Available online at www.sciencedirect.com

SCIENCE DIRECT®

Systems & Control Letters 55 (2006) 165–173

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

On the existence of stationary optimal policies for partially observed MDPs under the long-run average cost criterion

Shun-Pin Hsu[a, b], Dong-Ming Chuang[a, b], Ari Arapostathis[a, b, *]

[a] National Chi-Nan University, Electrical Engineering, 301 University Road, Puli, Nantou, Taiwan 545
[b] The University of Texas, Electrical and Computer Engineering, 1 University Station C0803, Austin, TX 78712, USA

Received 23 June 2004; received in revised form 8 June 2005; accepted 10 June 2005
Available online 2 August 2005

Abstract

This paper studies the problem of the existence of stationary optimal policies for finite state controlled Markov chains, with compact action space and imperfect observations, under the long-run average cost criterion. It presents sufficient conditions for existence of solutions to the associated dynamic programming equation, that strengthen past results. There is a detailed discussion comparing the different assumptions commonly found in the literature.

Keywords: Markov decision processes; Optimal control; Stochastic dynamic programming; Imperfect information

1. Introduction

Since the pioneering work by Bellman [2] in the 1950s, Markov decision processes (MDPs) have formed a very active research topic due to their wide applicability to problems in operations research, communication networks, economics, and other fields. Special attention has been paid to the study of models with partial observations. Whereas the finite state completely observed model is fully understood, the introduction of system state estimation results in some mathematical challenges, and there is still ample work that remains to be done on partially observed models. This is particularly true for the long-run average optimal control problem.

We adopt the model in [1,9] for a partially observed Markov decision process (POMDP) with finite state space $\boldsymbol{S} = \boldsymbol{X} \times \boldsymbol{Y}$, with $\boldsymbol{X} = \{1, 2, \dots, n\}$ and $\boldsymbol{Y} = \{1, 2, \dots, m\}$. We denote the set of natural numbers by $\mathbb{N}$, and the set of non-negative integers by $\mathbb{N}_0$. We use capital letters to denote random processes and variables and lower case letters to denote the elements of a space. Thus, we denote the state process by $\{X_t, Y_t\}_{t\in\mathbb{N}_0}$, and refer to the second component Y_t as the observation process. The action space $\boldsymbol{U}$ is assumed to be a compact metric space. The dynamics of the process are governed by a transition kernel on $\boldsymbol{X} \times \boldsymbol{Y}$ given $\boldsymbol{X} \times \boldsymbol{U}$, which may be interpreted as

$$Q_{ij}(y, u) := \text{Prob}\,(X_{t+1} = j, Y_{t+1} = y | X_t = i, U_t = u).$$

For fixed y and u, we view $Q(y, u)$ as an $n \times n$ substochastic matrix. We assume that $u \mapsto Q(y, u)$ is continuous, and that the running cost is a lower semi-continuous function $c : \boldsymbol{X} \times \boldsymbol{U} \to \mathbb{R}_+$. Only the second component of $\{Y_t\}$ is available for control purposes and reflecting this we call a sequence of controls $\{U_t\}$ admissible if for each t, U_t is conditionally independent of $\{X_{t'}, t' \leqslant t\}$ given $\{Y_{t'}, U_{t'-1}, t' \leqslant t\}$. The objective is to minimize over all admissible $\{U_t\}$ the cost

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}[c(X_t, U_t)]. \tag{1}$$

It is well known that for a POMDP model, one can derive a completely observed (CO) model which is equivalent to

[*] Corresponding author. The University of Texas, Electrical and Computer Engineering, 1 University Station C0803, Austin, TX 78712, USA.

E-mail address: ari@mail.utexas.edu (A. Arapostathis).

0167-6911/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.sysconle.2005.06.009

the original model in the sense that for every control policy in the POMDP model there corresponds a policy in the CO model that results in the same cost, and vice versa. For a discounted optimal control problem over an infinite horizon, the contraction property of the dynamic programming operator guarantees the existence of stationary optimal policies, and thus the problem is adequately characterized. In contrast, for the long-run average optimal control problem there are numerous examples for which the associated dynamic programming equation has no solution [12].

A commonly used method for studying the problem of existence of solutions to the average cost dynamic programming equation (ACOE) is the vanishing-discount method, an asymptotic method based on the solution of the much better understood discounted cost problem [6–8,10,12]. There are two main reasons why this approach is chosen. First, the theorem of Tauber which relates the asymptotic behavior of Cesaro sums to the behavior of its discounted averages. Second, it is well known, at least in the case of finite state and action models, that if the ACOE admits a bounded solution, then one such solution can always be obtained as the limit of a sequence of differential discounted value functions, as the discount factor tends to 1. In Theorem 7 we extend this result to models with a compact action space.

Most of the well known sufficient conditions for the existence of solutions to the ACOE, including Ross's *renewability* condition [14], Platzman's *reachability–detectability* condition [12], and Stettner's *positivity* condition [15], impose assumptions that need to be satisfied by all policies. As a result, these conditions fail in even some simple problems that are known to possess stationary optimal policies. We discuss various assumptions in detail in Section 3. Here, we present a brief description to help guide the reader. We use the following notation: For $k \in \mathbb{N}$, let $y^k = (y_1, \dots, y_k)$ and $u^k = (u_0, \dots, u_{k-1}) \in \boldsymbol{U}^k$ denote elements of $\boldsymbol{Y}^k$ and $\boldsymbol{U}^k$, respectively. Define

$$Q(y^k, u^k) = Q(y_1, u_0)Q(y_2, u_1) \cdots Q(y_k, u_{k-1}). \tag{2}$$

Similarly, if we let Y^k and U^k represent the vector-valued random variables $(Y_1, \dots, Y_k)$ and $(U_0, \dots, U_{k-1})$, respectively, the random variable $Q(Y^k, U^k)$ is defined in complete analogy to (2).

In [4], existence of stationary optimal policies is studied under the following assumption: For some $k \geqslant 1$,

$$\min_{(y^k,u^k)\in \boldsymbol{Y}^k \times \boldsymbol{U}^k} \left[\frac{\min_{ij} Q_{ij}(y^k, u^k)}{\max_{ij} Q_{ij}(y^k, u^k)} \right] > 0. \tag{3}$$

However, the proof supplied in [4] has an error (see [5]). In an earlier unpublished work, the same author considered the following condition: There exists $k \geqslant 1$, and $\Delta > 0$ such that, for all admissible $\{U_t\}$

$$\mathbb{P}(X_k = j | X_0 = i, U_{t-1}, Y_t, 1 \leqslant t \leqslant k) \geqslant \Delta \quad \forall i, j \in \boldsymbol{X}.$$

This condition is equivalent to the requirement that for all $y^k \in \boldsymbol{Y}^k$ and $u^k \in \boldsymbol{U}^k$,

$$Q_{ij}(y^k, u^k) \geqslant \Delta \sum_{\ell \in \boldsymbol{X}} Q_{i\ell}(y^k, u^k) \quad \forall i, j \in \boldsymbol{X}. \tag{4}$$

Clearly, (3) implies (4). Also (4) implies Platzman's conditions, which are discussed in Section 3. However, Platzman in [12] assumes that $\boldsymbol{U}$ is finite, and as far as we know, there is no satisfactory proof for the existence of solutions to the ACOE under (4), when $\boldsymbol{U}$ is a compact metric space. In this work, we establish the existence of solutions to the ACOE under conditions that are weaker than (4). These conditions are stated in Section 3 in a policy independent form as Assumption 4, and also in an even weaker form as Assumption 2, which needs only be satisfied over the stationary optimal policies of the discounted problem. As shown through two examples in Section 5, if some structural properties of the discounted optimal policies are known, these conditions can be readily verified.

