A Heuristic Policy for Maintaining Multiple Multi-State Systems

Mimi Zhang

School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

ABSTRACT

This work is concerned with the optimal allocation of limited maintenance resources among a collection of competing multi-state systems, and the dynamic of each multi-state system is modelled by a Markov chain. Determining the optimal dynamic maintenance policy is prohibitively difficult, and hence we propose a heuristic dynamic maintenance policy in which maintenance resources are allocated to systems with higher importance. The importance measure is well justified by the idea of subsidy, yet the computation is expensive. Hence, we further propose two modifications of the importance measure, resulting in two modified heuristic policies. The performance of the two modified heuristics is evaluated in a systematic computational study, showing exceptional competence.

KEY WORDS: approximate linear programming; expected discounted reward; partially observable Markov decision process;

1 Introduction

A partially observable Markov decision process (POMDP) is a generalization of a Markov decision process. A POMDP models a decision process in which it is assumed that the system's dynamic is determined by a Markov decision process, but the decision maker cannot directly observe the system's state. For a finite-state Markov decision process, the optimal policy can be expressed in a simple tabular form. When state uncertainty is introduced, the optimal policy for a POMDP is defined over a continuum of states. It is established in Madani et al. (1999) that

optimal planning without full observability is prohibitively difficult both in theory and practice, and many natural questions in this domain are undecidable. Consequently, approximate methods are required even for small-size problems. Existing efficient approximate methods are policy iteration (Hansen, 1998), point-based value iteration (Pineau et al., 2003), and approximate linear programming (Hauskrecht and Kveton, 2004). The current work investigates an even more difficult problem: optimally maintaining a collection of multi-state systems with limited maintenance resources, where the dynamic of each multi-state system is modelled by a Markov chain. That is, instead of one POMDP, the problem involves multiple independent POMDPs, and the state of a POMDP affects the action taken on another POMDP. Determining the optimal dynamic maintenance policy for multiple competing POMDPs is apparently impractical, and hence we develop a heuristic policy: at each decision epoch, we measure the importance of each system, and only systems with larger importance measures will receive their optimal actions.

Importance measures have been widely used as important decision-aiding indicators in various domains. For example, in risk analyses, importance measures are used in risk-informed decision-making (Tyrväinen, 2013); in reliability engineering, importance measures are used to prioritize components in a system for reliability improvement (Borgonovo et al., 2016). Recently, importance measures have been applied for maintenance optimization. Liu et al. (2014) proposed a maintenance strategy in which the component yielding the largest expected net revenue is selected for maintenance whenever the system reliability is below a threshold. To reduce system downtime, Wu et al. (2016) proposed a maintenance strategy that, when a component in a system is failed and under repair, a number of the other components are selected for preventive maintenance; the authors developed an importance measure for the selection of components for preventive maintenance. Dui et al. (2017) pointed out that the preventive maintenance time of a selected component may be longer than the maintenance time of the failed component, and that with the same reliability improvement on the system, different components may result in different preventive maintenance costs; the authors developed an importance measure taking into account the time and cost of preventive maintenance. With the objective of maximizing the throughput of a production system over a time interval, Ahmed and Liu (2019) developed two types of importance measures for prioritizing the critical components in the maintenance schedule. In the framework of condition-based maintenance, Do and Bérenguer (2020) developed an importance measure based on the conditional reliability of the system; that is, components are ranked according to their ability to improve the system's conditional reliability over a time interval.

Existing works on importance-measure based maintenance are all focused on ranking the com-

ponents. By contrast, this work is devoted to ranking systems. Within the POMDP framework, a multi-state system is treated as important (having a large importance measure) if the cost for not optimally maintaining the system is high. The importance measure defined in this work has the economic interpretation as a subsidy (for a positive importance measure) or a tax (for a negative importance measure); see Whittle (1988). Our sequential resource allocation and stochastic scheduling framework is very general, and can be applied to solve, e.g., the dynamic multichannel access problem (Liu and Zhao, 2010), multi-UAV dynamic routing (Ny et al., 2008), sequential selection of online ads (Yuan and Wang, 2012), etc.

In the upcoming sections, we will cover the following. In Section 2, we formulate the problem, define the importance measure, and point out its drawbacks. In Section 3, we introduce the two modified importance measures. We prove that the two measures are well defined and further give two interpretations of the second measure. In Section 4, the performance of the proposed heuristics is studied in computational experiments. Section 5 concludes.

2 Problem Formulation

POMDPs provide a rich framework for planning under both state transition uncertainty and observation uncertainty. A standard discrete-time POMDP can be defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, p^a_{ss'}, f^a_s(z), R^a_s, \theta)$:

- $\mathcal{S}$ is a finite set of states;
- $\mathcal{A}$ is a finite set of actions;
- $\mathcal{Z}$ is an observation space;
- $p^a_{ss'}$ is the probability of transitioning to state s' after taking action a, given that the current state is s ($s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$);
- $f^a_s(z)$ is the probability for observing z after taking action a, given that the current state is s ($z \in \mathcal{Z}$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$);
- R^a_s is the finite immediate reward by taking action a for state s ($s \in \mathcal{S}$ and $a \in \mathcal{A}$);
- $\theta \in (0,1)$ is a discount factor.

For an action a that cannot return any observation, it is equivalent to saying that action a always returns the same observation, denoted by "null", and $f^a_s(z = \text{null}) = 1$ for any state s. Ellis et al.

(1995) provided an application example of the POMDP to a one-lane, two-girder highway bridge. The condition of the bridge is characterized by five states, i.e., $\mathcal{S} = \{1, 2, 3, 4, 5\}$. The available actions are $\mathcal{A}$=$\{$doing nothing, visual inspection, nondestructive ultrasonic evaluation, cleaning and repainting corroded surfaces, repainting and strengthening deteriorated girders, extensive structural repair$\}$. An visual inspection yields one of three possible outcomes: good, fair, and poor. The ultrasonic technique is to measure web and flange thickness loss in girders, and the indicated results $\{$state 1, state 2, state 3, state 4, state 5$\}$ are error corrupted. Therefore, the observation space $\mathcal{Z}$ is a discrete set of eight observations. If, for example, the underlying state is $s = 1$ and the action taken is a =visual inspection, then $f_s^a(z = \text{good}) = 0.2$ and $f_s^a(z = \text{fair}) = 0.8$; if the underlying state is $s = 2$ and the action taken is a =nondestructive ultrasonic evaluation, then $f_s^a(z = \text{state 1}) = 0.05$, $f_s^a(z = \text{state 2}) = 0.9$, and $f_s^a(z = \text{state 3}) = 0.05$. State transitions satisfy the Markov property; for example, given $s = 1$ at time t and a =doing nothing, the probability $p_{ss'}^a$ for $s' = 2$ at time $t+1$ is 0.13, independent of all states and actions before time t.

Within the POMDP framework, the information on the system's true state is incomplete and encapsulated by a probability vector, called the *belief state*. A belief state at epoch t ($t = 0, 1, 2, \ldots$) is a (column) vector of probabilities: $\boldsymbol{b}^t = (b_s^t : s \in \mathcal{S})$, where b_s^t is the probability of the system being in state s at epoch t. We have $b_s^t \geq 0$ and $\sum_{s \in \mathcal{S}} b_s^t = 1$, and therefore the belief state space is a unit simplex, denoted by Δ. It is well-known that $\boldsymbol{b}^t$ summarizes all the information necessary for making decisions at epoch t (Sondik, 1978); that is, to make a decision at epoch t, we only need to know the belief state $\boldsymbol{b}^t$, instead of all the historical actions and observations.

The Markovian decision-making process is as follows. At time 0, the decision maker's belief state $\boldsymbol{b}_0$ characterizes the prior knowledge regarding the condition of the system before the beginning of the sequential decision making. At time point t ($t = 1, 2, \ldots$), the decision maker collects an observation z_t. According to the information at time $t-1$ (i.e., $\boldsymbol{b}^{t-1}$ and a_{t-1}) and the new information (i.e., z_t), the decision maker updates his belief regarding the system's current state s_t. According to the newly updated belief state $\boldsymbol{b}^t$, the decision maker then determines the action a_t. Likewise, at epoch $t+1$, the decision maker collects a new observation z_{t+1}, then updates the belief state $\boldsymbol{b}^{t+1}$ from ($\boldsymbol{b}^t$, a_t, z_{t+1}), and finally determines the action a_{t+1}.

The rule for determining the action a_t for the belief state $\boldsymbol{b}^t$ is called a *policy*. More formally, a policy π is a mapping from the belief state space to the action set ($\pi : \Delta \to \mathcal{A}$), and the optimal

policy π^* maximizes the value function (the expected discounted reward) for any given belief state:

$$\begin{aligned} V_{\pi^*}(\boldsymbol{b}^t) &= \mathrm{E}\left[R_{s_t}^{a_t} + \theta R_{s_{t+1}}^{a_{t+1}} + \theta^2 R_{s_{t+2}}^{a_{t+2}} + \cdots | \boldsymbol{b}^t, \pi^*\right] \\ &= \max_{a\in\mathcal{A}}\{\sum_{s\in\mathcal{S}} R_s^a b_s^t + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z|\boldsymbol{b}^t, a_t = a) V_{\pi^*}(\boldsymbol{b}^{t+1}) dz\}, \end{aligned} \tag{1}$$

where $\boldsymbol{b}^{t+1}$ is calculated from ($\boldsymbol{b}^t$, a_t, z_{t+1}) using Bayes' rule:

$$\begin{aligned} \Pr(s_{t+1} = s'|\boldsymbol{b}^t, a_t = a, z_{t+1} = z) &= \frac{\Pr(z_{t+1} = z|\boldsymbol{b}^t, a_t = a, s_{t+1} = s')\Pr(s_{t+1} = s'|\boldsymbol{b}^t, a_t = a)}{\Pr(z_{t+1} = z|\boldsymbol{b}^t, a_t = a)} \\ &= \frac{f_{s'}^a(z)\sum_{s\in\mathcal{S}} b_s^t p_{ss'}^a}{\sum_{s'\in\mathcal{S}} f_{s'}^a(z)\sum_{s\in\mathcal{S}} b_s^t p_{ss'}^a}. \end{aligned} \tag{2}$$

In the following, we write $\boldsymbol{b}^{t+1}$ and $\ell(\boldsymbol{b}, a, z)$ interchangeably to indicate that $\boldsymbol{b}^{t+1}$ is updated from $\boldsymbol{b}^t = \boldsymbol{b}$, $a_t = a$ and $z_{t+1} = z$. The optimal policy π^* is deterministic, stationary and Markovian (Blackwell, 1965). The optimum policy is defined over a continuum of states, yet does not have an analytic expression. Hence, different methods have been developed for approximating the optimal policy; see Hauskrecht (2000), de Farias and Roy (2003) and Shani et al. (2013).

