Available online at www.sciencedirect.com
SCIENCE DIRECT
automatica
Automatica 42 (2006) 1321 – 1330
www.elsevier.com/locate/automatica

A risk-sensitive approach to total productive maintenance☆

Abhijit Gosavi*

Department of Industrial Engineering, University at Buffalo, 317 Bell Hall, SUNY, Buffalo, NY 14260, USA

Received 16 March 2005; received in revised form 16 January 2006; accepted 4 February 2006
Available online 18 May 2006

Abstract

While risk-sensitive (RS) approaches for designing plans of total productive maintenance are critical in manufacturing systems, there is little in the literature by way of theoretical modeling. Developing such plans often requires the solution of a discrete-time stochastic control-optimization problem. Renewal theory and Markov decision processes (MDPs) are commonly employed tools for solving the underlying problem. The literature on preventive maintenance, for the most part, focuses on minimizing the *expected* net cost, and disregards issues related to minimizing risks. RS maintenance managers employ safety factors to modify the risk-neutral solution in an attempt to heuristically accommodate elements of risk in their decision making. In this paper, our efforts are directed toward developing a formal theory for developing RS preventive-maintenance plans. We employ the Markowitz paradigm in which one seeks to optimize a function of the expected cost and its variance. In particular, we present (i) a result for an RS approach in the setting of renewal processes and (ii) a result for solving an RS MDP. We also provide computational results to demonstrate the efficacy of these results. Finally, the theory developed here is of sufficiently general nature that can be applied to problems in other relevant domains.

Keywords: Variance; Renewal reward theorem; Markov decision processes; Maintenance; Risk sensitive

1. Introduction

Total productive maintenance (TPM) is a management initiative that has been widely embraced in the industry. A positive *strategic* outcome of such implementations is the reduced occurrence of unexpected machine breakdowns that disrupt production and lead to losses which can exceed millions of dollars annually. Additionally, frequent machine breakdowns indirectly can lead to a host of other problems, e.g., difficulties in meeting customer deadlines, which makes the transition from make-to-stock to make-to-order difficult (Suri, 1998) and magnifies the need to keep extra safety stocks, increasing inventory-holding costs (Askin & Goldberg, 2002). An important tool of a TPM program is the stochastic model used to determine the optimal time for preventive maintenance (PM) (Askin & Goldberg, 2002). PM can help reduce the frequency of unexpected repairs when the failure *rate* is of an increasing nature (Das & Sarkar, 1999; Lewis, 1994).

Renewal processes (Kao, 1997; Ross, 1992) and Markov decision processes (MDPs) (Bertsekas, 1995; Puterman, 1994) are frequently used as the underlying stochastic models in a TPM program. A critical drawback of a traditional approach in TPM is to use the *expected* value of the long-run cost as the objective function. Such an approach overlooks the risk associated with the occasional high cost that can occur in system optimized with respect to the expected cost. As a result, risk-sensitive (RS) managers, whom we interacted with in a local automobile industry, modify the predicted optimal (with respect to the expected cost) time for maintenance, τ^*, by using a factor of safety, η, where $\eta > 1$, such that the time for PM is then: τ^*/η. While this certainly results in a more *conservative* time for maintenance, it is a heuristic approach. What managers really need is a more sophisticated approach that would help them (i) quantify their risk sensitivity on a scale from 0 to 1 and (ii) determine the optimal maintenance time using a model that incorporates this factor. This clearly motivates the need for

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Qing Zhang under the direction of Editor Suresh Sethi.
* Tel.: +1 716 645 2357; fax: +1 716 645 3302.
E-mail address: agosavi@buffalo.edu.

0005-1098/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2006.02.006

embedding the well-known Markowitz criterion (Markowitz, 1952) within the stochastic model. Another significant demand of managers from the model is the ability to *quantify* risks in terms of dollars (or Euros) and hours—units that they are comfortable with. In particular, senior managers involved in developing long-term plans for an enterprise are familiar with the idea of using variance per unit time as a measure of risk in *strategic* decision making (see Ruefli, Collins, & Lacugna, 1999 for an extensive survey). Since TPM has a significant strategic impact on the organization, the units of risk in these calculations should ideally match those used in strategic management. One of the goals of this research is to develop models that can be conveniently used by managers.

A general cost function (objective function) using the Markowitz criterion is

$$g(\tau) = \mu_C + \theta\sigma^2 \quad \text{with } \theta > 0, \tag{1}$$

where μ_C and σ^2 denote the long-run mean and the long-run variance, respectively, of the net cost per unit time incurred from following a preventive maintenance plan that prescribes τ as the time for PM. An alternative formulation in terms of *rewards*, in which the objective function is *maximized*, is $g_R(\tau) = \mu_R - \theta\sigma^2$, with $\theta > 0$, where μ_R and σ^2 denote the long-run mean and variance of the net *reward* per unit time, respectively. Since $\mu_R = -\mu_C$, both formulations are *equivalent*.

Risk-neutral (RN) statistical models for PM use $\theta = 0$. Typically, θ is selected by experimentation by the manager and is a function of the variability in the system. A very large value for θ is undesirable, since that could produce a solution with a very low variability but also with a very high cost. This is because a very large value for θ amplifies the importance of the variance and diminishes that of the mean. A very low value for θ, on the other hand, is indicative of a manager who is neutral to risks. Clearly, the smaller the value of θ, the closer the model gets to becoming RN.

The time for PM, it must be understood, is the time since the last repair or PM. A common assumption is that the unit or the line is as good as new when it is repaired or preventively maintained. A typically made second assumption is that when the machine is not working, it is assumed not to age. We will stick to these two assumptions here. The main focus of this paper is to develop a theory when $\theta > 0$. The work of Chen and Jin (2003) also employs the Markowitz criterion, but their approach is quite different than ours; this will be clarified via our discussions below.

TPM plans for the production line in its entirety tend to be distinct from those for individual units that operate independently of the line. Most factories are full of such units, e.g., fork-lift trucks, electrical pumps, etc. We will develop separate models for the individual-unit scenario and the production-line scenario. For the case of the individual unit, we will present a renewal-theory model and for the case of the production line, we will present a more involved model based on MDPs. The analysis will involve presentation of some key results that could be applied to a large number of other management-science problems involving control theory. Thereafter, we will present results from computational experiments with both models. The remainder of this paper is organized as follows. Section 2 presents the renewal-theory model, and Section 3 presents the Markov decision model. Section 4 describes empirical work done using these two models, and Section 5 concludes the paper.

2. A renewal-theory model

A commonly used model in most TPM programs employs renewal theory and goes by the name "age-replacement". In the setting of renewal processes, every failure or a maintenance triggers a so-called *renewal event*. A classical result in renewal theory, called the renewal reward theorem, provides an expression for the expected reward per unit time in a renewal process. An extension of this concept to the variance in the rewards of the renewal process can be found in Chen and Jin (2003). However, we present a different result that leads to a significantly different mechanism for measuring variance. Our result was influenced by the need of managers in a local industry for measuring risk in practical units that could be explained to senior management; the unit of risk (variance) in our result is hence $dollar^2/hour$ or $Euro^2/hour$, which is the same as that for variance in rewards per unit time. As mentioned above, strategic managers are known to use variance to measure risk (Ruefli et al., 1999).