We also show that existence of solutions to the ACOE is guaranteed by the following condition, which in some sense is a dual to (4): There exists $k \geqslant 1$, and $\Delta > 0$ such that, for all $y^k \in \boldsymbol{Y}^k$ and $u^k \in \boldsymbol{U}^k$,

$$Q_{ij}(y^k, u^k) \geqslant \Delta \sum_{\ell \in \boldsymbol{X}} Q_{\ell j}(y^k, u^k) \quad \forall i, j \in \boldsymbol{X}. \tag{5}$$

Condition (5) is weaker than the positivity assumption in [15], and we use it in an even weaker form which is stated in Assumption 5 in Section 3.

In a recent series of papers [6–8] introduced the notion of "wide sense admissible" controls, and has used coupling arguments to obtain the "vanishing discount" limit. The method in [7,8] is fairly sophisticated and the results are applicable to models with non-finite state space. It is more appropriate to compare the results in this paper with [6], where the state space is assumed finite. There are two key differences between [6] and this paper: (a) the main assumption in [6] (Assumption 3.2) holds over all wide sense admissible controls, whereas the assumptions in this paper are stated over the class of stationary, discounted optimal controls, and (b) the vanishing discount limit results in a continuous limiting value function in [6], whereas this is not necessarily the case in this paper. Example 15 in Section 5 exhibits a model for which Assumption 6 in Section 3 holds but Assumption 3.2 of [6] fails.

The paper is organized as follows: In Section 2 we summarize the construction of the equivalent CO model. In Section 3 we state and compare various assumptions in the literature. Section 4 contains the main results of the paper. Finally, Section 5 demonstrates the results through three examples.

2. The equivalent CO model

In this section we summarize the equivalent CO model, following for the most part [1,9]. Let $\boldsymbol{\Psi} := \mathscr{P}(\boldsymbol{X})$ denote the set of probability distributions on $\boldsymbol{X}$. Let $\mathbf{1}$ denote the element of $\mathbb{R}^n$ with entries equal to 1 and $\bar{\psi}$ be the element of $\boldsymbol{\Psi}$ satisfying $\bar{\psi}(i) = 1/n$, for all $i \in \boldsymbol{X}$. Viewing $\psi \in \boldsymbol{\Psi}$ as a row vector, we define

$$V(\psi, y, u) := \psi Q(y, u)\mathbf{1},$$
$$T(\psi, y, u) := \begin{cases} \dfrac{\psi Q(y, u)}{V(\psi, y, u)} & \text{if } V(\psi, y, u) \neq 0, \\ \bar{\psi} & \text{otherwise.} \end{cases} \tag{6}$$

Eq. (6) is known as the filtering equation. Note that when $V(\psi, y, u) = 0$, then $T(\psi, y, u)$ can be arbitrarily chosen. Here, we choose to specify its value as $\bar{\psi}$ only for convenience, and we use this definition throughout the paper. The state space of the equivalent CO model is $\boldsymbol{\Psi} \times \boldsymbol{Y}$, with transition kernel given by

$$\mathscr{K}(B, y|\psi, u) = V(\psi, y, u)\mathbb{I}_B(T(\psi, y, u)), \quad B \in \mathscr{B}(\boldsymbol{\Psi}),$$

where $\mathscr{B}(\boldsymbol{\Psi})$ denotes the Borel σ-field of $\boldsymbol{\Psi}$, and $\mathbb{I}_B$ denotes the indicator function of B. The running cost of the CO model is chosen as

$$\tilde{c}(\psi, u) := \sum_{x \in \boldsymbol{X}} c(x, u)\psi(x), \quad \psi \in \boldsymbol{\Psi}, \;\; u \in \boldsymbol{U}.$$

If $\mu_0 \in \mathscr{P}(\boldsymbol{X} \times \boldsymbol{Y})$ is the initial state distribution of the POMDP model, then disintegrating this distribution as $\mu_0(x, y) = q_0(y)\psi_0(x|y)$, where q_0 is the marginal on $\boldsymbol{Y}$, we select the initial distribution of the equivalent CO model as $\mu \in \mathscr{P}(\boldsymbol{\Psi} \times \boldsymbol{Y})$ defined by

$$\mu(\psi, y) = q_0(y)\mathbb{I}_\psi(\psi_0(\cdot|y)). \tag{7}$$

In this manner we obtain a CO model $(\boldsymbol{\Psi} \times \boldsymbol{Y}, \boldsymbol{U}, \mathscr{K}, \tilde{c})$, with state process $\{\Psi_t, Y_t\}$.

The history spaces $\boldsymbol{H}_t$, $t \in \mathbb{N}_0$, of $(\boldsymbol{\Psi} \times \boldsymbol{Y}, \boldsymbol{U}, \mathscr{K}, \tilde{c})$ are defined by $\boldsymbol{H}_0 := \boldsymbol{\Psi} \times \boldsymbol{Y}$, and

$$\boldsymbol{H}_{t+1} := \boldsymbol{H}_t \times \boldsymbol{U} \times \boldsymbol{\Psi} \times \boldsymbol{Y}, \quad t \in \mathbb{N}_0.$$

In other words, an element $h_t \in \boldsymbol{H}_t$ is of the form

$$h_t = (\psi_0, y_0, u_0, \psi_1, y_1, u_1, \ldots, u_{t-1}, \psi_t, y_t).$$

An *admissible strategy* or *admissible policy* π is a sequence $\{\pi_t\}_{t=0}^{\infty}$ of Borel measurable stochastic kernels on $\boldsymbol{U}$ given $\boldsymbol{H}_t$. We denote the set of admissible policies by Π. An admissible policy is called *deterministic* if π_t is a point mass, and for such a policy we identify π_t with a measurable map from $\boldsymbol{H}_t$ into $\boldsymbol{U}$. Note that the model $(\boldsymbol{\Psi}, \boldsymbol{U}, \tilde{\mathscr{K}}, \tilde{c})$, with state process $\{\Psi_t\}$ and transition kernel

$$\tilde{\mathscr{K}}(B|\psi, u) = \sum_{y \in \boldsymbol{Y}} \mathscr{K}(B, y|\psi, u), \quad B \in \mathscr{B}(\boldsymbol{\Psi})$$

is also a completely observed MDP, and is equivalent to the POMDP model. A policy $\pi \in \Pi$ is called Markov if $\pi_t(\cdot|h_t) = \pi_t(\cdot|\psi_t)$, for all $t \in \mathbb{N}_0$, and in addition it is called stationary if $t \mapsto \pi_t$ is constant. We let Π_{M}, Π_{S}, and Π_{SD} denote the sets of all Markov, stationary, and stationary deterministic policies, respectively.