The current work is focused on the problem of optimally allocating limited effort (such as time, spares, maintenance personnel, etc.) among a collection of competing projects, and the dynamic of each project is modelled by an independent Markov chain. For example, a collection of multi-state systems competing for a limited number of spare parts. For illustrative purpose, we here consider the problem of maintaining a collection of M (> 1) multi-state systems with only $\kappa (< M)$ repairmen. Consequently, at each decision epoch, if there are more than κ systems whose optimal actions are not "doing nothing", we need to decide which κ systems will receive their optimal actions – the remaining $M - \kappa$ systems will all receive the do-nothing action. The optimal planning for a collection of competing POMDPs is prohibitively difficult due to the inherent complexity of the POMDP model. In fact, Papadimitriou and Tsitsiklis (1999) proved that such problems are PSPACE-hard. This motivates us to develop a heuristic policy: at each decision epoch, we measure the importance of each system, and only κ systems with larger importance measures will receive their optimal actions. Hereafter, we label the do-nothing action by the number 0; that is, $a_t = 0$ means that the action taken at time t is "doing nothing".

The importance measure defined in this work is inspired by the idea of subsidy for "doing nothing". We explain the idea through one POMDP/multi-state system. Assume that the decision maker will be given a subsidy whenever the action taken on the system is "doing nothing". For example, if the optimal action for the belief state $\boldsymbol{b}^t$ is "replacing a component". If the decision

maker instead takes the do-nothing action, he will be given a positive subsidy to offset the loss caused by not taking the optimal action for the belief state $\boldsymbol{b}^t$. Apparently, the decision maker is willing to trade "replacing a component" for "doing nothing" only when the subsidy is large enough to cover the loss. In other words, the minimal subsidy required by the decision maker reflects the importance of the optimal action for the belief state $\boldsymbol{b}^t$, and hence can be adopted as the importance measure of the system at time t.

We now formally define the importance measure. After including the subsidy w for the do-nothing action, let $V(\boldsymbol{b}^t; w)$ denote the new maximal expected discounted reward (EDR) for belief state $\boldsymbol{b}^t$:

$$\begin{aligned} V(\boldsymbol{b}^t; w) &= \max_{a_t \in \mathcal{A}} \{ w\delta(a_t = 0) + \sum_{s \in \mathcal{S}} R_s^{a_t} b_s^t + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}^t, a_t) V(\boldsymbol{b}^{t+1}; w) dz \} \\ &= \max_{a_t \in \mathcal{A}} \{ \sum_{s \in \mathcal{S}} [R_s^{a_t} + w\delta(a_t = 0)] b_s^t + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}^t, a_t) V(\boldsymbol{b}^{t+1}; w) dz \}, \end{aligned} \tag{3}$$

where $\delta(\cdot)$ is the indicator function. Equation (3) implies that the subsidy can be incorporated into the reward structure, and the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, p_{ss'}^a, f_s^a(z), R_s^a + w\delta(a = 0), \theta)$ is still a POMDP with a deterministic and stationary optimal policy. The optimal action for belief state $\boldsymbol{b}^t$ is

$$a(\boldsymbol{b}^t; w) = \arg\max_{a_t \in \mathcal{A}} \{ w\delta(a_t = 0) + \sum_{s \in \mathcal{S}} R_s^{a_t} b_s^t + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}^t, a_t) V(\boldsymbol{b}^{t+1}; w) dz \}. \tag{4}$$

We call the set of belief states $\mathcal{P}(w) = \{\boldsymbol{b} \in \Delta : a(\boldsymbol{b}; w) = 0\}$ as the inactive set. In other words, under subsidy w, if the belief state $\boldsymbol{b}^t \in \mathcal{P}(w)$, then the optimal action $a(\boldsymbol{b}^t; w)$ is "doing nothing". Intuitively, if the optimal action for a belief state $\boldsymbol{b}$ is "doing nothing" when the subsidy is w, then the optimal action for $\boldsymbol{b}$ will be "doing nothing" for any subsidy larger than w. Hence, we would expect that, if the action $a(\boldsymbol{b}; w_1)$ is "doing nothing", then $a(\boldsymbol{b}; w_2)$ is always "doing nothing" for $w_2 > w_1$; or, equivalently, if $\boldsymbol{b} \in \mathcal{P}(w_1)$ and $w_2 > w_1$, then $\boldsymbol{b} \in \mathcal{P}(w_2)$. Unfortunately, this is not always the case (Whittle, 1988): for an arbitrary POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, p_{ss'}^a, f_s^a(z), R_s^a, \theta)$, there may exist a subsidy $w_2 (> w_1)$ such that $\boldsymbol{b} \in \mathcal{P}(w_1)$ yet $\boldsymbol{b} \notin \mathcal{P}(w_2)$. In other words, the subsidy as an importance measure is not well defined for all POMDPs. The POMDPs whose inactive sets can only increase with the subsidy are called indexable:

Definition 1. *A POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, p_{ss'}^a, f_s^a(z), R_s^a, \theta)$ is called indexable if the inactive set $\mathcal{P}(w)$ increases from the empty set $\varnothing$ to the whole belief state space Δ as the subsidy w increases from $-\infty$ to $+\infty$.*

Definition 2. *If a POMDP (i.e., a multi-state system) is indexable, and its belief state at time t is $\boldsymbol{b}^t$, then its importance measure at time t, denoted by $I(\boldsymbol{b}^t)$, is the infimum subsidy w such that $a(\boldsymbol{b}^t; w) = 0$.*

Given that indexability does not always hold, we have to trade indexability for specific structural conditions. In Appendix A, we study a particular POMDP (with only two actions) for which the indexability always holds.

After defining the importance measure, we now come back to the problem of optimally allocating limited effort among M multi-state systems. Note that the M multi-state systems need not be identical; each multi-state system can be modelled by a different Markov chain. Suppose that all the M multi-state systems are indexable. At each decision epoch, if the number of positive importance measures is larger than κ, then only κ multi-state systems with larger importance measures will receive their optimal actions. If the number of positive importance measures is smaller than κ, then only multi-state systems with positive importance measures will receive their optimal actions.

Although the importance measure defined above is well justified by the notion of subsidy, it has two drawbacks: (1) The importance measure is only defined for indexable POMDPs. (2) The importance measure is computationally expensive; according to Equation (3), we have to try many candidate subsidy values for a belief state, and each trial calls the running of value iteration until convergence. Therefore, we below introduce two modified importance measures, both of which are defined for every POMDP and are computationally cheap.

3 Two Modified Importance Measures

3.1 Approximate Measure

The computational burden of the importance measure is mainly introduced by the difficulty in evaluating the value function $V_{\pi^*}(\boldsymbol{b})$. We hence propose to approximate the value function to the second order. Then the infimum subsidy calculated from the approximate value function will serve as an importance measure, called the approximate measure.

Recall that, given a policy π, the EDR for the POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, p^a_{ss'}, f^a_s(z), R^a_s, \theta)$ is

$$V_\pi(\boldsymbol{b}) = \mathrm{E}\left[R^{a_t}_{s_t} + \theta R^{a_{t+1}}_{s_{t+1}} + \theta^2 R^{a_{t+2}}_{s_{t+2}} + \cdots | \boldsymbol{b}^t = \boldsymbol{b}, \pi\right]. \tag{5}$$

The well-known myopic policy approximates the EDR $V_\pi(\boldsymbol{b})$ by $\langle \vec{R}^{\pi(\boldsymbol{b})}, \boldsymbol{b}\rangle$, where $\langle\cdot,\cdot\rangle$ is the inner product, and $\vec{R}^a = (R^a_s : s \in \mathcal{S})$ is a vector of rewards. We here propose a second-order approxima-

tion:

$$V_\pi(\boldsymbol{b}) \approx \mathrm{E}\left[R_{s_t}^{a_t} + \theta R_{s_{t+1}}^{a_{t+1}} | \boldsymbol{b}^t = \boldsymbol{b}, \pi\right] = \langle \vec{R}^{\pi(\boldsymbol{b})}, \boldsymbol{b} \rangle + \theta \mathrm{E}\left[R_{s_{t+1}}^{a_{t+1}} | \boldsymbol{b}^t = \boldsymbol{b}, \pi\right]. \tag{6}$$

Then the optimal value function $V_{\pi^*}(\cdot)$ is approximated by $V_2(\cdot)$:

$$V_2(\boldsymbol{b}) = \max_{a \in \mathcal{A}} \{ \langle \vec{R}^a, \boldsymbol{b} \rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}, a_t = a) \max_{a_{t+1} \in \mathcal{A}} \langle \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b}, a, z) \rangle dz \}. \tag{7}$$

For the POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, p_{ss'}^a, f_s^a(z), R_s^a + w\delta(a=0), \theta)$, the corresponding optimal value function is approximated by

$$V_2(\boldsymbol{b}; w) = \max_{a \in \mathcal{A}} \{ w\delta(a=0) + \langle \vec{R}^a, \boldsymbol{b} \rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}, a_t = a) \max_{a_{t+1} \in \mathcal{A}} \langle w\delta(a_{t+1}=0) + \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b}, a, z) \rangle dz \}. \tag{8}$$

The optimal action determined by the second-order approximation is

$$a_2(\boldsymbol{b}; w) = \arg\max_{a \in \mathcal{A}} \{ w\delta(a=0) + \langle \vec{R}^a, \boldsymbol{b} \rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}, a_t = a) \max_{a_{t+1} \in \mathcal{A}} \langle w\delta(a_{t+1}=0) + \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b}, a, z) \rangle dz \}. \tag{9}$$

Correspondingly, we can define an inactive set $\mathcal{P}_2(w) = \{\boldsymbol{b} \in \Delta : a_2(\boldsymbol{b}; w) = 0\}$. The approximate measure for belief state $\boldsymbol{b}$ is defined as the infimum subsidy w such that $a_2(\boldsymbol{b}; w) = 0$. The following proposition states that the approximate measure is well defined for every POMDP.

Proposition 1. *For any POMDP, the inactive set $\mathcal{P}_2(w)$ increases from the empty set $\varnothing$ to the whole belief state space Δ as the subsidy w increases from $-\infty$ to $+\infty$.*

Proof. The proof is given in Appendix B. □

Then the heuristic policy for the competing M multi-state systems operates as follows. At each decision epoch, if the number of positive approximate measures is larger than κ, then only κ systems with larger approximate measures will receive their optimal actions. If the number of positive approximate measures is smaller than κ, then only systems with positive approximate measures will receive their optimal actions. Although the values of the approximate measures are different from the values of the importance measures, it is the ordering of the importance/approximate measures that determines the policy. We expect that the ordering of the importance measures is most of the time preserved under our approximate approach.

We can further approximate the optimal value function $V_{\pi^*}(\cdot)$ to the third order:

$$V_3(\boldsymbol{b}) = \max_{a \in \mathcal{A}} \{ \langle \vec{R}^a, \boldsymbol{b} \rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}, a_t = a) V_2(\ell(\boldsymbol{b}, a, z)) dz \}, \tag{10}$$

and define an importance measure from the third-order approximation in a similar manner, which we might call the third-order measure. One may argue that the heuristic policy under the third-order measure is superior to the approximate-measure policy, as the third-order approximation $V_3(\boldsymbol{b})$ is closer to $V_{\pi^*}(\boldsymbol{b})$. However, as with the importance measure, the third-order measure is not well defined for every POMDP. The computational complexity of the approximate measure is much lower than that of the third-order measure. Moreover, the numerical study in Section 4 will reveal that the approximate-measure policy outperforms the third-order measure policy.