Consider a counting process, $\{N(t), t \geqslant 0\}$, and let T_n denote the time between the $(n-1)$th and the nth event in this process with $n \geqslant 1$. If $\{T_1, T_2, \ldots\}$ denotes a sequence of non-negative random variables that are independent and identically distributed, then the counting process is called a *renewal* process. When there is a reward associated with each renewal, we have a renewal reward process. Let R_n denote the reward associated with the nth renewal or nth cycle. We will let $R(t) = \sum_{n=1}^{N(t)} R_n$ denote the sum of the individual rewards earned by time t and $R^2(t) = \sum_{n=1}^{N(t)} (R_n)^2$ denote the sum of the square of the individual rewards earned by time t. Further, let

$$E[R] \equiv E[R_n], \quad E[R^2] \equiv E[(R_n)^2] \quad \text{and} \quad E[T] \equiv E[T_n],$$

where E denotes the expectation operator. The *long-run* variance in the rewards in Chen and Jin (2003) (see Lemma 1 of their paper) is $\lim_{t\to\infty} R^2(t)/t - [\lim_{t\to\infty} R(t)/t]^2$, which requires a subtraction of two quantities of which the first has the unit $dollar^2$ per $hour$ and the second $dollar^2$ per $hour^2$. We present the following definition for the long-run variance, which measures *a time average of the total variance in infinitely many renewals*: $\sigma^2 = \lim_{t\to\infty} V(t)/t$, where $V(t) = \sum_{n=1}^{N(t)} (R_n - E[R])^2$. Thus σ^2 represents the sum of the squared deviations of the cycle rewards from their means, along an infinite number of renewals, divided by the total duration of the renewals. Hence, it can be interpreted as the average variance per unit time measured over the long run. What is key is that the above definition produces a consistent unit of $dollar^2$ per hour for the variance. We now prove the following result to measure the variance in this style.

Theorem 1. *If* $|E[R]| < \infty$, $E[T] < \infty$, $E[R^2] < \infty$, *then with probability* 1 (*w.p.*1)

$$\sigma^2 \equiv \lim_{t\to\infty} \frac{V(t)}{t} = \frac{E[R^2] - (E[R])^2}{E[T]}. \tag{2}$$

Proof.

$$\begin{aligned}\sigma^2 \equiv \lim_{t\to\infty} \frac{V(t)}{t} &= \lim_{t\to\infty} \left(\frac{V(t)}{N(t)}\right)\left(\frac{N(t)}{t}\right) \\ &= \lim_{t\to\infty} \left(\frac{\sum_{n=1}^{N(t)} (R_n - E[R])^2}{N(t)}\right)\left(\frac{N(t)}{t}\right) \\ &= \frac{E[R^2] - (E[R])^2}{E[T]} \quad \text{w.p.1.} \end{aligned} \tag{3}$$

In the above, equality (3) follows from (i) the fact that

$$\begin{aligned}\frac{\sum_{n=1}^{N(t)} (R_n - E[R])^2}{N(t)} &= \frac{\sum_{n=1}^{N(t)} R_n^2}{N(t)} - 2\frac{\sum_{n=1}^{N(t)} R_n E[R]}{N(t)} \\ &\quad + \frac{\sum_{n=1}^{N(t)} (E[R])^2}{N(t)} \\ &= E[R^2] - (E[R])^2 \quad \text{w.p.1 as } t \to \infty,\end{aligned}$$

(ii) the elementary renewal theorem (see e.g., Ross, 1997, Proposition 7.1, p. 407), which implies that w.p.1 $\lim_{t\to\infty} N(t)/t = 1/E[T]$. □

The renewal reward theorem (see e.g., Ross, 1997, Proposition 7.3, p. 417) for the expected reward per unit time, which we will need in our analysis, is stated next.

Theorem 2. *If* $|E[R]| < \infty$, $E[T] < \infty$, *then w.p.*1

$$\mu_R \equiv \lim_{t\to\infty} \frac{R(t)}{t} = \frac{E[R]}{E[T]}.$$

Hence if C denotes the cost in one cycle, $\mu_C = E[C]/E[T]$.

We now develop an expression for the Markowitz function defined in Eq. (1). Let $F(.)$ denote the *cdf* and $f(.)$ the *pdf*. Further, let c_r denote the cost of one repair and c_m the cost of one PM. Then, if τ denotes the age for PM, the expected cost in a renewal cycle will be

$$E[C] = c_m(1 - F(\tau)) + c_r F(\tau). \tag{4}$$

Similarly, the expected value of the square of the costs in a renewal cycle will be

$$E_\tau[R^2] = E_\tau[C^2] = c_m^2(1 - F(\tau)) + c_r^2 F(\tau). \tag{5}$$

And finally, the expected length of the renewal cycle:

$$E[T] = \int_0^\tau (x + t_r) f(x)\,dx + (\tau + t_m)[1 - F(\tau)], \tag{6}$$

where t_r denotes the expected time for one repair and t_m denotes the expected time for PM. Then we have the following result:

Proposition 3. *If* $|E[R]| < \infty$, $E[T] < \infty$, $|E[R^2]| < \infty$, *then w.p.*1

$$\begin{aligned} g(\tau) = &\frac{c_m(1 - F(\tau)) + c_r F(\tau)}{\int_0^\tau (x + t_r) f(x)\,dx + (\tau + t_m)[1 - F(\tau)]} \\ &+\theta \frac{c_m^2(1-F(\tau)) + c_r^2 F(\tau) - [c_m(1-F(\tau)) + c_r F(\tau)]^2}{\int_0^\tau (x+t_r) f(x)\,dx + (\tau+t_m)[1-F(\tau)]}. \end{aligned}$$

Proof. The proof follows from Theorems 1 and 2, and Eqs. (1), (4)–(6). □

When a closed form expression is not available for $g(\tau)$, the latter can be numerically optimized to obtain τ. We present such computational experiments in Section 4.

A reviewer of this paper pointed out that in case R is constant but T is a random variable, the above mechanism for measuring risk will not work, because then we have that the variance $(E[R^2] - E^2[R])/E[T] = 0$. Note that the model in Chen and Jin (2003) will result in the following expression for variance $E[R^2]/E[T] - (E[R]/E[T])^2$, which will also equal 0 for $E[T] = 1$; but if this expression is used for $E[T] < 1$, the variance will be negative, since $E[R^2] = E^2[R]$ for a constant R. Hence it is stated in Chen and Jin (2003) that their model applies only when $E[T] > 1$. In a scenario where the renewal reward in a cycle is a constant but the time is not, one could use downside risk (Fishburn, 1977; Gan, Sethi, & Yan, 2005; Roy, 1952; Tesler, 1955) or perhaps the more classical utility functions (Von Neumann & Morgenstern, 1953). This also implies that variance is perhaps not a perfect measure for risk. Limitations of variance were pointed out by (Markowitz, 1952). The analysis of downside risk or classical utility functions is, however, beyond the scope of this paper.