For each initial distribution $\mu \in \mathscr{P}(\boldsymbol{\Psi} \times \boldsymbol{Y})$ as in (7) and $\pi \in \Pi$, there exists a unique probability measure $\mathbb{P}_\mu^\pi$ induced on the sample path of the process $(\Psi_t, Y_{t+1}, U_t,\ t \in \mathbb{N})$ [3]. We let $\mathbb{E}_\mu^\pi$ denote the associated expectation operator. We define the long-run average cost

$$J(\mu, \pi) := \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_\mu^\pi[\tilde{c}(\Psi_t, U_t)]$$

and the β-discounted cost, $\beta \in (0, 1)$,

$$J_\beta(\mu, \pi) := \sum_{t=0}^{\infty} \beta^t \mathbb{E}_\mu^\pi[\tilde{c}(\Psi_t, U_t)].$$

Minimizing $J(\mu, \pi)$ over all $\pi \in \Pi$ is equivalent to the optimization problem in (1). Without loss of generality, we may assume that $\mu \in \mathscr{P}(\boldsymbol{\Psi} \times \boldsymbol{Y})$ is a point mass concentrated at some $(\psi, y) \in \boldsymbol{\Psi} \times \boldsymbol{Y}$. Then, we may use without ambiguity the notation $\mathbb{P}_\psi^\pi$ instead of $\mathbb{P}_\mu^\pi$.

We next state the following well known characterization of β-discounted optimal policies [11, Chapter 2].

Lemma 1. *The β-discounted value function, defined by*

$$h_\beta(\psi) = \inf_{\pi \in \Pi} J_\beta(\psi, \pi) \tag{8}$$

satisfies the β-discounted cost optimality equation (*DCOE*):

$$h_\beta(\psi) = \min_{u \in \boldsymbol{U}} \left\{ \tilde{c}(\psi, u) + \beta \int_{\boldsymbol{\Psi}} h_\beta(\psi')\, \tilde{\mathscr{K}}(\mathrm{d}\psi'|\psi, u) \right\}. \tag{9}$$

Furthermore, $h_\beta(\psi)$ in (8) *is the unique solution to* (9), *in the space of bounded lower-semicontinuous functions, which is denoted by in $\mathscr{L}\mathscr{C}_b(\boldsymbol{\Psi})$, and the minimizer of* (9) *denoted by π_β is optimal, i.e., $h_\beta(\psi) = J_\beta(\psi, \pi_\beta)$.*

It is well known that h_β is concave on $\boldsymbol{\Psi}$ [12]. This property plays a crucial role in our analysis.

3. Assumptions

Various necessary conditions for the existence of a solution to the ACOE have been proposed in the literature. We present here two new conditions and compare them to the hypotheses used in [4,12,15].

Assumption 2 (*Interior accessibility*). *Define*

$$\boldsymbol{\Psi}_\varepsilon := \{\psi \in \boldsymbol{\Psi} : \psi(i) \geqslant \varepsilon,\ \forall\, i \in \boldsymbol{X}\}.$$

There exist constants $\varepsilon > 0$, $k_0 \in \mathbb{N}$ and $\beta_0 < 1$ such that if $\psi_*(\beta) \in \arg\min_{\psi \in \boldsymbol{\Psi}} h_\beta(\psi)$ then for each $\beta \in [\beta_0, 1)$ we have

$$\max_{1 \leqslant k \leqslant k_0} \mathbb{P}_{\psi_*(\beta)}^{\pi_\beta}(\Psi_k \in \boldsymbol{\Psi}_\varepsilon) \geqslant \varepsilon. \tag{10}$$

Since h_β is concave, it attains its minima on the extreme points of $\boldsymbol{\Psi}$, which is a finite set. This simplifies the verification of (10), as the examples in Section 5 show. Assumption 2 is stated in terms of the β-discounted optimal policy π_β. However, it can be replaced by a stronger condition, Assumption 4 below, which does not necessitate knowledge of π_β. First we need to introduce some notation.

Definition 3. For $y^k \in \boldsymbol{Y}^k$ and $u^k \in \boldsymbol{U}^k$, $k \in \mathbb{N}$, let $Q(y^k, u^k)$ be the expression defined in (2). Recall that $\bar{\psi}$ stands for the element in $\boldsymbol{\Psi}$, with $\bar{\psi}(i) = 1/n$, for all $i \in \boldsymbol{X}$. We use the notation

$$V(\psi, y^k, u^k) = \psi Q(y^k, u^k)\mathbf{1},$$

$$T(\psi, y^k, u^k) = \begin{cases} \dfrac{\psi Q(y^k, u^k)}{V(\psi, y^k, u^k)} & \text{if } V(\psi, y^k, u^k) \neq 0, \\ \bar{\psi} & \text{otherwise.} \end{cases}$$

Assumption 4. There exists $k_0 \in \mathbb{N}$ such that, for each $i \in \boldsymbol{X}$,

$$\max_{1 \leqslant k \leqslant k_0} \min_{u^k \in \boldsymbol{U}^k} \left\{ \max_{y^k \in \boldsymbol{Y}^k} \min_{j \in \boldsymbol{X}} Q_{ij}(y^k, u^k) \right\} > 0.$$

Perhaps a more transparent way of stating Assumption 4 is that for each $i \in \boldsymbol{X}$ and for each sequence $u^{k_0} = (u_0, \ldots, u_{k_0-1})$, there exists some $k \leqslant k_0$ and a sequence $y^k = (y_1, \ldots, y_k)$, such that $Q_{ij}(y^k, u^k) > 0$, for all $j \in \boldsymbol{X}$.

It is simple to show that (4) implies Assumption 4. Indeed, let $u^k \in \boldsymbol{U}^k$ be arbitrary and note that

$$\sum_{y^k \in \boldsymbol{Y}^k} \sum_{\ell \in \boldsymbol{X}} Q_{i\ell}(y^k, u^k) = 1$$

$$\Longrightarrow \max_{y^k \in \boldsymbol{Y}^k} \sum_{\ell \in \boldsymbol{X}} Q_{i\ell}(y^k, u^k) \geqslant \frac{1}{m^k}.$$

Hence (4) implies that

$$\max_{y^k \in \boldsymbol{Y}^k} Q_{ij}(y^k, u^k) \geqslant \frac{\Delta}{m^k} \quad \forall i, j \in \boldsymbol{X}.$$

Platzman utilizes a *reachability* condition, which amounts to the existence of $\delta_1 > 0$ and $k_0 \in \mathbb{N}$ such that

$$\sup_{\pi \in \Pi} \max_{0 \leqslant k \leqslant k_0} \mathbb{E}^{\pi}_{\mu}[\Psi_k(j)] \geqslant \delta_1 \quad \forall j \in \boldsymbol{X}, \ \forall \psi \in \boldsymbol{\Psi}, \tag{11}$$

together with a *detectability* condition that can be stated as follows: First define a metric D on $\boldsymbol{\Psi}$ by

$$D(\psi, \psi') = \max\{\mathrm{d}(\psi, \psi'), \mathrm{d}(\psi', \psi)\},$$

where

$$\mathrm{d}(\psi, \psi') = 1 - \min\left\{\frac{\psi(i)}{\psi'(i)} : \psi'(i) > 0\right\}.$$

Let e^i denote the row vector with the ith component equal to 1 and all other components equal to 0. If for an $n \times n$ substochastic matrix Q we set

$$\alpha(Q) := \max_{i, i' \in \boldsymbol{X}} \left\{ D\left(\frac{e^i Q}{e^i Q\mathbf{1}}, \frac{e^{i'} Q}{e^{i'} Q\mathbf{1}}\right) : e^i Q \neq 0, e^{i'} Q \neq 0 \right\},$$

then Platzman's detectability condition asserts that there is constant $\delta_2 < 1$ and $k \in \mathbb{N}$, such that

$$\mathbb{E}^{\pi}_{\psi}[\alpha(Q(Y^k, U^k))] \leqslant \delta_2 \quad \forall \psi \in \boldsymbol{\Psi} \ \forall \pi \in \Pi. \tag{12}$$

For a substochastic matrix $Q \neq 0$, $\alpha(Q) < 1$ if and only if Q is subrectangular, and this fact is utilized in the proofs of the results in [12]. Hence Assumption 4 does not imply (12). Also, as mentioned earlier the action space is assumed finite in [12].