To calculate the approximate measure, we need to numerically try different values of w. For a large enough subsidy $\hat{w}$ such that $0 = \arg\max_{a\in\mathcal{A}}\langle \vec{R}^a + \hat{w}\delta(a=0), \boldsymbol{b}\rangle$ for any $\boldsymbol{b}$, the optimal action at any decision epoch is always $a = 0$. Hence, we only need to search in the interval $(0, \hat{w})$ the minimal subsidy value such that the optimal action for $\boldsymbol{b}$ is $a = 0$. If the observation space $\mathcal{Z}$ is discrete, the approximate measure can be quickly determined. Otherwise, if the observation space is continuous, we can apply numerical integration on the grid of points $\{z_1, z_2, z_3, \ldots\}$ over the observation space $\mathcal{Z}$. Specifically, under subsidy w, the second-order approximation reads:

$$V_2(\boldsymbol{b}; w) \approx \max_{a\in\mathcal{A}}\{w\delta(a=0) + \langle \vec{R}^a, \boldsymbol{b}\rangle + \theta \sum_{z_i} \Pr(z_{t+1} = z_i | \boldsymbol{b}, a_t = a) \max_{a_{t+1}\in\mathcal{A}} \langle w\delta(a_{t+1}=0) + \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b}, a, z_i)\rangle dz_i\}. \tag{11}$$

3.2 Rate Measure

The rate measure for belief state $\boldsymbol{b}$, denoted by $\mathcal{I}(\boldsymbol{b})$, is the minimal subsidy w such that

$$0 = \arg\max_{a\in\mathcal{A}}\{w\delta(a=0) + \langle \vec{R}^a, \boldsymbol{b}\rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}, a_t = a) V_{\pi^*}(\ell(\boldsymbol{b}, a, z)) dz\}. \tag{12}$$

$\mathcal{I}(\boldsymbol{b})$ can be interpreted as a one-off subsidy as follows. Recall that the optimal action for $\boldsymbol{b}$ should be

$$\arg\max_{a\in\mathcal{A}}\{\langle \vec{R}^a, \boldsymbol{b}\rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}, a_t = a) V_{\pi^*}(\ell(\boldsymbol{b}, a, z)) dz\}. \tag{13}$$

However, due to competing multi-state systems, we have to take action $a = 0$. We assume that this is a one-time restriction, and we can still act optimally afterwards according to the optimal policy π^*. Under this assumption, the loss for taking action $a = 0$ (at time t only) is

$$V_{\pi^*}(\boldsymbol{b}) - [\langle \vec{R}^0, \boldsymbol{b}\rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}, a_t = 0) V_{\pi^*}(\ell(\boldsymbol{b}, 0, z)) dz]. \tag{14}$$

If we subsidize action $a = 0$ by the amount $\mathcal{I}(\boldsymbol{b})$, then the optimal action for state $\boldsymbol{b}$ will be $a = 0$. Therefore, we have

$$\mathcal{I}(\boldsymbol{b}) = V_{\pi^*}(\boldsymbol{b}) - [\langle \vec{R}^0, \boldsymbol{b}\rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z | \boldsymbol{b}, a_t = 0) V_{\pi^*}(\ell(\boldsymbol{b}, 0, z)) dz]. \tag{15}$$

We can utilize the above equation to calculate the rate measure, which requires very little effort.

A POMDP under the rate measure is apparently indexable: if

$$0 = \arg\max_{a\in\mathcal{A}}\{\mathcal{I}(\boldsymbol{b})\delta(a=0) + \langle\vec{R}^a,\boldsymbol{b}\rangle + \theta\int_{\mathcal{Z}}\Pr(z_{t+1}=z|\boldsymbol{b},a_t=a)V_{\pi^*}(\ell(\boldsymbol{b},a,z))dz\}, \tag{16}$$

then $0 = \arg\max_{a\in\mathcal{A}}\{w\delta(a=0) + \langle\vec{R}^a,\boldsymbol{b}\rangle + \theta\int_{\mathcal{Z}}\Pr(z_{t+1}=z|\boldsymbol{b},a_t=a)V_{\pi^*}(\ell(\boldsymbol{b},a,z))dz\}$ for any $w > \mathcal{I}(\boldsymbol{b})$; that is, with the subsidy increasing, the inactive set cannot decrease.

We here give another interpretation of $\mathcal{I}(\boldsymbol{b})$ utilizing the approximate linear programming technique (de Farias and Roy, 2003; Hauskrecht and Kveton, 2004). Consider the problem

$$\begin{aligned}(\text{P1}) \qquad \min_{V(\cdot)} \quad & \int_{\Delta} c(\boldsymbol{b})V(\boldsymbol{b})d\boldsymbol{b} \\ \text{s.t.} \quad & V(\boldsymbol{b}) \geq \langle\vec{R}^a,\boldsymbol{b}\rangle + \theta\int_{\mathcal{Z}}\Pr(z|\boldsymbol{b},a)V(\ell(\boldsymbol{b},a,z))dz, \quad \forall\,\boldsymbol{b}\in\Delta, a\in\mathcal{A}.\end{aligned}$$

Here, $c(\cdot)$ is an arbitrary positively valued function. It is clear that, for any positive function $c(\cdot)$, $V_{\pi^*}(\cdot)$ is the unique solution to problem (P1). The approximate linear programming method approximates the value function $V(\cdot)$ by a set of basis functions, in order to transform the problem into linear. With an aim of computing a coefficient vector $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_k)$ such that $V_{\pi^*}(\cdot)$ can be approximated closely by the given basis functions $\boldsymbol{\upsilon}(\boldsymbol{b}) = (\upsilon_1(\boldsymbol{b}),\ldots,\upsilon_k(\boldsymbol{b}))$: $V_{\pi^*}(\boldsymbol{b}) \approx \langle\boldsymbol{\beta},\boldsymbol{\upsilon}(\boldsymbol{b})\rangle$, we pose the following optimization problem

$$\begin{aligned}(\text{P2}) \qquad \min_{\boldsymbol{\beta}} \quad & \langle\boldsymbol{\beta}, \int_{\Delta} c(\boldsymbol{b})\boldsymbol{\upsilon}(\boldsymbol{b})d\boldsymbol{b}\rangle \\ \text{s.t.} \quad & \langle\boldsymbol{\beta},\boldsymbol{\upsilon}(\boldsymbol{b}) - \theta\int_{\mathcal{Z}}\Pr(z|\boldsymbol{b},a)\boldsymbol{\upsilon}(\ell(\boldsymbol{b},a,z))dz\rangle \geq \langle\vec{R}^a,\boldsymbol{b}\rangle, \quad \forall\,\boldsymbol{b}\in\mathcal{B}, a\in\mathcal{A},\end{aligned}$$

where we approximate the belief state space by a finite set, $\mathcal{B}$, of randomly sampled belief states. Define for shorthand $\tilde{\boldsymbol{\upsilon}} = \int_{\Delta} c(\boldsymbol{b})\boldsymbol{\upsilon}(\boldsymbol{b})d\boldsymbol{b}$ and $\bar{\boldsymbol{\upsilon}}(\boldsymbol{b},a) = \int_{\mathcal{Z}}\Pr(z|\boldsymbol{b},a)\boldsymbol{\upsilon}(\ell(\boldsymbol{b},a,z))dz$. Then the Lagrange dual function is

$$L(\boldsymbol{\beta},\lambda_{\boldsymbol{b},a}) = \langle\boldsymbol{\beta},\tilde{\boldsymbol{\upsilon}}\rangle - \sum_{a\in\mathcal{A}}\sum_{\boldsymbol{b}\in\mathcal{B}}\lambda_{\boldsymbol{b},a}[\langle\boldsymbol{\beta},\boldsymbol{\upsilon}(\boldsymbol{b})\rangle - \langle\vec{R}^a,\boldsymbol{b}\rangle - \theta\langle\boldsymbol{\beta},\bar{\boldsymbol{\upsilon}}(\boldsymbol{b},a)\rangle]. \tag{17}$$

The corresponding Lagrange dual problem is

$$\begin{aligned}(\text{P3}) \qquad \max_{\{\lambda_{\boldsymbol{b},a}\}} \quad & \sum_{a\in\mathcal{A}}\sum_{\boldsymbol{b}\in\mathcal{B}}\lambda_{\boldsymbol{b},a}\langle\vec{R}^a,\boldsymbol{b}\rangle \\ \text{s.t.} \quad & \tilde{\boldsymbol{\upsilon}} - \sum_{a\in\mathcal{A}}\sum_{\boldsymbol{b}\in\mathcal{B}}\lambda_{\boldsymbol{b},a}[\boldsymbol{\upsilon}(\boldsymbol{b}) - \theta\bar{\boldsymbol{\upsilon}}(\boldsymbol{b},a)] = 0; \\ & \lambda_{\boldsymbol{b},a} \geq 0, \quad \forall\,\boldsymbol{b}\in\mathcal{B}, a\in\mathcal{A}.\end{aligned}$$

Let $\boldsymbol{\beta}^*$ and $\{\lambda^*_{\boldsymbol{b},a} : a \in \mathcal{A}, \boldsymbol{b} \in \mathcal{B}\}$ denote the optimal primal and dual solutions. We note the following.

- The objective function of the dual problem (P3) indicates that $\lambda^*_{\boldsymbol{b},a}$ can be interpreted as the expected discounted time that action a is taken for belief state $\boldsymbol{b}$ under the optimal policy. By complementary slackness, we have $\lambda^*_{\boldsymbol{b},a} = 0$ for any non-optimal action a: $\langle \boldsymbol{\beta}^*, \boldsymbol{\upsilon}(\boldsymbol{b}) \rangle > \langle \vec{R}^a, \boldsymbol{b} \rangle + \theta \langle \boldsymbol{\beta}^*, \bar{\boldsymbol{\upsilon}}(\boldsymbol{b}, a) \rangle$. In other words, the optimal action for a belief point $\boldsymbol{b}$ is simply $\{a \in \mathcal{A} : \lambda^*_{\boldsymbol{b},a} > 0\}$.

- The Lagrange dual function (17) indicates that $\langle \boldsymbol{\beta}, \boldsymbol{\upsilon}(\boldsymbol{b}) \rangle - \langle \vec{R}^a, \boldsymbol{b} \rangle - \theta \langle \boldsymbol{\beta}, \bar{\boldsymbol{\upsilon}}(\boldsymbol{b}, a) \rangle$ is the rate of decrease in the objective function of the dual problem per unit increase in the value of $\lambda_{\boldsymbol{b},a}$ – the expected discounted time that action a is taken for $\boldsymbol{b}$. Therefore, we can define a rate measure as $\langle \boldsymbol{\beta}^*, \boldsymbol{\upsilon}(\boldsymbol{b}) \rangle - \langle \vec{R}^0, \boldsymbol{b} \rangle - \theta \langle \boldsymbol{\beta}^*, \bar{\boldsymbol{\upsilon}}(\boldsymbol{b}, 0) \rangle$, representing the cost of taking the inactive action $a = 0$ for belief state $\boldsymbol{b}$.

It is clear that the rate measure is exactly the one-off subsidy $\mathcal{I}(\boldsymbol{b})$, hence the name.