3. A model based on Markov decision processes

Consider a system of machines, e.g., forging machines, painting machines and rolls, in a production line. Typically, such systems are failure prone with an increasing failure rate. Let X denote the time to failure of the system. We will make the following assumptions, which are based on the system we studied: (i) when the line fails, it is usually repaired by the next day. (ii) The line is shut down for PM, usually, for the duration of the entire day. At the beginning of each day, the manager has two options: (i) continue with production and (ii) do PM. If the system state space on the nth day is defined by B_n, the number of days elapsed since the last repair or PM, i.e., the age of the line, then $\{B_n | n = 1, 2, \ldots\}$ is a Markov chain under any action. The age of the line will be assumed to be 0 after a repair or a PM. Formally, the chain satisfies the following Markov property under any action: $\Pr[B_{n+1} = j | B_n, B_{n-1}, \ldots, B_0] = \Pr[B_{n+1} | B_n]$.

We now introduce some notation. Let S denote the set of states, $A(i)$ the finite set of actions permitted in state i, and $\mu(i)$ the action chosen in state i when policy $\hat{\mu}$ is pursued, where $\bigcup_{i\in S} A(i) = A$. Further let $r(.,.,.) : S \times A \times S \to \Re$ denote the

one-step immediate reward and $p(.,.,.) : S \times A \times S \to [0,1]$ denote the associated transition probability. Then the *expected* immediate reward earned in state i when action a is chosen in it is $\bar{r}(i,a) = \sum_{j=1}^{|S|} p(i,a,j) r(i,a,j)$. Also, $p(i,a,j) = P_a(i,j)$, where $a \in \{c(\textit{continue}), m(\textit{maintain})\}$.

Definition 4. The long-run average (expected) reward of a policy $\hat{\mu}$ starting at state i is

$$\rho_{\hat{\mu}}(i) \equiv \lim_{k\to\infty} \frac{E_{\hat{\mu}}[\sum_{s=1}^{k} \bar{r}(x_s, \mu(x_s))|x_1 = i]}{k},$$

where x_s is the state occupied before the sth transition and $E_{\hat{\mu}}$ denotes the expectation induced by $\hat{\mu}$.

The above definition holds for an MDP in which the time spent in each transition is either assumed to be 1 or it is irrelevant to the model. Hence, the average reward has a unit of dollars per (state) transition. For a semi-Markov decision problem (SMDP), we need to modify the definition (see Appendix).

In general, we will use the notation $\vec{z}$ to denote a column vector whose ith element is $z(i)$. Also, $P_{\hat{\mu}}$ will denote the transition probability matrix associated with the policy $\hat{\mu}$. Let $\vec{r}_{\hat{\mu}}$ denote the column vector whose ith element is $\bar{r}(i, \mu(i))$. Then, from the definition above it follows that $\vec{\rho}_{\hat{\mu}} = (\lim_{k\to\infty}(1/k)\sum_{m=0}^{k-1} P_{\hat{\mu}}^m)\vec{r}_{\hat{\mu}}$. Let $\Pi_{\hat{\mu}}$ denote the steady-state probability of being in state i of the Markov chain of policy $\hat{\mu}$. We know that $\lim_{k\to\infty}\frac{1}{k}\sum_{m=0}^{k-1} P_{\hat{\mu}}^m$ exists for irreducible and recurrent Markov chains, and in fact, it follows that $\rho_{\hat{\mu}}(j) = \sum_{i\in S}\Pi_{\hat{\mu}}(i)\bar{r}(i,\mu(i))$ for any $j \in S$, i.e., $\rho_{\hat{\mu}}(j)$ is constant for every $j \in S$. A policy's variance in an MDP is defined in a seminal work (Filar, Kallenberg, & Lee, 1989) as follows.

Definition 5. The long-run variance of the reward of a policy $\hat{\mu}$ starting at state i in an MDP is $\sigma_{\hat{\mu}}^2(i) \equiv \lim_{k\to\infty} E_{\hat{\mu}}[\sum_{s=1}^{k}[\bar{r}(x_s,\mu(x_s)) - \rho_{\hat{\mu}}(x_s)]^2|x_1 = i]/k$.

See Appendix for the SMDP's definition. A variance-penalized MDP, in which Markov chains of all policies are positive recurrent and irreducible, then seeks to maximize $\rho_{\hat{\mu}}(i) - \theta\sigma_{\hat{\mu}}^2(i)$ over all $\hat{\mu}$ for every $i \in S$. It is not hard to show that

$$\rho = \rho_{\hat{\mu}}(i) - \theta\sigma_{\hat{\mu}}^2(i) \quad \forall i \in S. \tag{7}$$

3.1. A quadratic-programming model

Filar et al. (1989) presented a quadratic program (QP) for solving the variance-penalized MDP.

$$\text{Maximize} \sum_{i\in S}\sum_{a\in A(i)} [\bar{r}(i,a) - \theta\bar{r}^2(i,a)]x(i,a) + \theta\left[\sum_{i\in S}\sum_{a\in A(i)} \bar{r}(i,a)x(i,a)\right]^2 \quad \text{with } \theta \geqslant 0$$

such that

$$\sum_{a\in A(j)} x(j,a) - \sum_{i\in S}\sum_{a\in A(i)} p(i,a,j)x(i,a) = 0 \quad \forall j \in S, \tag{8}$$

$$\sum_{i\in S}\sum_{a\in A(i)} x(i,a) = 1 \quad \text{and } x(i,a) \geqslant 0 \quad \forall (i,a) \in (S, A(i)). \tag{9}$$

A key result in Filar et al. (1989) shows that the optimal policy, when Markov chains of all policies are positive recurrent and irreducible, is deterministic and stationary. Unfortunately, the QP given above is not easily solvable. See Filar et al. (1989, Remark 2.1, p. 152) about the openness of this problem from the solution perspective. We now discuss some computational approaches for solving the underlying variance-penalized MDP.

3.2. Computational approaches

One approach to avoid the QP is to exhaustively enumerate all the policies. However, for large-scale problems, this approach is infeasible. We present an approach based on *linearizing* the quadratic objective function by using a *surrogate* form for variance. Optimization could then be performed via linear programming. Our analysis of the resultant LP will also prove the existence of a deterministic policy for the surrogate objective function that we propose. More importantly, linearization paves the way for a computationally attractive approach based on dynamic programming (DP).

3.2.1. A linear-programming approximation

To define our surrogate form for a policy's variance, we need four functions $v(.,.,.)$, $\bar{v}(.,.)$, $w(.,.,.)$, and $\bar{w}(.,.)$. Let $v(.,.,.) : S \times A \times S \to \Re$ denote the (one-step) immediate variance, which would be defined as follows for $(i,j) \in S$ and $a \in A(i)$: $v(i,a,j) = [r(i,a,j) - \bar{r}(i,a)]^2$. Also, for any $i \in S$ and $a \in A(i)$, we define $\bar{v}(i,a) = \sum_{j=1}^{|S|} p(i,a,j)v(i,a,j)$ and for $\theta \geqslant 0$, $w(i,a,j) = r(i,a,j) - \theta v(i,a,j)$ and

$$\bar{w}(i,a) = \bar{r}(i,a) - \theta\bar{v}(i,a). \tag{10}$$

We now define one-step (or jump) variance.

