and uncertainty quantification. She has received several Best Paper Awards including the 2015 Best Applications Paper Award from IIE Transactions on Quality & Reliability Engineering. Prof. Byon is a member of IIE, INFORMS, IEEE, and ASQ.

References

Arnold, B.C. and Groeneveld, R.A. (1979) Bounds on expectations of linear systematic statistics based on dependent samples. *The Annals of Statistics*, **7**(1), 220–223.

Asmussen, S., Nerman, O. and Olsson, M. (1996) Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, **23**(4), 419–441.

Bian, L. and Gebraeel, N. (2014) Stochastic modeling and real-time prognostics for multi-component systems with degradation rate interactions. *IIE Transactions*, **46**(5), 470–482.

Byon, E. (2013) Wind turbine operations and maintenance: A tractable approximation of dynamic decision-making. *IIE Transactions*, **45**(11), 1188–1201.

Byon, E., Ntaimo, L. and Ding, Y. (2010) Optimal maintenance strategies for wind turbine systems under stochastic weather conditions. *IEEE Transactions on Reliability*, **59**(2), 393–404.

Castanier, B., Grall, A. and Berenguer, C. (2005) A condition-based maintenance policy with non-periodic inspections for a two-unit series system. *Reliability Engineering & System Safety*, **87**(1), 109–120.

Chen, N. and Tsui, K.L. (2013) Condition monitoring and remaining useful life prediction using degradation signals: Revisited. *IIE Transactions*, **45**(9), 939–952.

Choe, Y., Byon, E. and Chen, N. (2015) Importance sampling for reliability evaluation with stochastic simulation models. *Technometrics*, **57**(3), 351–361.

Dvorak, P. (2014) Should wind-turbine oil changes be interval based or condition based? Accessed 18 August 2015. Available at http://www.windpowerengineering.com/featured/business-news-projects/wind-turbine-oil-changes-interval-based-condition-based/

Ethier, S.N. and Kurtz, T.G. (1986) *Markov Processes: Characterization and Convergence*, first edition, John Wiley & Sons, Inc., Hoboken, NJ.

Folland, G.B. (1999) *Real Analysis: Modern Techniques and Their Applications*, second edition, John Wiley & Sons, Inc., New York, NY.

Guikema, S., Davidson, R. and Liu, H. (2006) Statistical models of the effects of tree trimming on power system outages. *IEEE Transactions on Power Delivery*, **21**(3), 1549–1557.

Hong, Y., Duan, Y., Meeker, W.Q., Stanley, D.L. and Gu, X. (2015) Statistical methods for degradation data with dynamic covariates information and an application to outdoor weathering data. *Technometrics*, **57**(2), 180–193.

Horsburgh, J.S., Reeder, S.L., Jones, A.S. and Meline, J. (2015) Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. *Environmental Modelling & Software*, **70**, 32–44.

Jiang, L., Feng, Q. and Coit, D. (2012) Reliability and maintenance modeling for dependent competing failure processes with shifting failure thresholds. *IEEE Transactions on Reliability*, **61**(4), 932–948.

Kharoufeh, J., Solo, C. and Ulukus, M. (2010) Semi-Markov models for degradation-based reliability. *IIE Transactions*, **42**(8), 599–612.

Ko, Y.M. and Byon, E. (2015) Reliability evaluation of large-scale systems with identical units. *IEEE Transactions on Reliability*, **64**(1), 420–434.

Kulkarni, V.G. (1995) *Modeling and Analysis of Stochastic Systems*, first edition, Chapman & Hall/CRC, Boca Raton, FL.

Kurtz, T. (1978) Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and their Applications*, **6**(3), 223–240.

Lee, G., Byon, E., Ntaimo, L. and Ding, Y. (2013) Bayesian spline method for assessing extreme loads on wind turbines. *Annals of Applied Statistics*, **7**(4), 2034–2061.

Liu, X., Li, J., Al-Khalifa, K.N., Hamouda, A.S., Coit, D.W. and Elsayed, E.A. (2013) Condition-based maintenance for continuously monitored degrading systems with multiple failure modes. *IIE Transactions*, **45**(4), 422–435.

Maillart, L.M. (2006) Maintenance policies for systems with condition monitoring and obvious failures. *IIE Transactions*, **38**(6), 463–475.

Maillart, L.M. and Zheltova, L. (2007) Structured maintenance policies on interior sample paths. *Naval Research Logistics*, **54**(6), 645–655.