4 Numerical Study

In this section, we numerically evaluate the performance of the approximate-measure policy and the rate-measure policy. We first compare the two heuristic policies with a random policy, and then compare the approximate-measure policy with the third-order measure policy and the myopic policy.

Suppose we have M identical systems (e.g., M wind turbines in a wind farm) and κ repairmen. Hence, at any decision epoch, at most κ systems can be maintained. Four actions $\{0,1,2,3\}$ are available to each system, with actions $a = 0$ and $a = 1$ respectively representing "doing nothing" and "replacement". Each system has 4 states, labelled by $\{1,2,3,4\}$ from the worst state to the pristine state. The transition matrix for action $a = 0$ takes the form

$$P^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ p^0_{21} & 1-p^0_{21} & 0 & 0 \\ p^0_{31} & p^0_{32} & 1-\sum p^0_{3j} & 0 \\ p^0_{41} & p^0_{42} & p^0_{43} & 1-\sum p^0_{4j} \end{bmatrix},$$

and $P^1 = (\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{1})$. The transition matrices for actions $a = 2$ and $a = 3$ take the form

$$P^2 = \begin{bmatrix} 1-\sum p_{1j}^2 & p_{12}^2 & p_{13}^2 & p_{14}^2 \\ 0 & 1-\sum p_{2j}^2 & p_{23}^2 & p_{24}^2 \\ 0 & 0 & 1-p_{34}^2 & p_{34}^2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } P^3 = \begin{bmatrix} 1-\sum p_{1j}^3 & p_{12}^3 & p_{13}^3 & p_{14}^3 \\ 0 & 1-\sum p_{2j}^3 & p_{23}^3 & p_{24}^3 \\ 0 & 0 & 1-p_{34}^3 & p_{34}^3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

All the probabilities $\{p_{21}^0, p_{31}^0, p_{32}^0, p_{41}^0, p_{42}^0, p_{43}^0\}$, $\{p_{12}^2, p_{13}^2, p_{14}^2, p_{23}^2, p_{24}^2, p_{34}^2\}$ and $\{p_{12}^3, p_{13}^3, p_{14}^3, p_{23}^3, p_{24}^3, p_{34}^3\}$ are randomly generated, subject to the following conditions:

- all the diagonal entries are positive, and
- action 3 is more efficient (and hence more costly) than action 2: $\sum_{k=1}^{j} p_{ik}^3 \leq \sum_{k=1}^{j} p_{ik}^2$, $\forall$ $j = 1,\ldots,4$ and $i = 1,\ldots,4$.

Correspondingly, the reward structure is specified as follows:

$$R = \begin{bmatrix} R_{s=1}^{a=0} & R_{s=1}^{a=2} & R_{s=1}^{a=3} & R_{s=1}^{a=1} \\ R_{s=2}^{a=0} & R_{s=2}^{a=2} & R_{s=2}^{a=3} & R_{s=2}^{a=1} \\ R_{s=3}^{a=0} & R_{s=3}^{a=2} & R_{s=3}^{a=3} & R_{s=3}^{a=1} \\ R_{s=4}^{a=0} & R_{s=4}^{a=2} & R_{s=4}^{a=3} & R_{s=4}^{a=1} \end{bmatrix} = \begin{bmatrix} -100 & 10 & 30 & 60 \\ 12 & 20 & 65 & 50 \\ 40 & 66 & 50 & 40 \\ 80 & 60 & 45 & 25 \end{bmatrix}.$$

The observation space $\mathcal{Z}$ is the (0, 1) interval, and the observation function $f_s^a(z)$ is a beta density which we assume depends only on the true state s, not the action a. Hence let the four observation density functions, corresponding to states $\{1,2,3,4\}$, be Beta(2, 8), Beta(8, 12), Beta(12, 8) and Beta(8, 2). The discount rate is 0.95.

The transition matrices used for the following performance evaluation are (rounded to two decimal places):

$$P^0 = \begin{bmatrix} 1.00 & 0 & 0 & 0 \\ 0.71 & 0.29 & 0 & 0 \\ 0.48 & 0.26 & 0.26 & 0 \\ 0.29 & 0.20 & 0.22 & 0.29 \end{bmatrix}, P^2 = \begin{bmatrix} 0.32 & 0.29 & 0.18 & 0.21 \\ 0 & 0.40 & 0.29 & 0.31 \\ 0 & 0 & 0.53 & 0.47 \\ 0 & 0 & 0 & 1.00 \end{bmatrix}, P^3 = \begin{bmatrix} 0.27 & 0.21 & 0.28 & 0.24 \\ 0 & 0.30 & 0.23 & 0.47 \\ 0 & 0 & 0.42 & 0.58 \\ 0 & 0 & 0 & 1.00 \end{bmatrix}.$$

The total discounted reward and the total EDR are employed as the criteria for evaluating different policies. At time 0, given the M belief states $(\boldsymbol{b}^0(1),\ldots,\boldsymbol{b}^0(M))$, measure the importance of each system and then take the portfolio of maintenance actions determined by the corresponding heuristic policy. M rewards $(R_{s_0^1}^{a_0^1}, R_{s_0^2}^{a_0^2}, \ldots, R_{s_0^M}^{a_0^M})$ are obtained for time 0. At time 1, update the belief

states according to the belief states at time 0, maintenance actions at time 0, and observations at time 1; measure the importance of each system and then take the implied portfolio of maintenance actions. M rewards $(R_{s_1^1}^{a_1^1}, R_{s_1^2}^{a_1^2}, \ldots, R_{s_1^M}^{a_1^M})$ are obtained for time 1. Repeat the procedure until arriving at time 90. (We approximate the infinite horizon by a finite horizon of 90 units of time as $0.95^{90} < 0.01$.) The total discounted reward is simply $\sum_{m=1}^{M}\sum_{t=0}^{90}\theta^t R_{s_t^m}^{a_t^m}$. By repeating the above procedure for 1000 times, we approximate the total EDR by the average of the 1000 total discounted rewards.

4.1 Evaluating the Two Heuristic Policies

Generally, the relative suboptimality gap is employed as the performance measure: $\frac{V(\boldsymbol{b}_{1:M}^0)-V_i(\boldsymbol{b}_{1:M}^0)}{V(\boldsymbol{b}_{1:M}^0)}$, where $\boldsymbol{b}_{1:M}^0 = (\boldsymbol{b}^0(1),\ldots,\boldsymbol{b}^0(M))$, $V(\boldsymbol{b}_{1:M}^0)$ is the total EDR under the optimal policy, and $V_i(\boldsymbol{b}_{1:M}^0)$ is the total EDR under a heuristic policy. However, evaluating the optimal policy is PSPACE hard. Hence, instead of the optimal policy, we compare with a random policy in which we randomly select κ out of all the systems that need be maintained. Let $V(\boldsymbol{b}_{1:M}^0)$ be the total EDR under the random policy.

Set M to be 10, and let κ in turn take a value from $\{2,4,6,8\}$. Randomly generate one set of M starting belief states $(\boldsymbol{b}^0(1),\ldots,\boldsymbol{b}^0(M))$. To calculate the original optimal action for any given belief state, the optimal value function for each system is approximated by a set of 10000 α-vectors (Hauskrecht, 2000). In Figures 1-3, the red solid curve corresponds to the approximate-measure policy, the blue dashed curve corresponds to the rate-measure policy, and the black dotdash curve corresponds to the random policy. Figure 1 plots the total discounted reward $\sum_{m=1}^{M}\sum_{t=0}^{90}\theta^t R_{s_t^m}^{a_t^m}$ for each of the 1000 repeats, and Table 1 gives the mean value of the 1000 total discounted rewards. As

Table 1: Average of the 1000 total discounted rewards for different κ values.

	$\kappa=2$	$\kappa=4$	$\kappa=6$	$\kappa=8$
Approximate	11900.10	12627.15	13106.80	13219.79
Rate	1303.41	12324.83	13108.82	13218.77
Random	-1251.84	9739.55	12748.75	13137.71

stated in Section 3.2, the performance of the rate-measure policy depends on the ratio $\frac{\kappa}{M}$, the larger the better. Figure 1 and Table 1 show that, when the ratio $\frac{\kappa}{M}$ is larger than 0.5, the rate-measure policy and the approximate-measure policy have the same performance. Hence, when the ratio is larger than 0.5, we can use only the rate measure, as calculating the rate measure is faster than

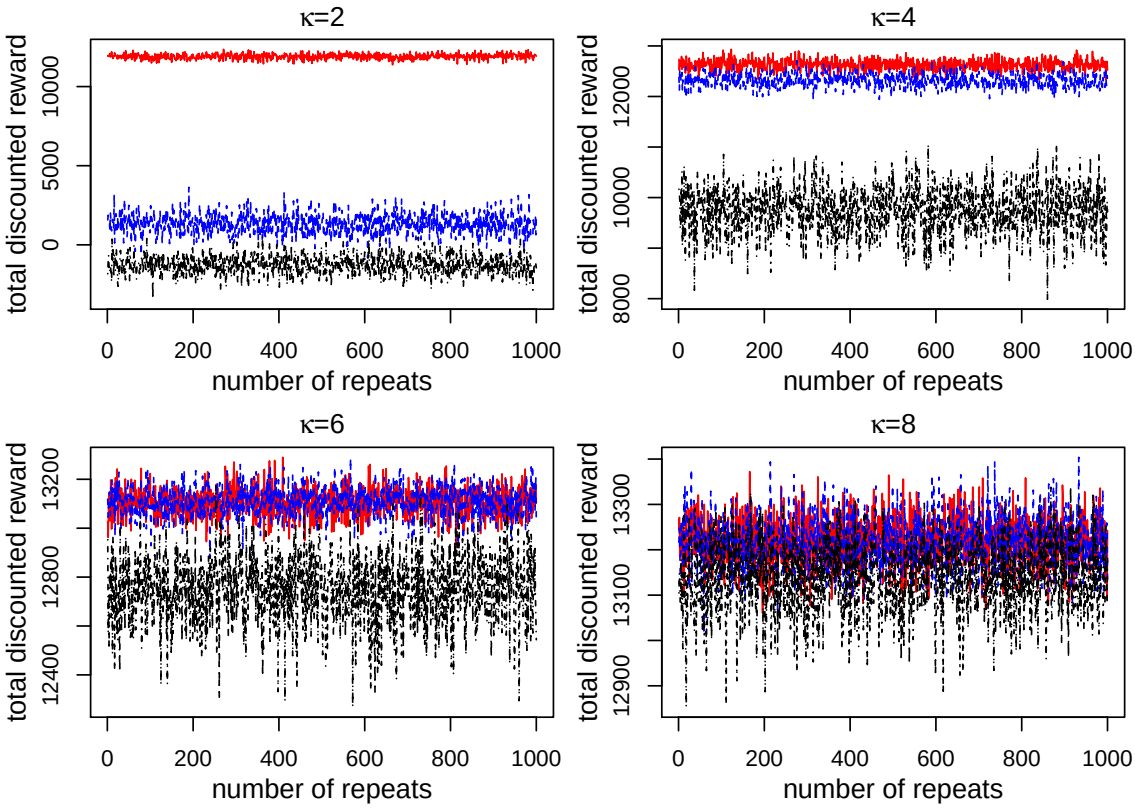

Figure 1: Total discounted rewards under 1000 repeats; each panel corresponds to a different κ value. The red solid curve corresponds to the approximate-measure policy, the blue dashed curve corresponds to the rate-measure policy, and the black dotdash curve corresponds to the random policy.

calculating the approximate measure. When the ratio is smaller than 0.5, the approximate-measure policy outperforms the rate-measure policy. In each case, the random policy performs the worst, with the 1000 total discounted rewards having low mean value and large variance.