Definition 6. The long-run one-step variance in the immediate rewards of a policy $\hat{\mu}$ starting at state i is

$$\kappa_{\hat{\mu}}^2(i) \equiv \lim_{k\to\infty} \frac{E_{\hat{\mu}}[\sum_{s=1}^{k} \bar{v}(x_s, \mu(x_s))|x_1 = i]}{k}. \tag{11}$$

Let $\vec{v}_{\hat{\mu}}$ denote the column vector whose ith element is $\bar{v}(i,\mu(i))$. Then, from the definition above it follows that $\vec{\kappa}_{\hat{\mu}}^2 = (\lim_{k\to\infty}(1/k)\sum_{m=0}^{k-1} P_{\hat{\mu}}^m)\vec{v}_{\hat{\mu}}$. Since $\lim_{k\to\infty}(1/k)\sum_{m=0}^{k-1} P_{\hat{\mu}}^m$ exists for irreducible and recurrent Markov chains, like in the case of average reward, it follows that $\kappa_{\hat{\mu}}^2(j) = \sum_{i\in S}\Pi_{\hat{\mu}}(i)\bar{v}(i,\mu(i))$ for any $j \in S$, and that κ^2 for a given policy is independent of the starting state. Our objective function, *also called Markowitz score*, for a policy $\hat{\mu}$ in the approximate-variance-penalized MDP, in which all policies

have irreducible and positive recurrent Markov chains, is

$$\phi_{\hat{\mu}} \equiv \rho_{\hat{\mu}} - \theta\kappa^2_{\hat{\mu}} = \sum_{i\in S} \Pi_{\hat{\mu}}\bar{w}(i,\mu(i)). \quad (12)$$

Consider the following LP: maximize

$$\sum_{i\in S}\sum_{a\in A(i)} \bar{r}(i,a)x(i,a) - \theta\sum_{i\in S}\sum_{a\in A(i)} \bar{v}(i,a)x(i,a), \quad \theta \geqslant 0$$

subject to (8) and (9). We now have the following result.

Proposition 7. (a) *Consider an MDP for which Markov chains of all policies are positive recurrent and irreducible. A vector whose* (i,a)*th element is* $x(i,a)$ *will satisfy* (8) *and* (9) *if and only if there exists a stationary policy such that* $x(i,a)$ *is equal to the limiting* (*steady-state*) *probability of being in state i and selecting action a when that stationary policy is used.*

(b) *An optimal solution to the LP corresponds to an optimal solution of the approximate-variance-penalized MDP.*

Proof. Part (a) follows directly from Theorem 8.8.2 Puterman (1994, p. 392). Now Part (a) implies that $x(i,a)$ is the limiting probability of selecting action a in state i under a stationary stochastic policy, and hence the objective function of the LP gives the objective function (12) associated with the same policy. Hence finding the optimal solution of the LP is equivalent to maximization of the objective function (12), thereby establishing Part (b). □

3.2.2. A dynamic programming approximation

In this section, we first analyze the existence of a DP solution for the problem posed above and then present a policy iteration (PI) algorithm for solving it. Finally, we analyze the convergence of the proposed algorithm.

The LP presented above shows that the approximate-variance-penalized MDP has a linear structure and suggests that DP is a possible route for solution. DP is computationally more efficient than an LP in solving MDPs (see Madani, 2000; Puterman, 1994, p. 223). For a DP-based solution, it is necessary to derive an optimality equation similar to the Bellman optimality equation for average reward, RN MDPs. The existence of such an equation is established via the following result.

Proposition 8. *If a scalar* ϕ *and an* $|S|$*-dimensional finite vector* $\vec{h}$ *satisfy for all* $i \in S$

$$\phi + h(i) \sum_{j\in S} p(i,\mu(i),j)[w(i,\mu(i),j) + h(j)], \quad (13)$$

then ϕ *is the Markowitz score associated with the policy* $\hat{\mu}$*. Furthermore if a scalar* ϕ^* *and an* $|S|$*-dimensional finite vector* $J(i)$ *satisfy for all* $i \in S$

$$\phi^* + J(i) = \max_{u\in A(i)} \left[\sum_{j\in S} p(i,u,j)[w(i,u,j) + J(j)]\right], \quad (14)$$

then ϕ^* *is the Markowitz score associated with the policy* $\hat{\mu}^*$ *that attains the max in the RHS of Eq.* (14)*. The policy* $\hat{\mu}^*$ *is the optimal policy, i.e., generates the maximum value for the Markowitz score.*

The proof will require a definition and a couple of lemmas.

Definition 9. If $\vec{h}$ denotes a vector whose ith component is denoted by $h(i)$, then we define the transformation $L_{\hat{\mu}}$ as

$$L_{\hat{\mu}}h(i) = \sum_{j\in S} p(i,\mu(i),j)[w(i,\mu(i),j) + h(j)] \quad \forall i \in S,$$

and the transformation L as

$$Lh(i) = \max_{a\in A(i)} \left[\sum_{j\in S} p(i,a,j)[w(i,a,j) + h(j)]\right] \quad \forall i \in S.$$

Lemma 10. *Given a vector* $\vec{h}$ *of dimension* $|S|$,

$$L^k_{\hat{\mu}}h(i) = E_{\hat{\mu}}\left[h(x_{k+1}) + \sum_{s=1}^{k}[\bar{w}(x_s,\mu(x_s))]|x_1 = i\right],$$

for all values of $i \in S$.

The proof is presented in the Appendix. The following proves the monotonicity of some transformations.

Lemma 11. *The transformation* $L_{\hat{\mu}}$ *is monotonic, i.e., given two vectors,* $\vec{J}$ *and* $\vec{J}'$*, which satisfy the relation* $J(i) \leqslant J'(i)$ *for every* $i \in S$ *the following is true for any positive integral* k*:* $L^k_{\hat{\mu}}J(i) \leqslant L^k_{\hat{\mu}}J'(i)$ *for every* $i \in S$.

The proof is presented in the Appendix. We can now prove Proposition 8.

Proof. Eq. (13) can be written in vector form as

$$\phi\vec{e} + \vec{h} = L_{\hat{\mu}}\vec{h}, \quad (15)$$

where $\vec{e}$ is an $|S|$-dimensional (column) vector whose every element equals 1. We will first prove that for $i = 1, 2, \ldots, |S|$,

$$L^k_{\hat{\mu}}h(i) = k\phi + h(i). \quad (16)$$

The above can be written in the vector form as

$$L^k_{\hat{\mu}}\vec{h} = k\phi\vec{e} + \vec{h}. \quad (17)$$

We will use an induction argument for the proof. From Eq. (15), the above is true when $k = 1$. Let us assume that the above is true when $k = m$. Then we have that

$$L^m_{\hat{\mu}}\vec{h} = m\phi\vec{e} + \vec{h}.$$

Using the transformation $L_{\hat{\mu}}$ on both sides of this equation, we have

$$\begin{aligned} L_{\hat{\mu}}\left(L^m_{\hat{\mu}}\vec{h}\right) &= L_{\hat{\mu}}(m\phi\vec{e} + \vec{h}) \\ &= m\phi\vec{e} + L_{\hat{\mu}}\vec{h} \\ &= m\phi\vec{e} + \phi\vec{e} + \vec{h} \quad \text{(using Eq. (15))} \\ &= (m+1)\phi\vec{e} + \vec{h}. \end{aligned}$$

Thus Eq. (17) is established using induction on k.

