Control and Cybernetics

VOL. 20 (1991) No. 4

Optimality of the NVI adaptive policy for a partially observed Markov decision model[1]

by

Enrique L. Sernik

Antenna Controls Department Electrospace Systems, Inc.
Richardson, Texas 75083
U.S.A

Steven I. Marcus

Systems Research Center & Electrical Engineering Department
University of Maryland
College Park, Maryland 20742
U.S.A.

We give, for a partially observed Markovian replacement model, sufficient conditions for the NVI adaptive policy to be asymptotically optimal when the performance index is the long-run average cost. We follow the approach in [1], and extend those results to the problem under study using some results obtained in [6] and [2] for the replacement problem.

[1] Research completed while the authors were with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78712. This research was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-86-0029, in part by the National Science Foundation under Grant ECS-8617860, in part by the Advanced Technology Program of the State of Texas, and in part by the Air Force Office of Scientific Research (AFSC) under Contract F49620-89-C-0044.

1. Introduction

In this work we obtain the asymptotic optimality of the non–stationary value iteration (NVI) adaptive policy, for a partially observed (PO) Markovian replacement process, when the performance index is the long–run average cost. The same ideas however, could be used in several other applications (see [20] for some examples) which share many of the general properties of the model to be described in the next section.

The NVI adaptive policy has been used in e.g., [8] and [11] in the discounted cost case (with complete separable metric state space), and in e.g., [1] and [13] in the average cost case (with denumerable state space). Below we describe the NVI adaptive policy for a PO production/replacement problem, the one considered by [18] and [22], when the performance index is the average cost, and prove its optimality. The proof follows the approach in [1], but the extension to the problem under study relies on some results obtained in [6] and [2] for the replacement problem. The NVI policy has also been considered for the general Borel state space case, and with the average cost criterion, in [9]. The proof of the optimality of the NVI policy in that case however, relies on several ergodicity conditions, some of which (e.g., [9, Ergodicity condition 3.1(4)]) are not satisfied in several applications (see [20]), including the replacement problem considered in this work (see also Remark 4.2 below).

This work is organized as follows. The replaeement problem is described in Section 2. In Section 3 we present a brief review of the parameter estimation algorithm developed in [21], [19], and to be used in the specification of the adaptive policy. Section 4 contains the main result of this work: it describes the NVI policy for the replacement problem, and shows its optimality. Section 5 contains some conclusions.

2. A Markovian Replacement Model

Consider the following production/replacement process. A machine that produces items at the beginning of distinct time periods $t = 0, 1, 2, \ldots$, can be in one of two states, according to its working condition. Let $\{x_t, t = 0, 1, 2, \ldots\}$ denote the state of the machine. x_t takes values in $X \equiv \{0, 1\} \equiv \{\text{good}, \text{bad}\}$. The quality of the item produced is a function of the underlying state, and it is assumed that the machine deteriorates with operation. The state of the machine is only partially observed, and the decision maker has two actions available in order to control the machine. Denote by $\{u_t, t = 0, 1, 2, \ldots\}$ the control process. u_t takes values in $U \equiv \{\text{produce}, \text{replace}\} \equiv \{0, 1\}$. There is a cost associated with each of these actions, as follows: the cost of the item produced is 0 if the machine is in the good state, and C if the machine is in the bad state; the cost of replacement, R, is assumed independent of the underlying state. Further, it is assumed that $0 < C < R < \infty$. Let $\bar{c}(u_t) \equiv (c(0, u_t), c(1, u_t))'$, $u_t \in U$, where $c: X \times U \to \mathbb{R}$, is the cost accrued when the machine is in state x_t, and action u_t is selected; $'$ denotes transpose. The observation process $\{y_t, t = 1, 2, 3, \ldots\}$ takes vatues in $Y \equiv \{0, 1\}$.

At the beginning of each time period a decision has to be made, based on the observations, on whether to replace the machine or not. If the machine is replaced, it will be in the good state at the end of the same period. It is further assumed that no item is produced during this period.

Assume for the moment that there is an underlying probability space $(\Omega, \mathcal{B}, \mathcal{P})$; it will be specified below. The state process evolves according to the transition probabilities $p_{x_t x_{t+1}}(u_t)$ defined by $p_{ij}(v) \equiv \mathcal{P}\{x_{t+1} = j | x_t = i, u_t = v\}$. The transition probability matrices $P(u_t)$, $u_t \in U$, with entries $p_{x_t x_{t+1}}(u_t)$, are given by:

$$P(0) = \begin{bmatrix} 1-\theta & \theta \\ 0 & 1 \end{bmatrix}, \qquad P(1) = \begin{bmatrix} 1-\theta & \theta \\ 1-\theta & \theta \end{bmatrix}, \qquad t \geq 0 \tag{1}$$

where θ, the unknown parameter, is the probability of machine failure in one time step, $\theta \in \Theta$, with Θ the parameter space (see Remark 4.1 below). The observation process is related to the state and the control processes by means of the conditional probabilities $q_{x_t y_{t+1}}(u_t)$, defined by $q_{ik}(v) \equiv \mathcal{P}\{y_{t+1} = k | x_t = i, u_t = v\}$. These probabilities are characterized by $q_v \in (0.5, 1.0)$, $v = 0, 1$, where q_v is the probability of making a correct observation when the action selected is v.