To further examine the influence of the ratio $\frac{\kappa}{M}$, we now fix κ at 12 and let M in turn take a value from $\{15, 20, 30, 60\}$, making the ratio $\frac{\kappa}{M}$ take the values $\{0.2, 0.4, 0.6, 0.8\}$. With the randomly generated initial belief states $(\boldsymbol{b}^0(1), \ldots, \boldsymbol{b}^0(M))$ being fixed, simulate a Markovian maintenance decision process until arriving at time 90, and then calculate the total discounted reward $\sum_{m=1}^{M} \sum_{t=0}^{90} \theta^t R_{s_t^m}^{a_t^m}$. Repeat the procedure for 1000 times to obtain 1000 total discounted rewards. Figure 2 plots the 1000 total discounted rewards, where each panel corresponds to a different M value. Table 2 lists the total EDRs. It is clear from Figure 2 and Table 2 that the performance of the rate-measure policy and the random policy depends on the ratio $\frac{\kappa}{M}$. The approximate-measure

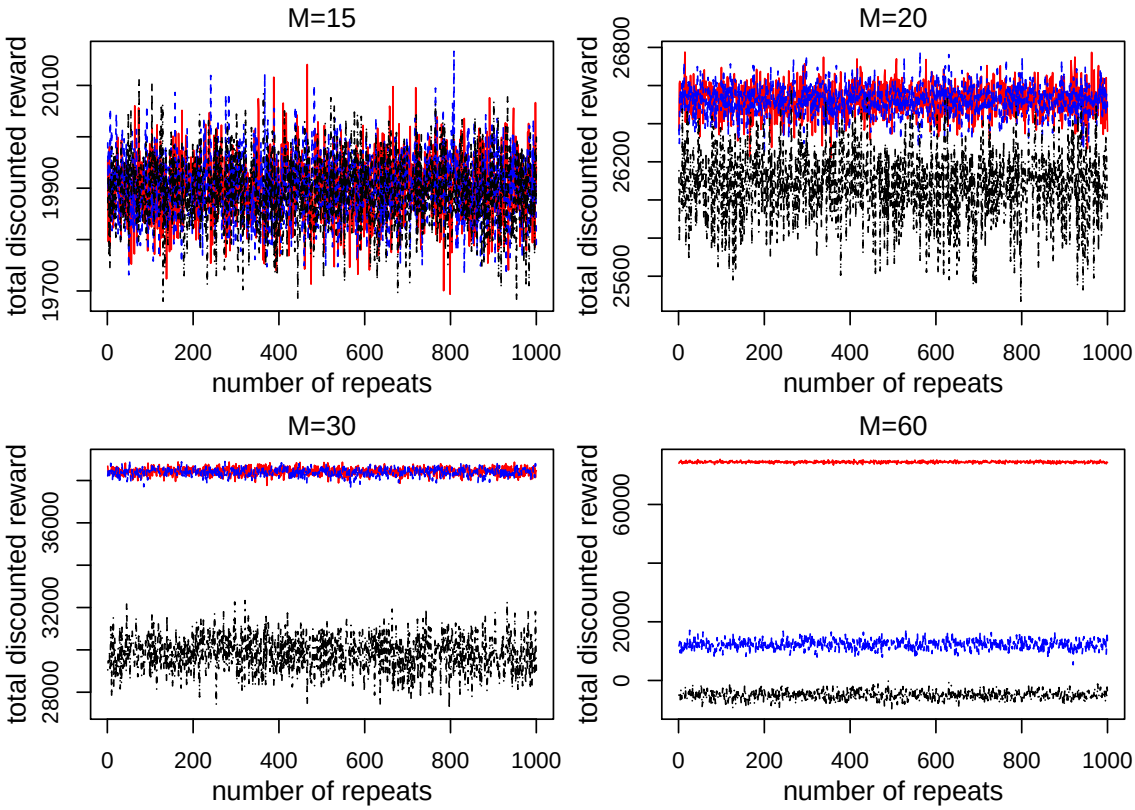

Figure 2: Total discounted rewards under 1000 repeats; each panel corresponds to a different M value. The red solid curve corresponds to the approximate-measure policy, the blue dashed curve corresponds to the rate-measure policy, and the black dotdash curve corresponds to the random policy.

Table 2: Average of the 1000 total discounted rewards for different M values.

	$M = 15$	$M = 20$	$M = 30$	$M = 60$
Approximate	19902.07	26522.33	38425.85	74462.71
Rate	19903.74	26522.28	38385.00	12174.22
Random	19895.64	26076.25	29840.14	-4998.60

policy outperforms the others when $\frac{\kappa}{M} < 0.5$; the large gap between the total discounted rewards of the random policy and the approximate-measure policy verifies the efficiency of the approximate-measure policy.

In each panel of Figures 1 and 2, the 1000 total discounted rewards are originated from one realization of $\boldsymbol{b}^0_{1:M}$. By averaging the 1000 total discounted rewards, we obtain an estimate of the

function value $V_i(\boldsymbol{b}^0_{1:M})$ or $V(\boldsymbol{b}^0_{1:M})$, i.e., the total EDR. Figure 3 plots the function values $V_i(\boldsymbol{b}^0_{1:M})$

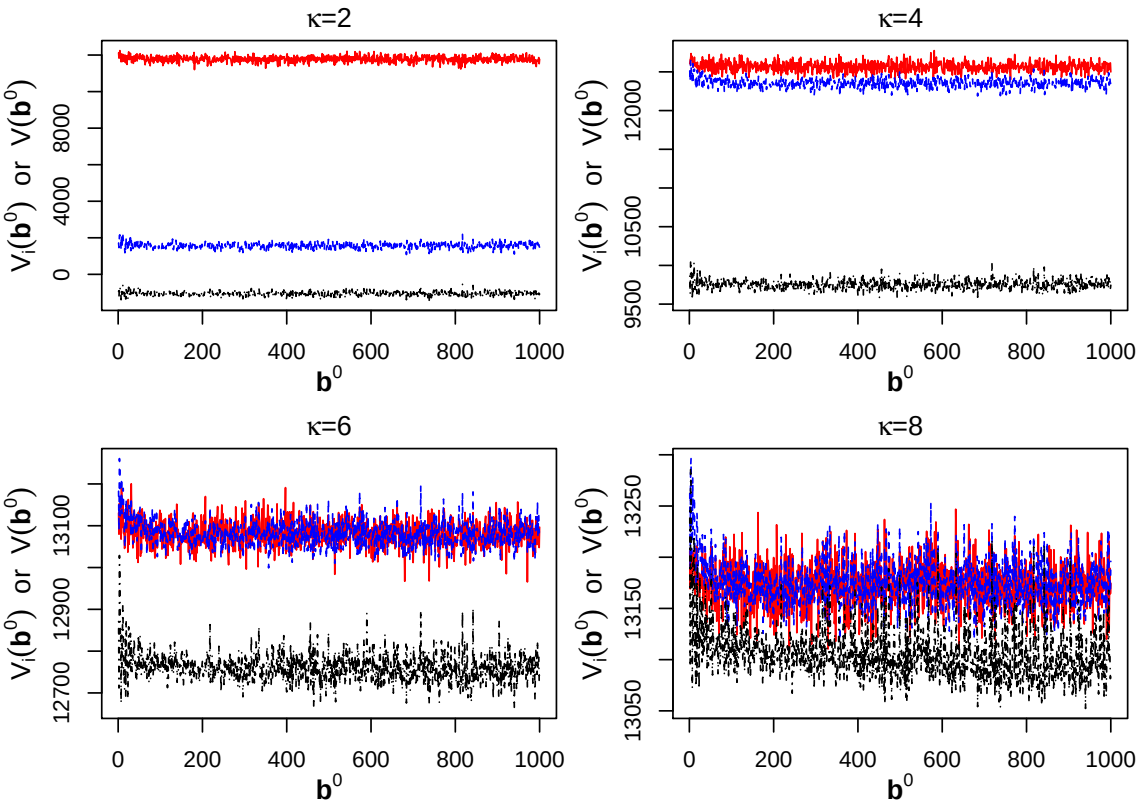

Figure 3: Function values of $V_i(\boldsymbol{b}^0)$ and $V(\boldsymbol{b}^0)$ for 1000 different $\boldsymbol{b}^0$; each panel corresponds to a different κ value. The red solid curve corresponds to the approximate-measure policy, the blue dashed curve corresponds to the rate-measure policy, and the black dotdash curve corresponds to the random policy.

and $V(\boldsymbol{b}^0_{1:M})$ for 1000 different sets of starting belief states. Consistent with Figures 1 and 2, Figure 3 indicates that

- when $\frac{\kappa}{M}$ is smaller than 0.5, the approximate-measure policy has the best performance;
- when $\frac{\kappa}{M}$ is larger than 0.5, the approximate-measure policy and the rate-measure policy have the same performance, but calculating the rate measure is faster than calculating the approximate measure;
- the large gap between $V_i(\boldsymbol{b}^0_{1:M})$ and $V(\boldsymbol{b}^0_{1:M})$ verifies the exceptional competence of the approximate measure.

To decide which importance measure to apply for a particular problem, one can calculate both the approximate measure and the rate measure for the first few decision epochs. If the two measures produce very similar total rewards, then it is safe to use only the rate measure for the following decision epochs. Note that, for either type, the M importance measures for the M systems can be calculated in parallel.

4.2 Comparing with the Third-Order Approximation

To further reveal the competence of the approximate measure, we here compare the approximate-measure policy with the myopic policy and the third-order measure policy.

We first fix κ at 12 and let M in turn take a value from $\{15, 20, 30, 60\}$. Randomly generate one set of M starting belief states $(\boldsymbol{b}^0(1), \ldots, \boldsymbol{b}^0(M))$; by simulating 1000 Markovian maintenance decision process, we obtain 1000 total discounted rewards $\sum_{m=1}^{M} \sum_{t=0}^{90} \theta^t R_{s_t^m}^{a_t^m}$. Figure 4 plots the 1000 total discounted rewards, where each panel corresponds to a different M value. The average of the 1000 total discounted rewards, i.e. the total EDR of the belief states $(\boldsymbol{b}^0(1), \ldots, \boldsymbol{b}^0(M))$, for each M value is given in Table 3. Instead of one single set of starting belief states, Figure 5 further plots

Table 3: Average of the 1000 total discounted rewards for different M values.