Using Lemma 10, we have for all i,

$$L_{\hat{\mu}}^k h(i) E_{\hat{\mu}} \left[h(x_{k+1}) + \sum_{s=1}^{k} [\bar{w}(x_s, \mu(x_s))] \middle| x_1 = i \right],$$

where $h(x_{k+1})$ is a finite quantity.

Using the above and Eq. (16), we have that

$$E_{\hat{\mu}} \left[h(x_{k+1}) + \sum_{s=1}^{k} [\bar{w}(x_s, \mu(x_s))] \middle| x_1 = i \right] = k\phi + h(i).$$

Therefore,

$$\frac{E_{\hat{\mu}}[h(x_{k+1})]}{k} + \frac{1}{k} E_{\hat{\mu}} \left[\sum_{s=1}^{k} [\bar{w}(x_s, \mu(x_s))] \middle| x_1 = i \right] = \phi + \frac{h(i)}{k}.$$

Taking limits as $k \to \infty$, since $\lim_{k\to\infty} D/k = 0$ for finite D, we have

$$\lim_{k\to\infty} \frac{1}{k} E_{\hat{\mu}} \left[\sum_{s=1}^{k} [\bar{w}(x_s, \mu(x_s))] \middle| x_1 = i \right] = \phi.$$

The definition of Markowitz score implies that the latter for the policy $\hat{\mu}$ is indeed ϕ, and the first part of the proposition is thus established. It follows directly from the above and Eq. (14) that ϕ^* is the Markowitz score associated with the policy $\hat{\mu}^*$ that attains the max in the RHS of Eq. (14).

We will now show that any policy that deviates from $\hat{\mu}^*$ will produce a Markowitz score lower than or equal to ϕ^*. This will establish that the policy $\hat{\mu}^*$ generates the maximum Markowitz score and is therefore an optimal policy. Thus, all we need to show is that a policy $\hat{\mu}$ which does not necessarily attain the max in Eq. (14) produces a Markowitz score less than or equal to ϕ^*.

Eq. (14) can be written in vector form as

$$\phi^* \vec{e} + \vec{J} = L(\vec{J}). \tag{18}$$

We will first prove that

$$L_{\hat{\mu}}^k \vec{J} \leqslant k\phi^* \vec{e} + \vec{J}. \tag{19}$$

As before, we use an induction argument. Now from Eq. (18), $L(\vec{J}) = \phi^* \vec{e} + \vec{J}$. But we know that $L_{\hat{\mu}}(\vec{J}) \leqslant L(\vec{J})$, which follows from the fact that any given policy may not attain the max in Eq. (14). Thus $L_{\hat{\mu}} \vec{J} \leqslant \phi^* \vec{e} + \vec{J}$. This proves that Eq. (19) holds when $k = 1$. Assuming that it holds when $k = m$, we have that $L_{\hat{\mu}}^m \vec{J} \leqslant m\phi^* \vec{e} + \vec{J}$.

Using the fact that $L_{\hat{\mu}}$ is monotonic from Lemma 11 it follows that

$$\begin{aligned} L_{\hat{\mu}}(L_{\hat{\mu}}^m)\vec{J} &\leqslant L_{\hat{\mu}}(m\phi^* \vec{e} + \vec{J}) \\ &= m\phi^* \vec{e} + L_{\hat{\mu}} \vec{J} \\ &\leqslant m\phi^* \vec{e} + \phi^* \vec{e} + \vec{J} \quad \text{(using Eq. (18))} \\ &= (m+1)\phi^* \vec{e} + \vec{J}. \end{aligned}$$

This establishes Eq. (19).

The following bears similarity to the proof of the first part of this proposition. Using Lemma 10 and Eq. (19), we have that for all $i \in S$,

$$E_{\hat{\mu}} \left[J(x_{k+1}) + \sum_{s=1}^{k} [\bar{w}(x_s, \mu(x_s))] \middle| x_1 = i \right] \leqslant k\phi^* + J(i).$$

From the same logic used for the first part of the proposition, dividing by k and taking the limit with $k \to \infty$, we have that

$$\lim_{k\to\infty} \frac{1}{k} E_{\hat{\mu}} \left[\sum_{s=1}^{k} [\bar{w}(x_s, \mu(x_s))] \middle| x_1 = i \right] \leqslant \phi^*.$$

In words, this means that the Markowitz score of the policy $\hat{\mu}$ is less than or equal to ϕ^*, which implies that the policy that attains the max in the RHS of Eq. (14) is indeed the optimal policy. □

Steps in the proposed PI algorithm

Step 1: Set k, the iteration number, to 0. Select an arbitrary policy. Let us denote the policy selected in the kth iteration by $\hat{\mu_k}$. Let $\hat{\mu^*}$ denote the optimal policy.

Step 2 (Policy evaluation): Solve the following linear system of equations:

$$h^k(i) = \bar{w}(i, \mu_k(i)) - \phi^k + \sum_{j=1}^{|S|} p(i, \mu_k(i), j) h^k(j).$$

The system of linear equations can be solved by setting one element of $\vec{h}^k$ to a fixed value, e.g., 0. The unknowns in this system are all the other elements of this vector and ϕ^k.

Step 3 (Policy improvement): Choose a new policy $\hat{\mu}_{k+1}$ such that

$$\mu_{k+1}(i) \in \arg\max_{a \in A(i)} \left[\bar{w}(i, a) + \sum_{j=1}^{|S|} p(i, a, j) h^k(j) \right].$$

If possible one should set $\hat{\mu}_{k+1} = \hat{\mu}_k$.

Step 4: If the new policy is identical to the old one, that is, if $\mu_{k+1}(i) = \mu_k(i)$ for each i then stop and set $\mu^*(i) = \mu_k(i)$ for every i. Otherwise, increment k by 1 and go back to the second step.

We now establish the algorithm's finite convergence.

Proposition 12. *The policy $\hat{\mu}^*$ generated by the algorithm above is an optimal policy if all policies have positive recurrent and irreducible Markov chains.*

Next, we present a lemma (see Appendix for proof) needed to prove Proposition 12.

Lemma 13. *Let $\Pi_{\hat{\mu}}(i)$ denote the limiting probability of the ith state in a Markov chain of the policy $\hat{\mu}$. If $\hat{\mu}$ has a positive recurrent and irreducible Markov chain,*

$$\sum_{i \in S} \Pi_{\hat{\mu}}(i) \left[\sum_{j \in S} p(i, \mu(i), j) h(j) - h(i) \right] = 0,$$

where $h(i)$ is a finite-valued scalar for all $i \in S$.

We present below the proof of Proposition 12.