Runggaldier and Stettner in [15], specialize the model to the case where $\{X_t\}$ is a controlled Markov chain with transition kernel P on $\boldsymbol{X}$ given $\boldsymbol{X} \times \boldsymbol{U}$, and the observations are governed by a kernel G on $\boldsymbol{Y}$ given $\boldsymbol{X}$. Representing P as a $n \times n$ matrix $P(u)$, $u \in \boldsymbol{U}$, and defining

$$\mathrm{O}(y) = \mathrm{diag}(G(y|1), \ldots, G(y|n)) \in \mathbb{R}^{n \times n},$$

then the transition kernel Q of the partially observed model takes the form $Q(y, u) = P(u)\mathrm{O}(y)$, $u \in \boldsymbol{U}$, $y \in \boldsymbol{Y}$. The following positivity condition is imposed in [15]:

$$\min_{i, j \in \boldsymbol{X}} \inf_{u, u' \in \boldsymbol{U}} \min_{\{k \in \boldsymbol{X} : [P(u)]_{ik} > 0\}} \frac{[P(u')]_{jk}}{[P(u)]_{ik}} > 0. \tag{13}$$

Note that (13) does not imply (11). However, (13) is stronger than (12).

Provided the running cost c is continuous, we can weaken Assumption 2 as follows.

Assumption 5 (*Relative interior accessibility*). The running cost c is continuous, and there exist $\varepsilon > 0$, $k_0 \in \mathbb{N}$, and $\beta_0 < 1$ such that for each β in $[\beta_0, 1)$ we have

$$\begin{aligned} &\max_{1 \leqslant k \leqslant k_0} \mathbb{P}^{\pi_\beta}_{\psi_*}(T(\psi_*, Y^k, U^k) \geqslant \varepsilon T(\psi^*, Y^k, U^k), \\ &V(\psi^*, Y^k, U^k) \geqslant \varepsilon V(\psi_*, Y^k, U^k)) \geqslant \varepsilon, \end{aligned} \tag{14}$$

where $\psi_* = \psi_*(\beta)$ and $\psi^* = \psi^*(\beta)$ are any pair of points in $\boldsymbol{\Psi}$ at which h_β attains its minimum and maximum, respectively.

Since ψ^* is not known, nor does it admit a finite set of values like ψ_*, we also state a condition that is independent of ψ^*.

Assumption 6. There exist constants $\varepsilon > 0$, $k_0 \in \mathbb{N}$ and $\beta_0 < 1$ such that for each β in $[\beta_0, 1)$ and all $\ell \in \boldsymbol{X}$

$$\begin{aligned} &\max_{1 \leqslant k \leqslant k_0} \mathbb{P}^{\pi_\beta}_{e^\ell}\Bigg(\frac{1}{\varepsilon} \sum_{j' \in \boldsymbol{X}} Q(Y^k, U^{k-1})_{ij'} \geqslant Q(Y^k, U^{k-1})_{\ell j} \\ &\quad \geqslant \varepsilon Q(Y^k, U^{k-1})_{ij}, \forall i, j \in \boldsymbol{X}\Bigg) \geqslant \varepsilon. \end{aligned} \tag{15}$$

Clearly, Assumption 6 implies Assumption 5. Note however that even if we require Assumption 6 to hold over all $\pi \in \Pi_{\mathrm{SD}}$, this does not imply detectability, i.e., condition (12), since (15) can hold even if the matrices involved are not subrectangular. Observe also that (15) follows from (5).

4. Main results

The results are stated in two theorems and two lemmas. Theorem 7 shows that existence of a bounded solution to the ACOE is equivalent to the uniform boundedness of the differential discounted value functions, which are defined by

$$\bar{h}_\beta(\psi) := h_\beta(\psi) - \inf_{\psi' \in \boldsymbol{\Psi}} h_\beta(\psi').$$

Lemmas 9 and 10 establish that under Assumptions 2 and 5, respectively, the family $\{\bar{h}_\beta : \beta \in (0,1)\}$ is uniformly bounded. Finally, Theorem 11 asserts the existence of stationary optimal policies.

Theorem 7. *There exists a solution* (ϱ, h), *with* $\varrho \in \mathbb{R}$ *and* $h : \boldsymbol{\Psi} \to \mathbb{R}_+$, *a bounded function, to the ACOE*

$$\varrho + h(\psi) = \min_{u \in \boldsymbol{U}} \left\{ \tilde{c}(\psi, u) + \int_{\boldsymbol{\Psi}} h(\psi') \tilde{\mathscr{K}}(\mathrm{d}\psi' | \psi, u) \right\} \tag{16}$$

if and only if $\{\bar{h}_\beta : \beta \in (0,1)\}$ *is uniformly bounded.*

Proof. We use the notation

$$\|c\|_\infty := \max_{x \in \boldsymbol{X}, u \in \boldsymbol{U}} c(x,u),$$
$$\operatorname{span}(h_\beta) := \sup_{\psi' \in \boldsymbol{\Psi}} h_\beta(\psi') - \inf_{\psi' \in \boldsymbol{\Psi}} h_\beta(\psi').$$

Necessity is standard and well known, since, as shown in [10], if h is a bounded solution to the ACOE then $\operatorname{span}(h_\beta) \leqslant 2\operatorname{span}(h)$ for all $\beta \in (0,1)$. To show sufficiency, suppose that for some constant $M_0 \in \mathbb{R}$, $\operatorname{span}(h_\beta) \leqslant M_0$ for all $\beta \in (0,1)$. Write (9) as

$$\begin{aligned}(1-\beta)h_\beta(\psi_*) + \bar{h}_\beta(\psi) \\ = \inf_{u \in \boldsymbol{U}} \left\{ \tilde{c}(\psi, u) + \beta \int_{\boldsymbol{\Psi}} \bar{h}_\beta(\psi') \, \tilde{\mathscr{K}}(\mathrm{d}\psi' | \psi, u) \right\}.\end{aligned} \tag{17}$$

Let $(h_\beta)_* := \inf_{\psi' \in \boldsymbol{\Psi}} h_\beta(\psi')$. Since $0 \leqslant h_\beta(\psi) \leqslant 1/(1-\beta)\|c\|_\infty$, for all $\psi \in \boldsymbol{\Psi}$, it follows that the set $\{(1-\beta)(h_\beta)_*, \ \beta \in (0,1)\}$ is uniformly bounded. By the Bolzano–Weierstrass Theorem, along some sequence $\{\beta_n\}_{n=0}^\infty$, tending to 1,

$$(1-\beta_n)(h_\beta)_* \longrightarrow \varrho \quad \text{as } n \to \infty$$

for some $\varrho \in \mathbb{R}$. Since the family $\{\bar{h}_\beta\}$ is concave and bounded, it is equi-Lipschitzian on each compact subset of the relative interior of each facet of $\boldsymbol{\Psi}$ (see [13, p. 88, Theorem. 10.6]). Since $\boldsymbol{\Psi}$ has finitely many facets, using Ascoli's theorem, we conclude that $\{\bar{h}_\beta\}$ converges pointwise along some subsequence $\{\beta'_n\}$ of $\{\beta_n\}$ to some function h. Define

$$F_\beta(\psi, u) := \tilde{c}(\psi, u) + \beta \int_{\boldsymbol{\Psi}} \bar{h}_\beta(\psi') \tilde{\mathscr{K}}(\mathrm{d}\psi' | \psi, u), \quad \beta < 1,$$

$$F_1(\psi, u) := \tilde{c}(\psi, u) + \int_{\boldsymbol{\Psi}} h(\psi') \tilde{\mathscr{K}}(\mathrm{d}\psi' | \psi, u).$$