	$M = 15$	$M = 20$	$M = 30$	$M = 60$
Third-Order	19934.66	26517.35	38515.46	71066.91
Approximate	19935.25	26518.42	38623.99	74389.85
Myopic	19932.43	26515.57	38482.29	13237.68

the total EDRs for 1000 different sets of starting belief states. The average of the 1000 total EDRs for the third-order measure policy is respectively 19768.63 (M=15), 26235.26 (M=20), 37912.11 (M=30), and 70445.24 (M=60); the average for the approximate-measure policy is respectively 19769.50 (M=15), 26237.64 (M=20), 38131.8 (M=30), and 74017.96 (M=60); the average for the myopic policy is respectively 19767.57 (M=15), 26231.82 (M=20), 37749.03 (M=30), and 11097.95 (M=60). From Figures 4 and 5 and Table 3, it is clear that the approximate-measure policy frequently gives a higher total EDR than the third-order measure policy. Particularly, when the ratio $\frac{\kappa}{M}$ is small, the approximate-measure policy always outmatches the third-order measure policy in terms of the total EDR. Therefore, we claim that the second-order approximation is superior to the third-order approximation: the computation for the second-order approximation is much

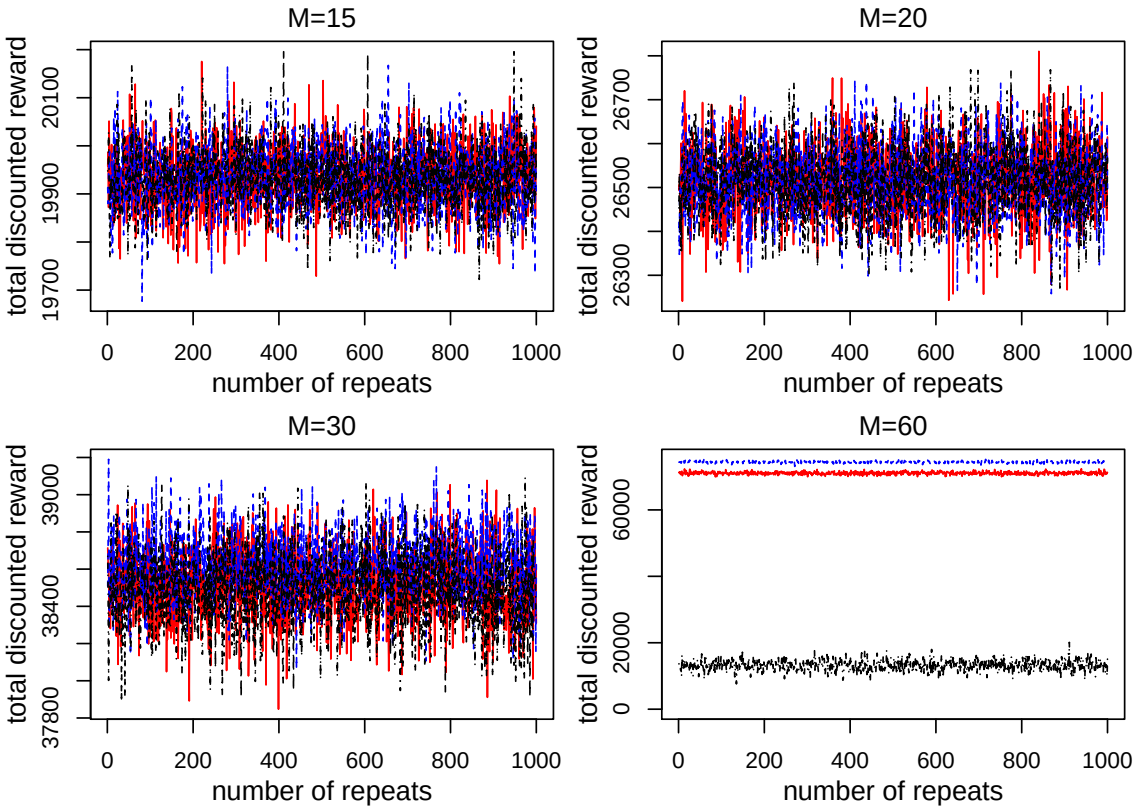

Figure 4: Total discounted rewards under 1000 repeats; each panel corresponds to a different M value. The red solid curve corresponds to the third-order measure policy, the blue dashed curve corresponds to the approximate-measure policy, and the black dotdash curve corresponds to the myopic policy.

less demanding. The large gap between the total discounted rewards of the myopic policy and the approximate-measure policy when $\frac{\kappa}{M} = 0.2$ further approves the dominance of the second-order approximation.

We then fix M at 10, and let κ in turn take a value from $\{2,4,6,8\}$. Figure 6 plots the total EDRs for 1000 different sets of starting belief states. The average of the 1000 total EDRs for the third-order measure policy is respectively 11796.36 (κ=2), 12569.76 (κ=4), 13084.63 (κ=6), and 13174.30 (κ=8); the average for the approximate-measure policy is respectively 12353.84 (κ=2), 12710.02 (κ=4), 13089.24 (κ=6), and 13179.97 (κ=8); the average for the myopic policy is respectively 2095.81 (κ=2), 12361.72 (κ=4), 13081.00 (κ=6), and 13172.86 (κ=8). Figure 6 further verifies the exceptional competence of the second-order approximation. The approximate-measure policy prevails over the third-order measure policy in terms of both total EDR and computational

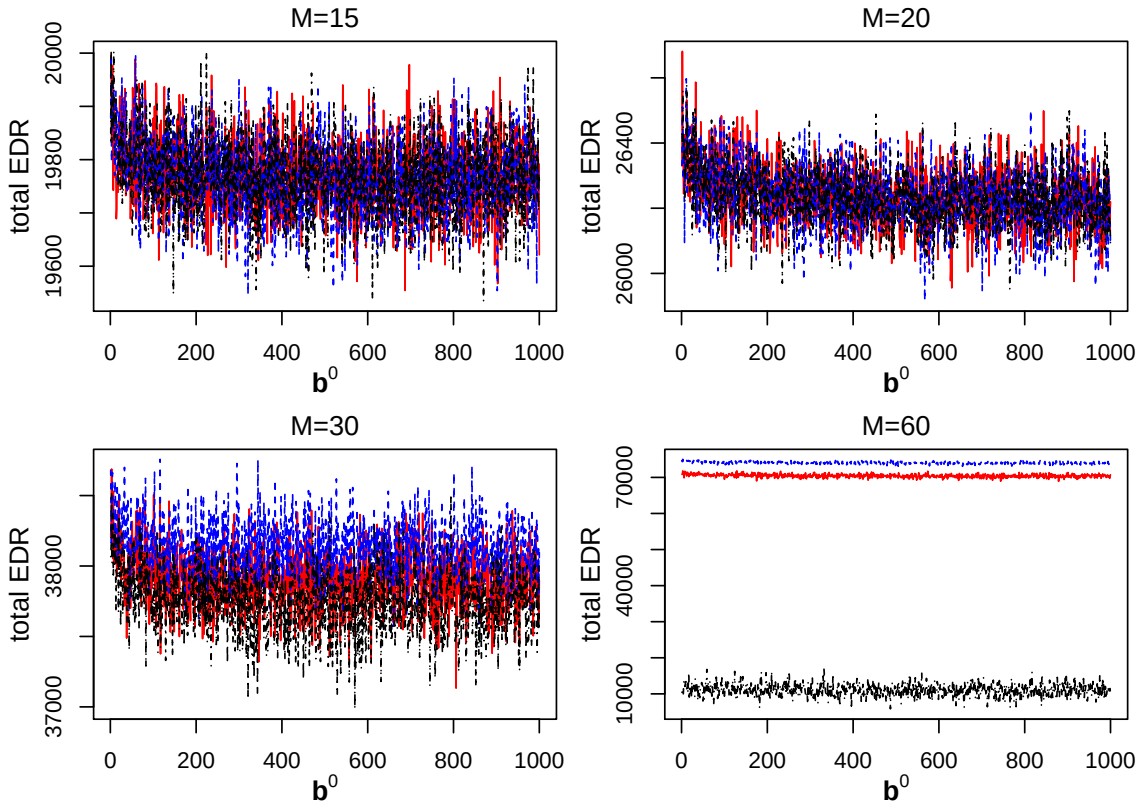

Figure 5: Total EDRs for 1000 different $\boldsymbol{b}^0$; each panel corresponds to a different M value. The red solid curve corresponds to the third-order measure policy, the blue dashed curve corresponds to the approximate-measure policy, and the black dotdash curve corresponds to the myopic policy.

cost. The myopic policy, though better than the random policy, still produces a much lower total EDR when $\frac{\kappa}{M} = 0.2$.

5 Conclusion and Further Research

This paper studies a PSPACE-hard problem: maintaining a collection of M (> 1) multi-state systems with only $\kappa (< M)$ repairmen. Two efficient importance measures are developed: the approximate measure and the rate measure. Under either heuristic policy, at each decision epoch, if the number of positive importance measures is larger than κ, then only κ multi-state systems with larger importance measures will receive their optimal actions. The two importance measures have the advantages that (1) they are well defined for POMDPs, and (2) the computation of the two measures is not demanding. Numerical studies showed that the approximate-measure policy has

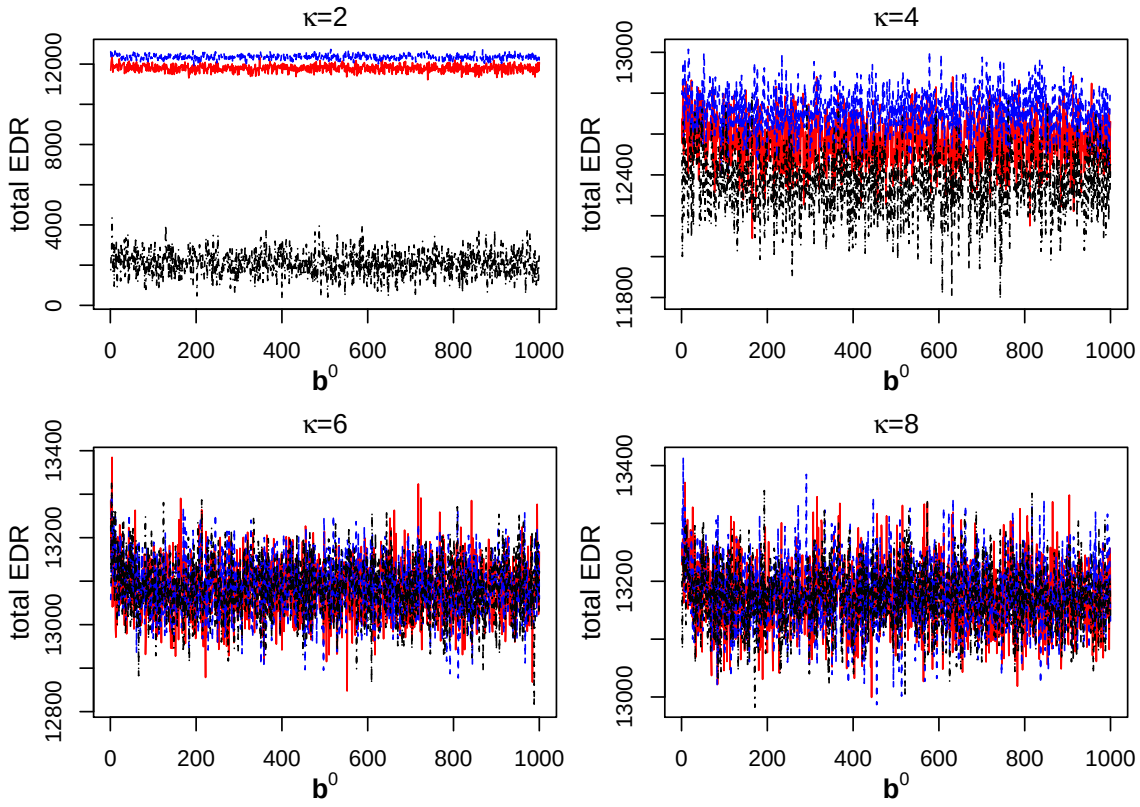

Figure 6: Total EDRs for 1000 different $\boldsymbol{b}^0$; each panel corresponds to a different κ value. The red solid curve corresponds to the third-order measure policy, the blue dashed curve corresponds to the approximate-measure policy, and the black dotdash curve corresponds to the myopic policy.

exceptional performance, and when the ratio $\frac{\kappa}{M}$ is large, the rate-measure policy is also outstanding. But calculating the rate measure is faster than calculating the approximate measure. Hence, the approximate measure and the rate measure can be applied to different settings. To decide which importance measure to use, one can calculate both importance measures for the first few decision epochs. If the two measures produce very similar total rewards, then one can switch to the rate measure for the following decision epochs. R codes for the above numerical study are available on request.