Let $p_{t/t} = (p^{(1)}_{t/t}, p^{(2)}_{t/t}) = (1 - \rho_{t/t}, \rho_{t/t})$ be the conditional probability distribution of the two states given the past observations and controls (the initial probabilities $p_{0/0} \equiv p_0 = (\mathcal{P}\{x_0 = 0\}, \mathcal{P}\{x_0 = 1\})$ are assumed given). That is, $p_{t/t}$ is defined as:

$$\begin{aligned} &\rho_{t/t} \equiv \mathcal{P}\{x_t = 1 | y_t, \ldots, y_1, u_{t-1}, \ldots, u_0\} \text{ for } t \geq 1, \\ &\text{and } p_{0/0} \equiv \rho_0 \equiv p^{(1)}_0. \end{aligned} \tag{2}$$

Using Bayes' rule one can obtain an expression for the updated conditional probability $\rho_{t+1/t+1}$, given y_{t+1}, u_t and prior distribution $\rho_{t/t}$. In our case, and for $t \geq 0$, the updated conditional probability is given by:

$$\rho_{t+1/t+1} = T(0, \rho_{t/t}, u_t) \cdot (1 - y_{t+1}) + T(1, \rho_{t/t}, u_t) \cdot y_{t+1} \tag{3}$$

where

$$\begin{aligned} T(0,\rho,0) &= ((1-\rho)\theta q_0 + \rho(1-q_0))/((1-\rho)q_0 + \rho(1-q_0)), \\ T(1,\rho,0) &= ((1-\rho)\theta(1-q_0) + \rho q_0)/((1-\rho)(1-q_0) + \rho q_0) \end{aligned} \tag{4}$$

and

$$T(0,\rho,1) = T(1,\rho,1) = \theta. \tag{5}$$

$T(k, \rho, v)$ is the updated conditional probability that the machine is in the bad state, given observation k, action v and prior distribution ρ.

Define $\Omega \equiv (X \times U \times Y)^\infty$ to be the canonical sample path space with elements in $X \times U \times Y$, and let $\mathcal{B}$ be the Borel σ–algebra obtained by endowing Ω with the discrete topology. We are interested in the infinite horizon case. Let h_{n+1}, $n \geq 0$, represent a generic element of H_{n+1}, the "history space" at time n, defined recursively by $H_{n+1} \equiv H_n \times U \times Y$, $H_0 \equiv \{p \in \mathbb{R}^2 \colon p^{(1)} = 1 - \rho, p^{(2)} = \rho,\ 0 \leq \rho \leq 1\}$. From [4] recall that a stochastic kernel $\mu_n(\cdot|\cdot)$ on U, given H_n, is a collection of probability distributions $\{\mu_n(\cdot|h_n), h_n \in H_n\}$ on U. An admissible control policy, denoted by $\overline{g}$, is a sequence of stochastic kernels $\{\mu_t(\cdot|\cdot)\}_{t \in \mathbb{N} \cup \{0\}}$, and if $\mu_t(\cdot|\cdot) = \mu(\cdot|\cdot)$ for all t, $\overline{g}$ is called stationary. The objective of the control problem is to find an optimal policy among the admissible policies, such that it minimizes the expected long-run average cost, given by:

$$J^{\overline{g}}(p_{0/0}) \equiv \limsup_{n\to\infty} \frac{1}{1+n} E^{\overline{g}}_{p_{0/0}} \left[\sum_{t=0}^{n} p_{t/t} \cdot \overline{c}(u_t) \right]. \tag{6}$$

Here the expectation is taken with respect to the unique probability measure $\mathcal{P}^{\overline{g}}_{p_{0/0}}$, on $\mathcal{B}$, induced by $p_{0/0}$ and a control policy $\overline{g}$ ([4, p. 140-144]). In the sequel it is understood that whenever a (given) control policy is under effect, the expectations and limits are taken with respect to (the appropriate marginal of) $\mathcal{P}^{\overline{g}}_{p_{0/0}}$. Similarly, the expressions in (1)–(5) are also interpreted with respect to $\mathcal{P}^{\overline{g}}_{p_{0/0}}$.

Furthermore, from [6] and [2], one has that when the parameters of the model are known, the optimal policies for this replacement problem can be characterized in terms of $\rho_{t/t}$, $t \geq 0$, by a so called control–limit ρ^*, $0 \leq \rho^* \leq 1$, such that for "all values of $\rho_{t/t} \leq \rho^*$, it is always optimal to let the machine produce, but for $\rho_{t/t} > \rho^*$, it is optimal to replace the machine".