Taking limits as $\beta'_n \to 1$ in (17) we obtain

$$\begin{aligned}\varrho + h(\psi) &= \lim_{n \to \infty} \inf_{u \in \boldsymbol{U}} F_{\beta'_n}(\psi, u) \\ &\leqslant \inf_{u \in \boldsymbol{U}} \lim_{n \to \infty} F_{\beta'_n}(\psi, u) = \inf_{u \in \boldsymbol{U}} F_1(\psi, u).\end{aligned} \tag{18}$$

Since h, being concave, is lower-semicontinuous, the 'inf' in the last line of (18) can be replaced by a 'min'. Fix $\hat{\psi} \in \boldsymbol{\Psi}$, and let $\hat{u}_n \in \arg\min_u \{F_{\beta'_n}(\hat{\psi}, u)\}$, $n = 1, 2, \ldots$, be an arbitrary sequence. Extract a convergent subsequence $\{u_n^*\} \subset \{\hat{u}_n\}$ and let u_∞^* denote its limit point. Also, let $\{\beta''_n\}$ denote the corresponding subsequence of $\{\beta'_n\}$, i.e.,

$$u_n^* \in \arg\min_u \{F_{\beta''_n}(\hat{\psi}, u)\}.$$

Since $T(\hat{\psi}, y, u)$ and $V(\hat{\psi}, y, u)$ are continuous in u, then given $\delta > 0$, we can select $n_\delta \in \mathbb{N}$, such that, on the set $\{y \in \boldsymbol{Y} : V(\hat{\psi}, y, u_\infty^*) \neq 0\}$,

$$T(\hat{\psi}, y, u_n^*) \geqslant (1-\delta) T(\hat{\psi}, y, u_\infty^*), \tag{19a}$$

$$V(\hat{\psi}, y, u_n^*) \geqslant (1-\delta) V(\hat{\psi}, y, u_\infty^*) \tag{19b}$$

for all $n \geqslant n_\delta$, and simultaneously, by the lower semicontinuity of $u \mapsto \tilde{c}(\hat{\psi}, u)$,

$$\tilde{c}(\hat{\psi}, u_n^*) \geqslant (1-\delta) \tilde{c}(\hat{\psi}, u_\infty^*) \quad \forall n \geqslant n_\delta. \tag{20}$$

By (19a)

$$\frac{1}{\delta} [T(\hat{\psi}, y, u_n^*) - (1-\delta) T(\hat{\psi}, y, u_\infty^*)] \in \boldsymbol{\Psi}.$$

Hence using the concavity of $\bar{h}_\beta$ we obtain

$$\begin{aligned}\bar{h}_\beta(T(\hat{\psi}, y, u_n^*)) &\geqslant \bar{h}_\beta(T(\hat{\psi}, y, u_\infty^*)) - \delta \operatorname{span}(\bar{h}_\beta) \\ &\geqslant \bar{h}_\beta(T(\hat{\psi}, y, u_\infty^*)) - \delta M_0\end{aligned} \tag{21}$$

for all $\beta \in (\beta_0, 1)$. Multiplying both sides of (21) by $V(\hat{\psi}, y, u_n^*)$, using (19b) to strengthen the inequality, summing over $y \in \boldsymbol{Y}$, and evaluating along the subsequence $\{\beta''_n\}$, we obtain,

$$\begin{aligned}&\int_{\boldsymbol{\Psi}} \bar{h}_{\beta''_n}(\psi') \tilde{\mathscr{K}}(\mathrm{d}\psi' | \hat{\psi}, u_n^*) \\ &\quad \geqslant (1-\delta) \int_{\boldsymbol{\Psi}} \bar{h}_{\beta''_n}(\psi') \tilde{\mathscr{K}}(\mathrm{d}\psi' | \hat{\psi}, u_\infty^*) - \delta M_0.\end{aligned} \tag{22}$$

Adding (20) and (22) yields

$$F_{\beta''_n}(\hat{\psi}, u_n^*) \geqslant (1-\delta) F_{\beta''_n}(\hat{\psi}, u_\infty^*) - \delta M_0.$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty}\inf_{u\in\boldsymbol{U}} F_{\beta''_n}(\hat{\psi},u) &= \lim_{n\to\infty} F_{\beta''_n}(\hat{\psi},u_n^*)\\ &\geqslant (1-\delta)\lim_{n\to\infty} F_{\beta''_n}(\hat{\psi},u_\infty^*) - \delta M_0\\ &= (1-\delta)\ F_1(\hat{\psi},u_\infty^*) - \delta M_0\\ &\geqslant (1-\delta)\inf_{u\in\boldsymbol{U}} F_1(\hat{\psi},u) - \delta M_0.\end{aligned} \tag{23}$$

Since (23), holds for arbitrary $\hat{\psi}\in\boldsymbol{\Psi}$ and $\delta>0$, then together with (18) we deduce

$$\varrho + h(\psi) = \min_{u\in\boldsymbol{U}} F_1(\psi,u),$$

thus obtaining (16). □

Remark 8. Since the discounted value functions are concave, Theorem 7 asserts that if the ACOE admits a bounded solution, then it admits a concave one as well.

Lemma 9. *Under Assumption* 2, $\{\bar{h}_\beta, \beta\in[\beta_0,1)\}$ *is uniformly bounded on* $\boldsymbol{\Psi}$.

Proof. Dropping the dependence on β, in order to simplify the notation, let ψ_* be a point in $\boldsymbol{\Psi}$ at which h_β attains its minimum. By (9), for each positive integer $k\in\mathbb{N}$, we have

$$\begin{aligned}h_\beta(\psi_*) &= \mathbb{E}^{\pi_\beta}_{\psi_*}\left[\sum_{t=0}^{k-1}\beta^t\tilde{c}(\Psi_t,U_t)+\beta^k h_\beta(\Psi_k)\right]\\ &\geqslant \beta^k\mathbb{E}^{\pi_\beta}_{\psi_*}[h_\beta(\Psi_k)].\end{aligned} \tag{24}$$

By (24),

$$\begin{aligned}\operatorname{span}(h_\beta) &\leqslant \sup_{\psi\in\boldsymbol{\Psi}} h_\beta(\psi) - \beta^k\mathbb{E}^{\pi_\beta}_{\psi_*}[h_\beta(\Psi_k)]\\ &= (1-\beta^k)\sup_{\psi\in\boldsymbol{\Psi}} h_\beta(\psi) + \beta^k\mathbb{E}^{\pi_\beta}_{\psi_*}\\ &\quad\times\left[\sup_{\psi\in\boldsymbol{\Psi}} h_\beta(\psi) - h_\beta(\Psi_k)\right].\end{aligned} \tag{25}$$

Note that if $\psi'\in\boldsymbol{\Psi}_\varepsilon$ then

$\tilde{\psi} := \frac{1}{1-\varepsilon}(\psi'-\varepsilon\psi)\in\boldsymbol{\Psi} \quad \forall\psi\in\boldsymbol{\Psi}.$