As future work, it is necessary to further provide provable bounds or establish asymptotic optimality of the proposed heuristics. Moreover, we found that if the actions can be ordered in certain way, then the ranking of the approximate importance measures is often the same with the ranking of the optimal actions; in other words, the rank of the optimal action indicates the importance of the multi-state system at the decision epoch. More study need be taken to examine

under which conditions such a relationship holds.

Appendix A A Two-Action Maintenance Problem

We here study a two-action maintenance problem: available maintenance actions are either "doing nothing" or "replacement". Arrange the states w.r.t. the level of degradation: the first state represents the worst machine condition, while the last state represents the pristine condition.

In the context of machine maintenance, if the do-nothing action is taken, then the condition of the machine will degrade. Hence, the transition matrix for the do-nothing action, denoted by $P^0 = (p^0_{ss'})$, is a lower triangular matrix; the main diagonal entries are smaller than 1 except the first entry. For a belief state $\boldsymbol{b}$, if we take the non-optimal action $a = 0$, then at the following epoch, action $a = 0$ will still be non-optimal. In other words, if the machine is in need of replacement but we do nothing, then at the following epoch the machine becomes more deteriorated, and hence replacement becomes more urgent.

The action "replacement" (labelled by the number 1) restores the machine condition to brand new. Hence, the transition matrix for the action "replacement", denoted by $P^1 = (p^1_{ss'})$, has the structure that the last column is the vector $\mathbf{1}$ while all the other entries are 0. Then it is readily to prove that

$$\ell(\boldsymbol{b}, a = 1, z) = \frac{F^1(z)(P^1)^T\boldsymbol{b}}{\mathbf{1}^T F^1(z)(P^1)^T\boldsymbol{b}} = (0,\ldots,0,1)^T, \quad \forall\, \boldsymbol{b} \text{ and } z,$$

where $F^a(z) = \mathrm{diag}(f^a_s(z) : s \in \mathcal{S})$ is a diagonal matrix, $\mathbf{1} = (1,\ldots,1)$ is the column vector of 1's, and T is the transpose operator. That is, after the "replacement" action, our belief state changes to $(0,\ldots,0,1)$ – we actually know that the machine is now in the pristine state. We write $\boldsymbol{e}$ as a notational shorthand for $(0,\ldots,0,1)$.

For the POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, p^a_{ss'}, f^a_s(z), R^a_s + w\delta(a=0), \theta)$, define the (stationary) stopping time

$$t_w := \min\{t : t \geq 1, \text{ the action at time } t \text{ is replacement.}\}.$$

Define two vectors of rewards: $\vec{R}^0 = (R^0_s : s \in \mathcal{S})$ and $\vec{R}^1 = (R^1_s : s \in \mathcal{S})$. Denote $\vec{\mathcal{R}}^0 = \vec{R}^0 + w$ and $\vec{\mathcal{R}}^1 = \vec{R}^1$. Let π^*_w be the optimal policy for the POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, p^a_{ss'}, f^a_s(z), R^a_s + wI(a=0), \theta)$. We have

$$\begin{aligned}
V_{\pi^*_w}(\boldsymbol{b}) &= \mathrm{E}[\mathcal{R}^{a_0}_{s_0} + \theta\mathcal{R}^{a_1}_{s_1} + \ldots + \theta^{t_w}\mathcal{R}^1_{s_{t_w}} | \boldsymbol{b}^0 = \boldsymbol{b}, \pi^*_w] + \mathrm{E}[\theta^{t_w+1}(\mathcal{R}^{a_{t_w+1}}_{s_{t_w+1}} + \theta\mathcal{R}^{a_{t_w+2}}_{s_{t_w+2}} + \cdots) | \boldsymbol{b}^0 = \boldsymbol{b}, \pi^*_w] \\
&= \mathrm{E}[\mathcal{R}^{a_0}_{s_0} + \theta\mathcal{R}^{a_1}_{s_1} + \ldots + \theta^{t_w}\mathcal{R}^1_{s_{t_w}} | \boldsymbol{b}^0 = \boldsymbol{b}, \pi^*_w] + \mathrm{E}[\theta^{t_w+1} V_{\pi^*_w}(\boldsymbol{e}) | \boldsymbol{b}^0 = \boldsymbol{b}, \pi^*_w] \\
&= \sup_{\tau \geq 1} \left\{ \mathrm{E}[\mathcal{R}^{a_0}_{s_0} + \theta\mathcal{R}^0_{s_1} + \ldots + \theta^{\tau-1}\mathcal{R}^0_{s_{\tau-1}} + \theta^\tau \mathcal{R}^1_{s_\tau} | \boldsymbol{b}^0 = \boldsymbol{b}, \pi^*_w] + \mathrm{E}[\theta^{\tau+1} | \boldsymbol{b}^0 = \boldsymbol{b}, \pi^*_w] V_{\pi^*_w}(\boldsymbol{e}) \right\}.
\end{aligned}$$

If we take the do-nothing action for $\boldsymbol{b}^0$ and follow the optimal policy afterwards, then the EDR is

$$\sup_{\tau\geq 1}\{w\frac{1-\mathrm{E}[\theta^\tau|\boldsymbol{b}^0=\boldsymbol{b}]}{1-\theta}+\mathrm{E}[R^0_{s_0}+\theta R^0_{s_1}+\ldots+\theta^{\tau-1}R^0_{s_{\tau-1}}+\theta^\tau R^1_{s_\tau}|\boldsymbol{b}^0=\boldsymbol{b}]+\mathrm{E}[\theta^{\tau+1}|\boldsymbol{b}^0=\boldsymbol{b}]V_{\pi^*_w}(\boldsymbol{e})\}.$$

For notational convenience, define

$$R(\boldsymbol{b},\tau)=\mathrm{E}[R^0_{s_0}+\theta R^0_{s_1}+\ldots+\theta^{\tau-1}R^0_{s_{\tau-1}}+\theta^\tau R^1_{s_\tau}|\boldsymbol{b}^0=\boldsymbol{b}].$$

If we take the replacement action for $\boldsymbol{b}^0$ and follow the optimal policy afterwards, then the EDR is $\langle\vec{R}^1,\boldsymbol{b}\rangle+\theta V_{\pi^*_w}(\boldsymbol{e})$. Hence, action $a=0$ is optimal for $\boldsymbol{b}^0$ if and only if

$$\sup_{\tau\geq 1}\{w\frac{1-\mathrm{E}[\theta^\tau|\boldsymbol{b}^0=\boldsymbol{b}]}{1-\theta}+R(\boldsymbol{b},\tau)+\mathrm{E}[\theta^{\tau+1}|\boldsymbol{b}^0=\boldsymbol{b}]V_{\pi^*_w}(\boldsymbol{e})\}\geq\langle\vec{R}^1,\boldsymbol{b}\rangle+\theta V_{\pi^*_w}(\boldsymbol{e}),$$

which is equivalent to

$$\sup_{\tau\geq 1}\frac{R(\boldsymbol{b},\tau)-\langle\vec{R}^1,\boldsymbol{b}\rangle}{1-\mathrm{E}[\theta^\tau|\boldsymbol{b}^0=\boldsymbol{b}]}\geq\theta V_{\pi^*_w}(\boldsymbol{e})-\frac{w}{1-\theta}.$$

The l.h.s. is independent of w, while the r.h.s. is decreasing in w. Therefore, the inactive set increases with the subsidy w.

Remark 1. *For any action $a\in\mathcal{A}$, define the action region $D^a_\pi=\{\boldsymbol{b}:\pi(\boldsymbol{b})=a\}$. It is easily seen that the set of belief states where it is optimal to take action 1 is convex (and therefore connected): For any belief states $\boldsymbol{b}_1,\boldsymbol{b}_2\in D^1_{\pi^*_w}$ and any $\rho\in[0,1]$, we have*

$$\begin{aligned}V_{\pi^*_w}(\rho\boldsymbol{b}_1+(1-\rho)\boldsymbol{b}_2)&\leq\rho V_{\pi^*_w}(\boldsymbol{b}_1)+(1-\rho)V_{\pi^*_w}(\boldsymbol{b}_2)\\&=\rho\langle\vec{R}^1,\boldsymbol{b}_1\rangle+\rho\theta V_{\pi^*_w}(\boldsymbol{e})+(1-\rho)\langle\vec{R}^1,\boldsymbol{b}_2\rangle+(1-\rho)\theta V_{\pi^*_w}(\boldsymbol{e})\\&=\langle\vec{R}^1,\rho\boldsymbol{b}_1+(1-\rho)\boldsymbol{b}_2\rangle+\theta V_{\pi^*_w}(\boldsymbol{e})\\&\leq V_{\pi^*_w}(\rho\boldsymbol{b}_1+(1-\rho)\boldsymbol{b}_2),\end{aligned}$$

where we have used the fact that $V_{\pi^*_w}(\cdot)$ is a convex function. Thus all the inequalities above are equalities, and $\rho\boldsymbol{b}_1+(1-\rho)\boldsymbol{b}_2\in D^1_{\pi^*_w}$. The region $D^0_{\pi^*_w}$, however, can be disconnected. Under suitable conditions, the optimal policy π^*_w can be characterized by a single curve, which partitions the belief state space Δ into two connected regions $D^0_{\pi^*_w}$ and $D^1_{\pi^*_w}$ (Krishnamurthy, 2016, Chapter 12). Then the importance measure for a belief state $\boldsymbol{b}$ is the value w making the switching curve passing through $\boldsymbol{b}$. The curve can be estimated via simulation based stochastic approximation algorithms.