Proof. We will need to prove that the sequence of Markowitz scores produced by the PI algorithm proposed above is an increasing one, i.e., if ϕ^k denotes the Markowitz score (determined using the policy evaluation step) in the kth iteration of the PI algorithm, then $\phi^{k+1} \geqslant \phi^k$.

If this is true, the sequence of Markowitz scores is an increasing sequence until a policy repeats. Since the number of states and actions is finite, there is a finite number of policies, and the convergence criterion $\hat{\mu}_{k+1} = \hat{\mu}_k$ must be satisfied at some finite value of k. When the policy repeats, we have that $\hat{\mu}_{k+1} = \hat{\mu}_k$ which means from the policy improvement and the policy evaluation steps that the policy $\hat{\mu}_k$ satisfies the Bellman optimality equation. From Proposition 8, this implies that $\hat{\mu}_k$ is the optimal policy.

We now prove $\phi^{k+1} \geqslant \phi^k$. From Proposition 8, we know that the term ϕ in the Bellman equation for a policy $\hat{\mu}$ equals the Markowitz score associated with the policy. Hence we can write an expression for the Markowitz score in the $(k+1)$th iteration of the algorithm as

$$\begin{aligned}
\phi^{k+1} &= \sum_{i\in S} \Pi_{\hat{\mu}_{k+1}}(i)\bar{w}(i,\mu_{k+1}(i)) \quad \text{(from Eq. (12))} \\
&= \phi^k + \sum_{i\in S} \Pi_{\hat{\mu}_{k+1}}(i)[\bar{w}(i,\mu_{k+1}(i)) - \phi^k] \\
&= \phi^k + \sum_{i\in S} \Pi_{\hat{\mu}_{k+1}}(i)[\bar{w}(i,\mu_{k+1}(i)) - \phi^k] \\
&\quad + \sum_{i\in S} \Pi_{\hat{\mu}_{k+1}}(i)\left[\sum_{j\in S} p(i,\mu_{k+1}(i),j)h^k(j) - h^k(i)\right] \\
&\quad \text{(using Lemma 13)} \\
&= \phi^k + \sum_{i\in S} \Pi_{\hat{\mu}_{k+1}}(i)\left[\bar{w}(i,\mu_{k+1}(i)) - \phi^k\right. \\
&\quad \left. + \sum_{j\in S} p(i,\mu_{k+1}(i),j)h^k(j) - h^k(i)\right]. \qquad (20)
\end{aligned}$$

The policy improvement step implies that μ_{k+1} is chosen in a way such that for each $i \in S$

$$\begin{aligned}
&\bar{r}(i,\mu_{k+1}(i)) - \theta\bar{v}(i,\mu_{k+1}(i)) + \sum_{j\in S} p(i,\mu_{k+1}(i),j)h^k(j) \\
&\quad \geqslant \bar{r}(i,\mu_k(i)) - \theta\bar{v}(i,\mu_k(i)) + \sum_{j\in S} p(i,\mu_k(i),j)h^k(j),
\end{aligned}$$

which implies that for each $i \in S$

$$\begin{aligned}
&\bar{w}(i,\mu_{k+1}(i)) + \sum_{j\in S} p(i,\mu_{k+1}(i),j)h^k(j) \\
&\quad \geqslant \bar{w}(i,\mu_k(i)) + \sum_{j\in S} p(i,\mu_k(i),j)h^k(j),
\end{aligned}$$

from which it follows that for each $i \in S$

$$\begin{aligned}
&\bar{w}(i,\mu_{k+1}(i)) + \sum_{j\in S} p(i,\mu_{k+1}(i),j)h^k(j) - \phi^k - h^k(i) \\
&\quad \geqslant \bar{w}(i,\mu_k(i)) + \sum_{j\in S} p(i,\mu_k(i),j)h^k(j) - \phi^k - h^k(i).
\end{aligned}$$

But the policy evaluation stage of the PI algorithm implies that the RHS of the above inequality equals 0. Thus for each $i \in S$,

$$0 \leqslant \bar{w}(i,\mu_{k+1}(i)) + \sum_{j\in S} p(i,\mu_{k+1}(i),j)h^k(j) - \phi^k - h^k(i).$$

From the above it follows that

$$\begin{aligned}
0 \leqslant \sum_{i\in S} \Pi_{\hat{\mu}_{k+1}} &\left[\bar{w}(i,\mu_{k+1}(i)) + \sum_{j\in S} p(i,\mu_{k+1}(i),j)h^k(j)\right. \\
&\left. - \phi^k - h^k(i)\right].
\end{aligned}$$

The above and (20) imply $\phi^{k+1} \geqslant \phi^k$. □

4. Computational experiments

The renewal-theory model: We now describe experiments with our renewal-theory model. We chose the gamma distribution to model the time to failure since it has an increasing failure rate (Lewis, 1994), making PM useful. The input parameters and output results are described in Tables 1 and 2, respectively. Our results show the (i) the optimized time for maintenance, τ (both under RS and RN conditions), (ii) the optimized objective function value of (1) and the objective function value associated with the RN strategy and (iii) the improvement obtained from pursuing RS strategies, which is defined as $Imp = (g(\tau_{RN}) - g(\tau_{RS}))/g(\tau_{RN}) \times 100$, where $g(.)$ is as defined in Eq. (1).

Experiments indicate that values exceeding 0.4 for θ produced policies that had very high costs, although their variance was also smaller. It can be noted that the RS solution invariably recommends a *shorter* PM interval suggesting that the use of a safety factor (η), discussed previously, is not practical since it is not unique. Fig. 1 plots for Case 4, $g(t)$ versus t when θ=0 (RN) and when θ is non-zero, show how the optimal point changes. The computer programs were written in MATLAB.

The MDP model: Under the action of continuing production, the one-step transition probability of going from state i to state j will be denoted by $P_c(i,j)$, and the same for the action of PM will be denoted by $P_m(i,j)$. Then, if d denotes the number of days since the last repair or maintenance, it follows that

$$P_c(d,d+1) = 1 - P_c(d,0) \quad \text{for } d = 0,1,2,\infty, \qquad (21)$$

and all other values of $P_c(i,j)$ will equal 0. The above (21) follows from the fact that the transition from d to 0 occurs because the line must have failed before the next production was completed. Also, $P_m(d,0) = 1$ for all values of d and $P_m(.,.)$ will equal 0 otherwise. This follows from the fact that when

Table 1
Input data for experiments with the risk-sensitive renewal-theory model

Case	c_r	c_m	t_r	t_m	θ	Gamma(n, λ)
1	5	2	50	12.5	0.2	(8, 0.08)
2	10	1	25	5	0.3	(10, 0.09)
3	5	2	50	12.5	0.3	(7, 0.06)
4	10	1	25	5	0.2	(5, 0.05)
5	5	2	50	12.5	0.3	(12, 0.15)
6	10	1	25	5	0.2	(10, 0.10)
7	5	2	50	12.5	0.3	(11, 0.12)
8	10	1	25	5	0.2	(6, 0.07)