Hence, since h_β is concave,

$$\begin{aligned}h_\beta(\psi)-h_\beta(\psi') &\leqslant (1-\varepsilon)[h_\beta(\psi)-h_\beta(\tilde{\psi})]\\ &\leqslant (1-\varepsilon)\operatorname{span}(h_\beta)\end{aligned} \tag{26}$$

for all $\psi\in\boldsymbol{\Psi}$ and $\psi'\in\boldsymbol{\Psi}_\varepsilon$. Fix $\beta\in[\beta_0,1)$. By Assumption 2, there exists $k'\leqslant k_0$ such that $\mathbb{P}^{\pi_\beta}_{\psi_*(\beta)}(\Psi_{k'}\in\boldsymbol{\Psi}_\varepsilon)\geqslant\varepsilon$. Therefore, by (26),

$$\begin{aligned}&\mathbb{E}^{\pi_\beta}_{\psi_*}\left[\sup_{\psi\in\boldsymbol{\Psi}} h_\beta(\psi) - h_\beta(\Psi_{k'})\right]\\ &\quad\leqslant \mathbb{P}^{\pi_\beta}_{\psi_*(\beta)}(\Psi_{k'}\in\boldsymbol{\Psi}_\varepsilon)(1-\varepsilon)\operatorname{span}(h_\beta)\\ &\qquad + (1-\mathbb{P}^{\pi_\beta}_{\psi_*(\beta)}(\Psi_{k'}\in\boldsymbol{\Psi}_\varepsilon))(1-\varepsilon)\operatorname{span}(h_\beta)\\ &\quad\leqslant (1-\varepsilon^2)\operatorname{span}(h_\beta).\end{aligned} \tag{27}$$

Thus, by (25) and (27),

$$\begin{aligned}\operatorname{span}(h_\beta) &\leqslant (1-\beta^{k'})\sup_{\psi\in\boldsymbol{\Psi}} h_\beta(\psi)\\ &\quad + (1-\varepsilon^2)\operatorname{span}(h_\beta).\end{aligned} \tag{28}$$

Since

$$\begin{aligned}(1-\beta^{k'})\sup_{\psi\in\boldsymbol{\Psi}} h_\beta(\psi) &\leqslant (1+\beta+\cdots+\beta^{k'-1})\|c\|_\infty\\ &\leqslant k_0\|c\|_\infty,\end{aligned}$$

inequality (28) yields

$$\operatorname{span}(h_\beta)\leqslant\frac{k_0}{\varepsilon^2}\|c\|_\infty \quad \forall\beta\in[\beta_0,1),$$

and the proof is complete. □

Note that every pair $(\pi,\psi)\in\Pi_{\mathrm{SD}}\times\boldsymbol{\Psi}$ induces in a natural manner a non-stationary policy $\Phi(\pi,\psi)=\{\varphi_t : t\in\mathbb{N}_0\}$, where each φ_t is a measurable map from $\boldsymbol{H}_t$ into $\boldsymbol{U}$, as follows: For $t=0$, set $\psi_0=\psi$ and $\varphi_0=\pi(\psi_0)$, and define inductively

$$\begin{aligned}&\psi_t = T(\psi_{t-1},y_t,\pi(\psi_{t-1})),\\ &\varphi_t = \pi(\psi_t).\end{aligned}$$

Thus, with (π,ψ) viewed as a parameter, $\varphi_t=\varphi_t(y_1,\ldots,y_t)$ is a measurable map from $\boldsymbol{Y}^t$ into $\boldsymbol{U}$.

Lemma 10. *Under Assumption* 5, $\{\bar{h}_\beta, \beta\in[\beta_0,1)\}$ *is uniformly bounded on* $\boldsymbol{\Psi}$.

Proof. Fix $k\in\mathbb{N}$ at the value that attains the maximum in (14). Let $y^k\mapsto u_*^k(y^k)$ be the map from the sequence of observations $y^k\in\boldsymbol{Y}^k$ to the sequence of actions of the policy $\Phi(\pi_\beta,\psi_*)$, which was defined in the paragraph preceding Lemma 10. In order to simplify the notation define

$$\begin{aligned}&V_*(y^k) := V(\psi_*,y^k,u_*^k(y^k)),\\ &V^*(y^k) := V(\psi^*,y^k,u_*^k(y^k)),\end{aligned}$$

and analogously, $T_*(y^k)$ and $T^*(y^k)$. Since $\Phi(\pi_\beta,\psi_*)$ is suboptimal relative to the initial distribution ψ^* and optimal relative to ψ_*, we obtain from (9)

$$h_\beta(\psi^*) \leqslant k_0\|c\|_\infty + \beta^k \sum_{y^k\in \boldsymbol{Y}^k} V^*(y^k)\, h_\beta(T^*(y^k)), \tag{29a}$$

$$h_\beta(\psi_*) \geqslant \beta^k \sum_{y^k\in\boldsymbol{Y}^k} V_*(y^k)\, h_\beta(T_*(y^k)). \tag{29b}$$

Let

$$\mathscr{Y} := \{y^k \in \boldsymbol{Y}^k : T_*(y^k) \geqslant \varepsilon T^*(y^k),\ V^*(y^k)\geqslant \varepsilon V_*(y^k)\}.$$

Observe that by the definition of V_*

$$\mathbb{P}^{\pi_\beta}_{\psi_*}(Y^k = y^k) = V_*(y^k).$$

Hence, by Assumption 5,

$$\sum_{y^k\in\mathscr{Y}} V_*(y^k)\geqslant \varepsilon. \tag{30}$$

Also, if $y^k\in\mathscr{Y}$ then

$$\tilde\psi(y^k) := \frac{T_*(y^k)-\varepsilon T^*(y^k)}{1-\varepsilon} \in \boldsymbol{\Psi}.$$

Decomposing the summation in (29b) over the sets $\mathscr{Y}$ and $\mathscr{Y}^c = \boldsymbol{Y}^k\setminus\mathscr{Y}$, then replacing $T_*(y^k)$ by $\varepsilon T^*(y^k)+(1-\varepsilon)\tilde\psi(y^k)$ in those terms that $y^k\in\mathscr{Y}$, and using convexity, we obtain

$$\begin{aligned} h_\beta(\psi_*) \geqslant{}& \beta^k \sum_{y^k\in\mathscr{Y}^c} V_*(y^k)\, h_\beta(T_*(y^k)) \\ &+ \beta^k \sum_{y^k\in\mathscr{Y}} V_*(y^k)[\varepsilon h_\beta(T^*(y^k)) \\ &+ (1-\varepsilon) h_\beta(\tilde\psi(y^k))]. \end{aligned} \tag{31}$$

Subtracting (31) from (29a), we get

$$\begin{aligned} &\operatorname{span}(h_\beta) \\ &\quad\leqslant k_0\|c\|_\infty + \beta^k \sum_{y^k\in\mathscr{Y}^c} V^*(y^k)\, h_\beta(T^*(y^k)) \\ &\qquad - \beta^k \sum_{y^k\in\mathscr{Y}^c} V_*(y^k)\, h_\beta(T_*(y^k)) \\ &\qquad + \beta^k \sum_{y^k\in\mathscr{Y}} [V^*(y^k)-\varepsilon V_*(y^k)]\, h_\beta(T^*(y^k)) \\ &\qquad - \beta^k \sum_{y^k\in\mathscr{Y}} (1-\varepsilon) V_*(y^k)\, h_\beta(\tilde\psi(y^k)). \end{aligned} \tag{32}$$