Appendix B Proof of Proposition 1

Given $\boldsymbol{b}^t = \boldsymbol{b}$ and $a_t = a$, the observation space $\mathcal{Z}$ can be divided into $|\mathcal{A}|$ different sets $\{\mathcal{Z}^{\tilde{a}}_{\boldsymbol{b},a} : \tilde{a} \in \mathcal{A}\}$ such that

$$\max_{a_{t+1}\in\mathcal{A}} \langle \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b},a,z)\rangle = \langle \vec{R}^{\tilde{a}}, \ell(\boldsymbol{b},a,z)\rangle, \quad \text{for any } z \in \mathcal{Z}^{\tilde{a}}_{\boldsymbol{b},a}.$$

Then we have

$$\begin{aligned} V_2(\boldsymbol{b}) &= \max_{a\in\mathcal{A}}\{\langle \vec{R}^a, \boldsymbol{b}\rangle + \theta \sum_{\tilde{a}\in\mathcal{A}} \int_{\mathcal{Z}^{\tilde{a}}_{\boldsymbol{b},a}} \Pr(z_{t+1} = z|\boldsymbol{b}, a_t = a)\langle \vec{R}^{\tilde{a}}, \ell(\boldsymbol{b},a,z)\rangle dz\} \\ &= \max_{a\in\mathcal{A}}\{\langle \vec{R}^a, \boldsymbol{b}\rangle + \theta \sum_{\tilde{a}\in\mathcal{A}} \int_{\mathcal{Z}^{\tilde{a}}_{\boldsymbol{b},a}} \langle P^a F^a(z)\vec{R}^{\tilde{a}}, \boldsymbol{b}\rangle dz\} \\ &= \max_{a\in\mathcal{A}}\langle \vec{R}^a + \theta P^a \sum_{\tilde{a}\in\mathcal{A}} F^a(\mathcal{Z}^{\tilde{a}}_{\boldsymbol{b},a})\vec{R}^{\tilde{a}},\ \boldsymbol{b}\rangle, \end{aligned}$$

where $F^a(\mathcal{Z}^{\tilde{a}}_{\boldsymbol{b},a})$ is a diagonal matrix with the main diagonal entries $\{\int_{\mathcal{Z}^{\tilde{a}}_{\boldsymbol{b},a}} f^a_s(z)dz : s \in \mathcal{S}\}$.

Let the optimal action be denoted by $\ddot{a}$: $\ddot{a} = \arg\max_{a\in\mathcal{A}}\langle \vec{R}^a + \theta P^a \sum_{\tilde{a}\in\mathcal{A}} F^a(\mathcal{Z}^{\tilde{a}}_{\boldsymbol{b},a})\vec{R}^{\tilde{a}},\ \boldsymbol{b}\rangle$. Now we subsidize action $\ddot{a}$ by the amount w. Then the observation space $\mathcal{Z}$ will be divided into $|\mathcal{A}|$ new sets $\{\mathcal{Z}^{w,\tilde{a}}_{\boldsymbol{b},a} : \tilde{a} \in \mathcal{A}\}$ such that

$$\max_{a_{t+1}\in\mathcal{A}} \langle w\boldsymbol{\delta}(a_{t+1} = \ddot{a}) + \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b},a,z)\rangle = \begin{cases} w + \langle \vec{R}^{\ddot{a}}, \ell(\boldsymbol{b},a,z)\rangle, & \forall z \in \mathcal{Z}^{w,\ddot{a}}_{\boldsymbol{b},a}; \\ \langle \vec{R}^{\tilde{a}}, \ell(\boldsymbol{b},a,z)\rangle, & \forall z \in \mathcal{Z}^{w,\tilde{a}}_{\boldsymbol{b},a} \text{ and } \tilde{a} \neq \ddot{a}. \end{cases}$$

The second-order approximate function $V_2(\boldsymbol{b};w)$ can be written into

$$V_2(\boldsymbol{b};w) = \max_{a\in\mathcal{A}} \left\{ w\boldsymbol{\delta}(a = \ddot{a}) + \langle \vec{R}^a + \theta P^a \sum_{\tilde{a}\in\mathcal{A}} F^a(\mathcal{Z}^{w,\tilde{a}}_{\boldsymbol{b},a})\vec{R}^{\tilde{a}} + w\theta P^a F^a(\mathcal{Z}^{w,\ddot{a}}_{\boldsymbol{b},a})\mathbf{1},\ \boldsymbol{b}\rangle \right\}.$$

If the optimal action is $a \in \mathcal{A}/\{\ddot{a}\}$, then

$$w\theta\langle P^a F^a(\mathcal{Z}^{w,\ddot{a}}_{\boldsymbol{b},a})\mathbf{1},\ \boldsymbol{b}\rangle + \langle \vec{R}^a + \theta P^a \sum_{\tilde{a}\in\mathcal{A}} F^a(\mathcal{Z}^{w,\tilde{a}}_{\boldsymbol{b},a})\vec{R}^{\tilde{a}},\ \boldsymbol{b}\rangle \geq w + \langle \vec{R}^{\ddot{a}} + \theta P^{\ddot{a}} \sum_{\tilde{a}\in\mathcal{A}} F^{\ddot{a}}(\mathcal{Z}^{w,\tilde{a}}_{\boldsymbol{b},\ddot{a}})\vec{R}^{\tilde{a}} + w\theta P^{\ddot{a}} F^{\ddot{a}}(\mathcal{Z}^{w,\ddot{a}}_{\boldsymbol{b},\ddot{a}})\mathbf{1},\ \boldsymbol{b}\rangle.$$

On one hand, we have

$$w\theta\langle P^a F^a(\mathcal{Z}^{w,\ddot{a}}_{\boldsymbol{b},a})\mathbf{1},\ \boldsymbol{b}\rangle \leq w\theta\langle P^a F^a(\mathcal{Z})\mathbf{1},\ \boldsymbol{b}\rangle = w\theta < w.$$

On the other hand,

$$\begin{aligned}
&\langle \vec{R}^{\ddot{a}} + \theta P^{\ddot{a}} \sum_{\tilde{a}\in\mathcal{A}} F^{\ddot{a}}(\mathcal{Z}^{w,\tilde{a}}_{\boldsymbol{b},\ddot{a}})\vec{R}^{\tilde{a}} + w\theta P^{\ddot{a}} F^{\ddot{a}}(\mathcal{Z}^{w,\ddot{a}}_{\boldsymbol{b},\ddot{a}})\mathbf{1},\ \boldsymbol{b}\rangle \\
&= \langle \vec{R}^{\ddot{a}}, \boldsymbol{b}\rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z|\boldsymbol{b}, a_t = \ddot{a}) \max_{a_{t+1}\in\mathcal{A}} \langle w\boldsymbol{\delta}(a_{t+1} = \ddot{a}) + \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b}, \ddot{a}, z)\rangle dz \\
&\geq \langle \vec{R}^{\ddot{a}}, \boldsymbol{b}\rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z|\boldsymbol{b}, a_t = \ddot{a}) \max_{a_{t+1}\in\mathcal{A}} \langle \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b}, \ddot{a}, z)\rangle dz \\
&\geq \langle \vec{R}^{a}, \boldsymbol{b}\rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z|\boldsymbol{b}, a_t = a) \max_{a_{t+1}\in\mathcal{A}} \langle \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b}, a, z)\rangle dz,
\end{aligned}$$

and

$$\langle \vec{R}^{a} + \theta P^{a} \sum_{\tilde{a}\in\mathcal{A}} F^{a}(\mathcal{Z}^{w,\tilde{a}}_{\boldsymbol{b},a})\vec{R}^{\tilde{a}},\ \boldsymbol{b}\rangle \leq \langle \vec{R}^{a}, \boldsymbol{b}\rangle + \theta \int_{\mathcal{Z}} \Pr(z_{t+1} = z|\boldsymbol{b}, a_t = a) \max_{a_{t+1}\in\mathcal{A}} \langle \vec{R}^{a_{t+1}}, \ell(\boldsymbol{b}, a, z)\rangle dz.$$

Therefore, we claim that

$$\arg\max_{a\in\mathcal{A}} \left\{ w\boldsymbol{\delta}(a = \ddot{a}) + \langle \vec{R}^{a} + \theta P^{a} \sum_{\tilde{a}\in\mathcal{A}} F^{a}(\mathcal{Z}^{w,\tilde{a}}_{\boldsymbol{b},a})\vec{R}^{\tilde{a}} + w\theta P^{a} F^{a}(\mathcal{Z}^{w,\ddot{a}}_{\boldsymbol{b},a})\mathbf{1},\ \boldsymbol{b}\rangle \right\} = \ddot{a},$$

and hence the inactive set $\mathcal{P}_2(w)$ increases with the subsidy w.

References

Ahmed, A. A. A. and Liu, Y. (2019). Throughput-based importance measures of multistate production systems. *International Journal of Production Research*, 57(2):397–410.

Blackwell, D. (1965). Discounted dynamic programming. *The Annals of Mathematical Statistics*, 36(1):226–235.

Borgonovo, E., Aliee, H., Glaß, M., and Teich, J. (2016). A new time-independent reliability importance measure. *European Journal of Operational Research*, 254(2):427 – 442.

de Farias, D. P. and Roy, B. V. (2003). The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865.

Do, P. and Bérenguer, C. (2020). Conditional reliability-based importance measures. *Reliability Engineering & System Safety*, 193:106633.

Dui, H., Si, S., and Yam, R. C. (2017). A cost-based integrated importance measure of system components for preventive maintenance. *Reliability Engineering & System Safety*, 168:98 – 104.

Ellis, H., Jiang, M., and Corotis, R. B. (1995). Inspection, maintenance, and repair with partial observability. *Journal of Infrastructure Systems*, 1(2):92–99.

Hansen, E. A. (1998). Solving pomdps by searching in policy space. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 211–219.

Hauskrecht, M. (2000). Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 13(1):33–94.

Hauskrecht, M. and Kveton, B. (2004). Linear program approximations for factored continuous-state markov decision processes. In *In Advances in Neural Information Processing Systems 16*, pages 895–902.

Krishnamurthy, V. (2016). *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambridge University Press.

Liu, B., Xu, Z., Xie, M., and Kuo, W. (2014). A value-based preventive maintenance policy for multi-component system with continuously degrading components. *Reliability Engineering & System Safety*, 132:83 – 89.

Liu, K. and Zhao, Q. (2010). Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567.

Madani, O., Hanks, S., and Condon, A. (1999). On the undecidability of probabilistic planning and infinite-horizon partially observable markov decision problems. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, pages 541–548.

Ny, J. L., Dahleh, M., and Feron, E. (2008). Multi-uav dynamic routing with partial observations using restless bandit allocation indices. In *2008 American Control Conference*, pages 4220–4225.

Papadimitriou, C. H. and Tsitsiklis, J. N. (1999). The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305.

Pineau, J., Gordon, G., and Thrun, S. (2003). Point-based value iteration: An anytime algorithm for pomdps. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1025 – 1032.

Shani, G., Pineau, J., and Kaplow, R. (2013). A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51.

Sondik, E. J. (1978). The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):282–304.

Tyrväinen, T. (2013). Risk importance measures in the dynamic flowgraph methodology. *Reliability Engineering & System Safety*, 118:35 – 50.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298.

Wu, S., Chen, Y., Wu, Q., and Wang, Z. (2016). Linking component importance to optimisation of preventive maintenance policy. *Reliability Engineering & System Safety*, 146:26 – 32.

Yuan, S. and Wang, J. (2012). Sequential selection of correlated ads by pomdps. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 515–524.