Table 2
Results from experiments with the risk-sensitive renewal-theory model

Case	τ_{RS}	τ_{RN}	$g(\tau_{RS})$	$g(\tau_{RN})$	*Imp* (%)
1	58.75	77.24	0.0623	0.0673	7.43
2	49.81	61.35	0.0941	0.1046	10.04
3	65.17	90.76	0.0715	0.0792	9.72
4	26.73	36.32	0.0560	0.0632	11.39
5	144.86	172.97	0.0331	0.0352	5.96
6	45.83	55.22	0.0831	0.0901	7.77
7	54.58	69.21	0.0768	0.0837	8.98
8	32.55	43.38	0.1207	0.1346	10.33

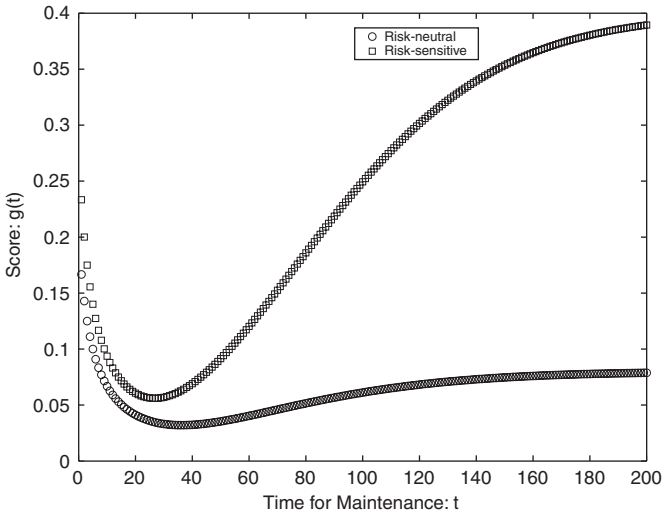

Fig. 1. A plot of the risk-sensitive and the risk-neutral objective functions for Case 4. It clearly shows how the optimal point is different in the two scenarios.

the decision of maintenance is made, the system transitions to state 0 with certainty.

By definition, the underlying Markov chains are infinite-dimensional, which can be approximated via a truncation procedure based on an augmented north-west-corner procedure matrix (Freedman, 1983; Senata, 1967). See Zhao, Braun, and Li (1999) for usage in MDPs. On the basis of the data gathered, we found that $P_c(i, 0) \approx 1$ for some $i = N^*$ and hence $|S| \approx N^* + 1$. Hence, the transition probability matrix of every chain in this problem is assumed to be truncated to its north west corner in which there are $(N^* + 1)$ rows and $(N^* + 1)$ columns. We experimented with different values for θ and the set of transition probabilities. If c_m denotes the cost of maintenance and c_r that of repair, then it follows that for all d, $r(d, \mathit{maintain}, 0) = -c_m$ and $r(d, \mathit{produce}, 0) = -c_r$. The input data and the results are described in Table 3. The set of transition probabilities in our experiments was generated using the following law: $p(d, \mathit{produce}, d+1)=\psi^d$ for $d=0, 1, \ldots, |S|-2$ and $p(|S|-1, \mathit{produce}, 0) = 1$. We used a number of different values for ψ in our experiments. Although our transition probabilities were generated in this style for experimentation, our DP model is *general*.

Table 3
Input data and results with the risk-sensitive MDP model

Case	ψ	c_m	c_r	θ	ϕ^{PI}_{RS}	ϕ^*_{RS}	OG
1	0.90	2	3	0.1	−0.7292	−0.7292	0
2	0.92	2	5	0.2	−1.0870	−1.0870	0
3	0.95	3	4	0.1	−0.8465	−0.8312	1.84
4	0.95	3	4	0.2	−1.0603	−1.032	2.74
5	0.93	2	4	0.1	−0.7987	−0.7980	0.087
6	0.89	3	6	0.2	−1.7723	−1.7723	0
7	0.88	8	10	0.1	−4.02	−3.88	3.6
8	0.96	2	4	0.2	−3.61	−3.21	12.46

For the data we obtained, $N^* = 30$ was a reasonable value.

In general, the transition probabilities for a system such as ours can be estimated as follows. Let $K(d)$ denote the number of instances in which the system transitions from the dth day to the next day without failure when the production action is chosen at the start of the dth day and $\bar{K}(d)$ denote the number of instances in which the machine fails during the dth day when the production action is chosen at the start of the dth day. The values of both K and $\bar{K}$ can be obtained from actual observations of the systems. Then, from the maximum likelihood principle, as $K(.)$ and $\bar{K}(.)$ approach infinity, $p(d, \mathit{produce}, d+1) \approx K(d)/(K(d) + \bar{K}(d))$.

An unbiased evaluation of our proposed PI algorithm would require comparing the algorithm's performance with that of an optimal algorithm. Hence for all the systems we considered, the optimal solutions were determined via an exhaustive search of the policy space using *the exact* objective function in (7)—not the surrogate. Fig. 2 plots the exact objective function and the surrogate objective function for Case 3. The graph shows that *the surrogate mimics the exact function reasonably well*. Although we do not plot all the cases, due to lack of space, this is true of all of our tests. As is clear from our empirical results, our proposed PI algorithm provides either optimal or near-optimal solutions. For the actual objective function, we remind the reader, a DP approach does not appear to be feasible, and at best one could use a QP (Filar et al., 1989).

In Table 3 ϕ^*_{RS} denotes the value of the optimized objective function in (7) and ϕ^{PI}_{RS} denotes the value of the objective function in (7) associated with the "optimal" policy of our PI algorithm. The optimality gap (in percent) can hence be defined as follows: $OG = (\phi^*_{RS} - \phi^{PI}_{RS})/\phi^*_{RS} \times 100$. In our

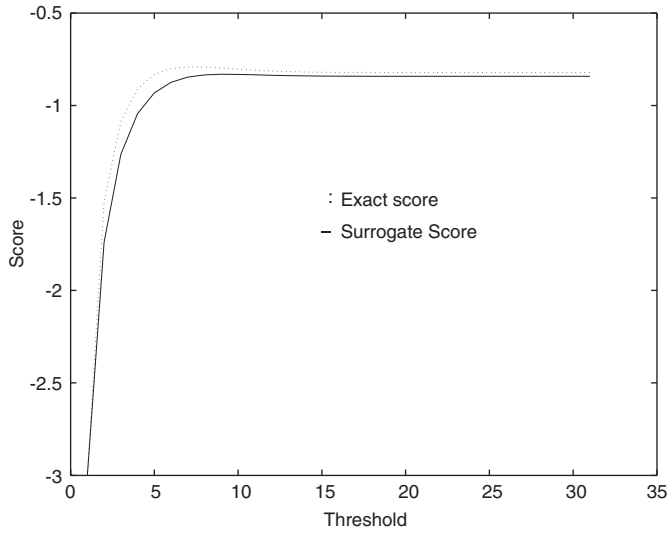

Fig. 2. A plot (for Case 3) of the exact risk-sensitive objective function (Filar et al., 1989) and the surrogate objective function.

experiments, this gap ranged from 0% to 12.46%. Our PI algorithm converged in no more than three iterations in each case tested, which takes less than 1s on a UNIX SUN-Blade Machine (C program). This assures us that the DP route is worth pursuing since a complex non-linear program has been solved approximately in a short computational time, which via other methods of non-linear programming could possibly require more time.