Using the identity

$$\sum_{y^k\in\boldsymbol{Y}^k} V_*(y^k) = \sum_{y^k\in\boldsymbol{Y}^k} V^*(y^k),$$

to add and subtract terms as needed, (32) is equivalent to

$$\begin{aligned} &\operatorname{span}(h_\beta) \\ &\quad\leqslant k_0\|c\|_\infty + \beta^k \sum_{y^k\in\mathscr{Y}^c} V^*(y^k)[h_\beta(T^*(y^k)) - h_\beta(\psi_*)] \\ &\qquad - \beta^k \sum_{y^k\in\mathscr{Y}^c} V_*(y^k)[h_\beta(T_*(y^k)) - h_\beta(\psi_*)] \\ &\qquad + \beta^k \sum_{y^k\in\mathscr{Y}} [V^*(y^k)-\varepsilon V_*(y^k)][h_\beta(T^*(y^k)) \\ &\qquad - h_\beta(\psi_*)] - \beta^k \sum_{y^k\in\mathscr{Y}} (1-\varepsilon) V_*(y^k) \\ &\qquad \times [h_\beta(\tilde\psi(y^k)) - h_\beta(\psi_*)]. \end{aligned} \tag{33}$$

Strengthening the inequality in (33) by discarding the third and the fifth terms on the right-hand side which are negative, and then evaluating at $\beta=1$, we obtain

$$\begin{aligned} \operatorname{span}(h_\beta) &\leqslant k_0\|c\|_\infty + \sum_{y^k\in\mathscr{Y}^c} V^*(y^k)\operatorname{span}(h_\beta) \\ &\quad + \sum_{y^k\in\mathscr{Y}} [V^*(y^k)-\varepsilon V_*(y^k)]\operatorname{span}(h_\beta) \\ &= k_0\|c\|_\infty + \sum_{y^k\in\boldsymbol{Y}^k} V^*(y^k)\operatorname{span}(h_\beta) \\ &\quad - \varepsilon \sum_{y^k\in\mathscr{Y}} V_*(y^k)\operatorname{span}(h_\beta) \\ &\leqslant k_0\|c\|_\infty + (1-\varepsilon^2)\operatorname{span}(h_\beta), \end{aligned}$$

where the last inequality is due to (30). Therefore,

$$\operatorname{span}(h_\beta) \leqslant \frac{k_0}{\varepsilon^2}\|c\|_\infty \quad \forall \beta\in(\beta_0,1),$$

and the proof is complete. □

Theorem 11. *Under either Assumptions* 2 *or* 5, *there exist a constant* ϱ *and a concave function* $h:\boldsymbol{\Psi}\to\mathbb{R}_+$ *such that* (ϱ,h) *is a solution of* (16). *Furthermore*, $\varrho=\inf_{\pi\in\Pi} J(\mu,\pi)$. *If* $\pi^*:\boldsymbol{\Psi}\to\boldsymbol{U}$ *is a measurable selector of the minimizer in* (16), *then* $\pi^*\in\Pi_{\mathrm{SD}}$ *is optimal, i.e.,* $J(\mu,\pi^*)=\varrho$, *for all initial distributions* μ.

Proof. Using Lemmas 9 and 10 and Theorem 7, we conclude that there exists a solution (ϱ,h) to (16), with h a concave function. Then, since the function inside the minimum in (16) is lower-semicontinuous, there exists a measurable map $\pi^*:\boldsymbol{\Psi}\to\boldsymbol{U}$, which is a selector from the set-valued minimizer. Hence $\pi^*\in\Pi_{\mathrm{SD}}$. Optimality of π^* follows as in [11, Chapter 3]. □

5. Examples

We examine the well known machine replacement problem to some detail. This example fails to fulfill the positivity assumption in [15] and the detectability condition in [12]. However, as we show it satisfies Assumption 2. The description of this problem is the following.

Example 12. Consider a system with state space $\boldsymbol{X}=\{0,1\}$ where 0 is the 'good' state and 1 is the 'down' state. The action space is $\boldsymbol{U}=\{0,1\}$ where 0 means to operate the system and 1 to repair the system. If the system is in state 0 and is operated then it fails with some positive probability. If the system is in state 1, it stays in that state, unless it is repaired, in which case the system moves to state 0 with a high probability (but not with certainty). Therefore, the state transition probability $P(u)$ of the system is specified by

$$P(0)=\begin{bmatrix}1-\eta & \eta\\ 0 & 1\end{bmatrix},\quad P(1)=\begin{bmatrix}1 & 0\\ \alpha & 1-\alpha\end{bmatrix},$$

where $\eta\in(0,1)$ is the one-step probability of failure and $\alpha\in(0.5,1]$ is the probability of success of the repair operation. Suppose the correct observation rate is $q\in(0.5,1)$, i.e., $\boldsymbol{Y}=\{0,1\}$, and the observations evolve according to a kernel $G(y|x)$, with $G(y|x)=q$, if $x=y$, and $G(y|x)=1-q$, if $x\neq y$. If we define

$$\mathrm{O}(0):=\begin{bmatrix}q & 0\\ 0 & 1-q\end{bmatrix},\quad \mathrm{O}(1):=\begin{bmatrix}1-q & 0\\ 0 & q\end{bmatrix},$$

then the transition kernel Q of the partially observed system is given by

$$Q(y,u)=P(u)\mathrm{O}(y),\quad u,y\in\{0,1\}.$$

The cost function $c(x,u)$ is given by $c(0,0)=0$, $c(1,0)=C$, and $c(j,1)=R$, $j\in\boldsymbol{X}$, satisfying $0<C<R$.

If ψ is represented by $[1-pp]$ where $p\in[0,1]$ is the probability that the system is down, then the β-discounted value function h_β satisfies:

$$h_\beta(\psi) = \min\left\{Cp+\beta\sum_{y\in\boldsymbol{Y}}V(\psi,y,0)h_\beta(T(\psi,y,0)),\right.\\ \left. R+\beta\sum_{y\in\boldsymbol{Y}}V(\psi,y,1)h_\beta(T(\psi,y,1))\right\}.$$

We next show that this example satisfies Assumption 2.

Theorem 13. *In Example* 12, *the β-discounted optimal policy satisfies*

(i) $\pi_\beta(\psi_*)=0$.
(ii) $\psi_*=[1\,0]$.

Proof. To show (i) we argue by contradiction. Suppose that $\pi_\beta(\psi_*)=1$. Then

$$h_\beta(\psi_*)=R+\beta\sum_{y=0}^{1}V(\psi_*,y,1)h_\beta(T(\psi_*,y,1))\\ \geqslant R+\beta h_\beta(\psi_*).$$

Thus $h_\beta(\psi_*)\geqslant R/(1-\beta)$. Consider the policy $\hat{\pi}=\{\pi_t\}_{t=0}^\infty$, with $\pi_t=0$ for all t. Under this policy, the machine is never repaired so the incurred value function $h_\beta^{\hat{\pi}}(\psi_*)$ does not exceed $\sum_{t=0}^\infty(\beta^tC)=C/(1-\beta)$. This contradicts the optimality of π_β, and hence $\pi_\beta(\psi_*)=0$.