Nicolai, R.P. and Dekker, R. (2008) Optimal maintenance of multi-component systems: A review, in *Complex System Maintenance Handbook*, Springer, London, UK, pp. 263–286.

Perez, E., Ntaimo, L. and Ding, Y. (2015) Multi-component wind turbine modeling and simulation for wind farm operations and maintenance. *Simulation*, **91**(4), 360–382.

Shafiee, M. and Finkelstein, M. (2015) A proactive group maintenance policy for continuously monitored deteriorating systems: Application to offshore wind turbines. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, **229**, pp. 373–384.

Song, S., Coit, D., Feng, Q. and Peng, H. (2014) Reliability analysis for multi-component systems subject to multiple dependent competing failure processes. *IEEE Transactions on Reliability*, **63**(1), 331–345.

Tian, Z. and Liao, H. (2011) Condition based maintenance optimization for multi-component systems using proportional hazards model. *Reliability Engineering & System Safety*, **96**(5), 581–589.

Wang, J.B. and Wei, C.M. (2011) Parallel machine scheduling with a deteriorating maintenance activity and total absolute differences penalties. *Applied Mathematics and Computation*, **217**(20), 8093–8099.

Wang, Y. and Pham, H. (2011) A multi-objective optimization of imperfect preventive maintenance policy for dependent competing risk systems with hidden failure. *IEEE Transactions on Reliability*, **60**(4), 770–781.

Williams, D. (1991) *Probability with Martingales*, first edition, Cambridge University Press, New York, NY.

Yampikulsakul, N., Byon, E., Huang, S., Sheng, S. and You, M. (2014) Condition monitoring of wind power system with nonparametric regression analysis. *IEEE Transactions on Energy Conversion*, **29**(2), 288–299.

Ye, Z.S., Wang, Y., Tsui, K.L. and Pecht, M. (2013) Degradation data analysis using Wiener processes with measurement errors. *IEEE Transactions on Reliability*, **62**(4), 772–780.

Appendix

Proof of Lemma 1. Since $X_j^{i,\eta}(t)/\eta \to \bar{x}_j^i(t)$ *a.s.* uniformly on bounded time intervals, it follows that the stopping time $\tau^\eta \to \bar{\tau}$ *a.s.* (Ethier and Kurtz, 1986). □

Proof of Lemma 2. Assuming that there are η number of independent and homogeneous units, let Z_k be the lifetime of unit k. Therefore, the Z_ks are independently and identically distributed and Z_k is the sum of the duration for which unit k spends at each degradation state. Then, from our assumption on transition distributions, we obtain

$$\mathrm{E}[Z_k] = \sum_{i=1}^{M-1} \mathrm{E}[\text{Time spent at state } i] \equiv \mu < \infty, \tag{A1}$$

$$\mathrm{Var}[Z_k] = \sum_{i=1}^{M-1} \mathrm{Var}[\text{Time spent at state } i] \equiv \sigma^2 < \infty. \tag{A2}$$

We define

$$\kappa^\eta = \inf\left\{t : \left(\gamma_M - \frac{X_1^{M,\eta}(t)}{\eta}\right) \leq 0\right\} \tag{A3}$$

$$= \inf\{t : X_1^{M,\eta}(t) \geq \lceil \eta\gamma_M \rceil\}, \tag{A4}$$

where the ceiling function $\lceil a \rceil$ yields the smallest integer not less than a. Then, we obtain $\tau^\eta \leq \kappa^\eta$ because κ^η is one of the M possible hitting times defined in Equation (9).

Now, we show that $\mathrm{E}[\kappa^\eta] \leq K$ for all η and some $K < \infty$. Let $Z_{(k)}$ be the kth-order statistic. Then, with $\kappa^\eta = Z_{(\lceil \eta\gamma^M \rceil)}$, we obtain

$$\mathrm{E}[\tau^\eta] \leq \mathrm{E}[\kappa^\eta] \tag{A5}$$

$$= \mathrm{E}[Z_{(\lceil \eta\gamma^M \rceil)}] \tag{A6}$$

$$\leq \mu + \sigma\sqrt{\frac{\lceil \eta\gamma^M \rceil - 1}{\eta - \lceil \eta\gamma^M \rceil + 1}}$$

by Arnold and Groeneveld (1979) (A7)