5. Conclusions

The literature on RS PM is limited. We developed two mathematically sound models, based on renewal theory and MDPs, for RS TPM. We need to point out that while our models were developed for optimizing a combination of mean and variance, they could be adapted easily for combining mean and the standard deviation, which have the same units. Our renewal-theory model was motivated by an industrial need for a model that *quantifies risk in tractable units*, e.g., *dollar*^{2} per hour. For the MDP model, we developed a surrogate objective function that closely mimicked the exact objective function, computationally. What is interesting is that for the surrogate, we were able to develop a computationally attractive DP approach, whose convergence we were able to show. Both models developed above are of a sufficiently general nature, and can be applied to other problems in management science. Other problem domains for our models that we will pursue in future work are supply chain management and airline revenue management.

Acknowledgments

The author thanks the anonymous reviewers, the associate editor, and the special-issue editor, Professor S. Sethi, for their useful comments.

Appendix A

Proof of Lemma 10. The proof follows from a result in Bertsekas (1995, Vol. 2, p. 6) by replacing r by w and discount factor by 1. □

Proof of Lemma 11. The proof follows from Lemma 1.1.1 of Bertsekas (1995, Vol. 2, p. 7) via replacing $\bar{r}(.,)$ by $\bar{w}(,)$. □

Proof of Lemma 13. From a result (Puterman, 1994, Theorem A.2, p. 592), one has that for all $j \in S$, $\sum_{i\in S}\Pi_{\hat{\mu}}(i)p(i,\mu(i),j) - \Pi_{\hat{\mu}}(j) = 0$. Hence $\sum_{j\in S}[\sum_{i\in S}\Pi_{\hat{\mu}}(i)p(i,\mu(i),j)h(i) - \Pi_{\hat{\mu}}(j)h(i)] = 0$, which after rearranging terms, becomes $0 = \sum_{i\in S}\Pi_{\hat{\mu}}(i)\sum_{j\in S}p(i,\mu(i),j)h(j) - \sum_i \Pi_{\hat{\mu}}(i)h(i)$. □

A.1. Definitions for SMDPs

Let $t(i,a,j)$ denote the time taken for a transition from i to j when action a is selected in i. Also, let $\bar{t}(i,a) = \sum_{j=1}^{|S|}p(i,a,j)t(i,a,j)$. Then we define three quantities $\alpha_{\hat{\mu}}(i) \equiv \lim_{k\to\infty}E_{\hat{\mu}}[\sum_{s=1}^{k}\bar{r}(x_s,\mu(x_s))|x_1 = i]/k$, $\beta_{\hat{\mu}}(i) \equiv \lim_{k\to\infty}E_{\hat{\mu}}[\sum_{s=1}^{k}\bar{t}(x_s,\mu(x_s))|x_1 = i]/k$, and $\gamma_{\hat{\mu}}(i) \equiv \lim_{k\to\infty}E_{\hat{\mu}}[\sum_{s=1}^{k}\bar{r}^2(x_s,\mu(x_s))|x_1 = i]/k$. Then for irreducible and recurrent Markov chains, from Theorem 7.5 (Ross, 1992, p. 160), the long-run average reward of a policy $\hat{\mu}$ in an SMDP (Definition 4) starting at state i is $\rho_{\hat{\mu}}(i) = \alpha_{\hat{\mu}}(i)/\beta_{\hat{\mu}}(i)$ and from Theorem 1, the corresponding long-run variance of rewards of the policy $\hat{\mu}$ (Definition 5), starting at state i, is $\sigma^2_{\hat{\mu}}(i) = \gamma_{\hat{\mu}(i)}/\beta_{\hat{\mu}}(i) - (\alpha_{\hat{\mu}}(i))^2/\beta_{\hat{\mu}}(i)$.

References

Askin, R., & Goldberg, J. (2002). *Design and analysis of lean production systems*. New York: Wiley.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control*. Belmont, MA, USA: Athena Scientific.

Chen, Y., & Jin, J. (2003). Cost-variability-sensitive preventive maintenance considering management risk. *IIE Transactions*, *35*(12), 1091–1102.

Das, T. K., & Sarkar, S. (1999). Optimal preventive maintenance in a production inventory system. *IIE Transactions*, *31*, 537–551.

Filar, J., Kallenberg, L., & Lee, H. (1989). Variance-penalized Markov decision processes. *Mathematics of Operations Research*, *14*(1), 147–161.

Fishburn, P. (1977). Mean-risk analysis with risk associated with below-target returns. *The American Economic Review*, *67*(2), 116–126.

Freedman, D. (1983). *Approximating countable Markov Chains*. 2nd ed., New York, NY: Springer.

Gan, X., Sethi, S., & Yan, H. (2005). Channel coordination with a risk-neutral supplier and downside-risk-averse retailer. *Production and Operations Management*, *14*(1), 80–89.

Kao, E. (1997). *An introduction to stochastic processes*. Belmont, CA: Duxbury Press.

Lewis, E. E. (1994). *Introduction to reliability engineering*. New York: Wiley.

Madani, O. (2000). *Complexity results for infinite-horizon Markov decision processes*. Ph.D. thesis, University of Washington, Department of Computer Science, U.S.A.

Markowitz, H. (1952). Porfolio selection. *Journal of Finance*, *7*(1), 77–91.

Puterman, M. L. (1994). *Markov decision processes*. New York: Wiley Interscience.

Ross, S. M. (1992). *Applied probability models with optimization applications*. New York: Dover.

Ross, S. M. (1997). *Introduction to probability models*. CA: Academi Press.

Roy, A. (1952). Safety first and the holding of assets. *Econometrica*, *20*(3), 431–449.

Ruefli, T., Collins, J., & Lacugna, J. (1999). Risk measures in strategic management research: Auld Lang Syne. *Strategic Management Journal*, *20*, 167–194.

Senata, E. (1967). Finite approximation to infinite non-negative matrices. *Proceedings of the Cambridge Philosophical Society*, *63*, 983–992.

Suri, R. (1998). *Quick response manufacturing*. Portland, OR: Productivity Press.

Tesler, L. (1955). Safety-first and hedging. *Review of Economic Studies*, *23*, 1–16.

Von Neumann, J., & Morgenstern, O. (1953). *The theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Zhao, Y., Braun, W., & Li, W. (1999). Northwest corner and banded matrix approximations to a Markov chain. *Naval Research Logistics*, *46*, 187–197.

Abhijit Gosavi obtained a Ph.D. (Industrial Engineering) in 1999, an M.S. (Mechanical Engineering) in 1995, and a B.S. (Mechanical Engineering) in 1992. Since 2003, he has been working as an assistant professor in the department of industrial engineering in the University at Buffalo, SUNY. His research interests are primarily in control theory, simulation-based optimization, revenue management, and quality systems. His papers have appeared in leading journals, such as *Management Science*, *Machine Learning*, and *Systems and Control Letters*.