Next we show (ii). Since h_β is convex, the only candidates for the minimizer ψ_* are [0 1] and [1 0]. Suppose $\psi_*=[0\,1]$. By (i) $\pi_\beta(\psi_*)=0$, and therefore, if the initial distribution is ψ_* the machine stays in the down state and the incurred cost is $h_\beta(\psi_*)=C/(1-\beta)$. However, if the initial distribution is [1 0], then under the policy $\hat{\pi}$ defined earlier, the incurred cost is at most $\beta C/(1-\beta)$, which is less than $h_\beta(\psi_*)$. This leads to contradiction and proves that $\psi_*=[1\,0]$. □

Using Theorem 13, if

$$\varepsilon=\min\{(1-q)\eta,(1-q)(1-\eta)\},$$

then

$$\mathbb{P}_{\psi_*}^{\pi_\beta}(\Psi_1(j)\geqslant\varepsilon,\forall j\in\boldsymbol{X})=1\quad\forall\beta\in(0,1).$$

Consequently, Assumption 2 is satisfied. However, due to the zeros in the transition kernels, this example does not satisfy the positivity assumption proposed in [15]. Similarly Platzman's detectability condition is not satisfied, since this condition needs to be met by *all* admissible policies, and the policy $\hat{\pi}$ used in the proof of Theorem 13 provides a counterexample.

Now we study a modified version of Example 12 to compare some of the assumptions discussed in Section 3.

Example 14. The state space is $\boldsymbol{X}=\{0,1,2\}$, and the states 0, 1 and 2 are interpreted as good, *in need of maintenance*, and *down*, respectively. The action space is the set {0, 1} where 0 means *operate* and 1 *repair*. Assume that the running cost satisfies

$$0\leqslant c(0,0)<c(1,0)<c(2,0)<c(j,1)<\infty\quad\forall j\in\boldsymbol{X}.$$

Operating the machine causes it to deteriorate statistically over time, and when the repair action is chosen, the machine's state may be improved. This is reflected in the state transition probabilities, which are selected as follows

$$P(0)=\begin{bmatrix}\theta_1 & \theta_2 & 1-\theta_1-\theta_2\\ 0 & \theta_3 & 1-\theta_3\\ 0 & 0 & 1\end{bmatrix},\quad P(1)=\begin{bmatrix}1 & 0 & 0\\ \theta_4 & 1-\theta_4 & 0\\ \theta_5 & 1-\theta_5 & 0\end{bmatrix}.$$

We assume that the parameters θ_i are non-zero. The observation kernel takes the general form

$$O(y) = \begin{bmatrix} q_{1y} & 0 & 0 \\ 0 & q_{2y} & 0 \\ 0 & 0 & q_{3y} \end{bmatrix},$$

with $\sum_{y \in \boldsymbol{Y}} q_{iy} = 1$ for each $i \in \boldsymbol{X}$.

Using arguments similar to those in Example 12 we can show that $\pi_\beta(\psi_*) = 0$ where $\psi_* = \operatorname{argmin}_{\psi \in \boldsymbol{\Psi}} h_\beta(\psi)$ and $\psi_* = [1\,0\,0]$ or $[0\,1\,0]$ (this depends on the transition parameters and the cost function). When $\psi_* = [1\,0\,0]$, we distinguish the following two cases:

(C1) There exists an observation $y \in \boldsymbol{Y}$ such that q_{1y}, q_{2y}, and q_{3y} are all positive.

(C2) The observation y=2 identifies the state with certainty, i.e.,

$$O(2) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

while the other two observation kernels are given by

$$O(0) = \begin{bmatrix} q & 0 & 0 \\ 0 & 1-q & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad O(1) = \begin{bmatrix} 1-q & 0 & 0 \\ 0 & q & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where we assume $q \in (.5, 1)$.

In (C1) Assumption 2 is satisfied but the *renewability* condition in [12] fails. In (C2) the information state $[0\,0\,1]$ is recurrent, and thus the renewability condition in [12] is satisfied. However, under any policy, the information state ψ_t lies on the boundary of $\boldsymbol{\Psi}$ for all $t \geqslant 1$, and hence Assumption 2 fails. Finally, we note that in both cases, neither the positivity assumption in [15] nor the detectability condition in [12] is satisfied due to the appearance of zeros in the transition kernels, but Assumption 5 is met as can be established by arguments similar to those used for Example 12.

Example 15. Consider Example 12 with η=0 and α=1. It is fairly simple to verify that the policy $\pi^*(\psi) = 0$, if $\psi = [1\,0]$, and $\pi^*(\psi) = 1$, otherwise, is average cost optimal (and also β-discounted optimal for β sufficiently large). Note then that Assumption 6 holds. Furthermore, the pair (ϱ, h) with $\varrho = 0$ and $h(\psi) = 0$, if $\psi = [1\,0]$, and $h(\psi) = R$, otherwise, is a solution of (16). Since h is discontinuous Assumption 3.2 of [6] fails. This can also be directly verified by considering a pair of chains with initial laws $[1\,0]$ and $[0\,1]$, respectively, governed by the policy $\hat{\pi}$, defined in the proof of Theorem 13.

Acknowledgements

This research was supported in part by the National Science Foundation under the grants ECS-0218207 and ECS-02225448, and in part by the Office of Naval Research through the Electric Ship Research and Development Consortium.

References

[1] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, S.I. Marcus, Discrete-time controlled Markov processes with average cost criterion: A survey, SIAM J. Control Optim. 31 (3) (1993) 282–344.

[2] R. Bellman, Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957.

[3] D.P. Bertsekas, S.E. Shreve, Stochastic Optimal Control: The Discrete Time Case, Academic Press, New York, 1978.

[4] V.S. Borkar, Ergodic control of partially observed Markov chains, Systems Control Lett. 34 (4) (1998) 185–189.

[5] V.S. Borkar, Erratum to: Ergodic control of partially observed Markov chains [Systems Control Lett. 34(4) (1998) 185–189], Systems Control Lett. 37(3) (1999) 181.

[6] V.S. Borkar, Average cost dynamic programming equations for controlled Markov chains with partial observations, SIAM J. Control Optim. 39 (3) (2000) 673–681.

[7] V.S. Borkar, Dynamic programming for ergodic control with partial observations, Stochastic Process. Appl. 103 (2003) 293–310.

[8] V.S. Borkar, A. Budhiraja, A further remark on dynamic programming for partially observed Markov decision processes, Stochastic Process. Appl. 112 (2004) 79–93.

[9] E.B. Dynkin, A.A. Yushkevich, Controlled Markov Processes, Grundlehren der mathematischen Wissenschaften, vol. 235, Springer, New York, 1979.

[10] E. Fernández-Gaucherand, A. Arapostathis, S.I. Marcus, Remarks on the existence of solutions to the average cost optimality equation in Markov decision processes, Systems Control Lett. 15 (1990) 425–432.

[11] O. Hernández-Lerma, Adaptive Markov Control Processes, Applied Mathematical Sciences, vol. 79, Springer, New York, 1989.

[12] L.K. Platzman, Optimal infinite-horizon undiscounted control of finite probabilistic systems, SIAM J. Control Optim. 18 (1980) 362–380.

[13] R.T. Rockafellar, Convex Analysis, Princeton Mathematical Series, vol. 28, Princeton University Press, Princeton, NJ, 1946.

[14] S.M. Ross, Arbitrary state Markovian decision processes, Ann. Math. Statist. 6 (1968) 2118–2122.

[15] W.J. Runggaldier, L. Stettner, Approximations of discrete time partially observed control problems, Applied Mathematics Monographs, no. 6, Giardini Editori E Stampatori in Pisa, 1994.