$$\leq \mu + \sigma\sqrt{\frac{\eta\gamma^M}{\eta - \eta\gamma^M}} \tag{A8}$$

$$= \mu + \sigma\sqrt{\frac{\gamma^M}{1-\gamma^M}} \tag{A9}$$

$$= K < \infty, \tag{A10}$$

where

$$K = \mu + \sigma\sqrt{\frac{\gamma^M}{1-\gamma^M}}.$$

Therefore, by the dominated convergence theorem (Williams, 1991), we have

$$\lim_{\eta\to\infty} \mathrm{E}[\tau^\eta] = \mathrm{E}\left[\lim_{\eta\to\infty} \tau^\eta\right] = \bar{\tau}, \tag{A11}$$

where the last equality is due to Lemma 1. □

Proof of Lemma 3. Since $|X_j^{i,\eta}(\tau^\eta)/\eta| \leq 1$ for all η, $\mathrm{E}[X_j^{i,\eta}(\tau^\eta)/\eta] \leq 1 < \infty$. Then, by the dominated convergence theorem:

$$\lim_{\eta\to\infty} \mathrm{E}\left[\frac{X_j^{i,\eta}(\tau^\eta)}{\eta}\right] = \mathrm{E}\left[\lim_{\eta\to\infty} \frac{X_j^{i,\eta}(\tau^\eta)}{\eta}\right] = \bar{x}_j^i(\bar{\tau}). \tag{A12}$$

□

Proof of Lemma 4.

$$\mathrm{E}\left[\int_0^{\tau^\eta} \frac{X_1^{M,\eta}(t)}{\eta} dt\right]$$

$$= \mathrm{E}\left[\int_0^{\bar{\tau}} \frac{X_1^{M,\eta}(t)}{\eta} dt\right] + \mathrm{E}\left[\int_{\bar{\tau}}^{\tau^\eta} \frac{X_1^{M,\eta}(t)}{\eta} dt\right]. \tag{A13}$$

We will look at each term in the right-hand side of Equation (A13) separately. For the first term, we get

$$\lim_{\eta\to\infty} \mathrm{E}\left[\int_0^{\bar{\tau}} \frac{X_1^{M,\eta}(t)}{\eta} dt\right]$$

$$= \lim_{\eta\to\infty} \int_0^{\bar{\tau}} \mathrm{E}\left[\frac{X_1^{M,\eta}(t)}{\eta}\right] dt \text{ by Fubini theorem} \tag{A14}$$

$$= \int_0^{\bar{\tau}} \lim_{\eta\to\infty} \mathrm{E}\left[\frac{X_1^{M,\eta}(t)}{\eta}\right] dt$$

by dominated convergence theorem (A15)

$$= \int_0^{\bar{\tau}} \bar{x}_1^M(t) dt. \tag{A16}$$

For the second term, we obtain

$$\left|\mathrm{E}\left[\int_{\bar{\tau}}^{\tau^\eta} \frac{X_1^{M,\eta}(t)}{\eta} dt\right]\right| \leq \mathrm{E}\left[\left|\int_{\bar{\tau}}^{\tau^\eta} \frac{X_1^{M,\eta}(t)}{\eta} dt\right|\right]$$

$$\leq E[|\tau^\eta - \bar{\tau}|]. \tag{A17}$$

Taking limits on both sides of Equation (A17) yields

$$\lim_{\eta\to\infty} \left|\mathrm{E}\left[\int_{\bar{\tau}}^{\tau^\eta} \frac{X_1^{M,\eta}(t)}{\eta} dt\right]\right| \leq \lim_{\eta\to\infty} E[|\tau^\eta - \bar{\tau}|] \tag{A18}$$

$$= E[\lim_{\eta\to\infty} |\tau^\eta - \bar{\tau}|] \text{ by dominated convergence theorem} \tag{A19}$$

$$(\text{since } E[\tau^\eta] < K \text{ and } |\tau^\eta - \bar{\tau}| \leq \tau^\eta + \bar{\tau}) \tag{A20}$$

$$= 0. \tag{A21}$$

Therefore, we have

$$\lim_{\eta\to\infty} \mathrm{E}\left[\int_0^{\tau^\eta} \frac{X_1^{M,\eta}(t)}{\eta} dt\right] = \int_0^{\bar{\tau}} \bar{x}^M(t) dt. \tag{A22}$$

□

Proof of Theorem 1. From Lemmas 1 to 4, the theorem is proved. □