Finally, let t_k, $k \geq 1$, be the k^{th} replacement time (i.e., the k^{th} time when the machine is replaced). The process evolves in regenerative cycles, and since for the model considered, if replaced at the beginning of period t_k, $k \geq 1$, the machine will be in the good state at the end of that period ([8, p. 587]), it follows that the observations $\{y_{t_k+1}\}_{k\geq 1}$ are irrelevant in the sense that the state of the machine is perfectly known at times $t_k + 1$, $k \geq 1$. Thus, $\mathcal{P}^{\overline{g}}_{p_{0/0}}\{y_{t_k+1} = 1\} = 1 - q_0$ and $\mathcal{P}^{\overline{g}}_{p_{0/0}}\{y_{t_k+1} = 0\} = q_0$ do not depend on θ, meaning that the sequence of observations $\{y_{t_k+1}\}_{k\geq 1}$ provides no information about the unknown parameter θ. This property of the model was taken into account in the specification of the parameter estimation algorithms (see [21], [19] for details).

3. The Parameter Estimation Algorithm

Denote by θ_0 the (unknown) true value of the parameter. It is assumed that θ_0 is constant and that $\theta_0 \in \Theta$. The estimation algorithm to be shown below is based on the minimization of the expected value of the square of the prediction error, and so it will be referred to as the *pe* algorithm. It takes into account an arbitrary (random) number of observations made in each of the regenerative cycles.

The *pe* algorithm, for $n \geq 1$, is specified by (see [21], [19]):

$$\hat{\theta}_{n+1} = \pi_\Theta \left\{ \hat{\theta}_n + \frac{1}{n+1} R^{-1}_{n+1} \cdot \psi_n(\hat{\theta}_n) \cdot \epsilon_n(\hat{\theta}_n) \right\}, \tag{7}$$

$$R_{n+1} = R_n + \frac{1}{n+1} \left[\psi_n(\hat{\theta}_n) \cdot \psi'_n(\hat{\theta}_n) - R_n \right], \tag{8}$$

$\hat{\theta}_1 \in \Theta$, $R_1 = 1$, where: $\epsilon_n(\cdot)$ is the prediction error at the n^{th} replacement time, defined as the difference between the vector of observations available in the n^{th} regenerative cycle, and its expected value; $\psi(\cdot)$ is defined as the negative of the partial derivative of the prediction error with respect to the unknown parameter. That is, the search for the true value of the parameter is in the direction of the negative gradient of the prediction error, and the step size is proportional to the magnitude of this error. The gains R_n, $n \geq 1$, minimize the variance of the estimates (see [3, Remark 3, p. 643], or [17, Remark 3.3, p. 150]); π_Θ is a projection operator that takes into account that θ is a probability.

In [21], [19], by means of the ODE method (see e.g., [16]), it was shown that the *pe* algorithm converges w.p.1 (with respect to the unique probability measure induced by a given admissible policy) to θ_0. It is also shown in [21] and [19] that θ_0 is the unique, globally asymptotically stable limit point associated with the *pe* algorithm. We refer the reader to [21], [19] for details. In the sequel we will use the *pe* algorithm to specify the NVI adaptive policy associated with the replacement problem under consideration.

4. NVI Adaptive Policy

We let $\hat{c}(\rho, u, \theta) \equiv p \cdot \overline{c}(u) = (1-\rho, \rho) \cdot \overline{c}(u)$ to show not only that these costs are functions of ρ (recall that we are using the fact that p is completely characterized by ρ, and so the optimization problem (6) reduces to a scalar one), but also that they could be explicit functions of θ. Although that is not the case for the particular model of the replacement problem that we have been considering, the results to be presented also apply when the immediate costs depend on θ. Thus, from now on, we work under this new asumption.

In order to describe an "iteration" of the NVI adaptive policy for the replacement problem of Section 2, we introduce the functions $h_t\colon [0,1] \times \Theta \to \mathbb{R}$, $t \geq 0$, as follows:

$$\begin{aligned} h_0(\rho, \theta) &\equiv \min_{u \in U}\{\hat{c}(\rho, u, \theta)\} \\ h_{t+1}(\rho, \theta) &\equiv \min_{u \in U}\{\hat{c}(\rho, u, \theta) + \sum_{k=0}^{1} D(k, \rho, u, \theta) h_t(T(k, \rho, u, \theta), \theta)\}, \end{aligned} \tag{9}$$

for $t = 0, 1, 2, \ldots$, where $D(k, \rho, u, \theta)$ is, for each $\theta \in \Theta$, the probability that the next observation will be k, given the probability distribution ρ and control action u. Also, define the sequence of functions $\nu_t\colon [0,1] \times \Theta \to U$, $t \geq 0$, by:

$$\begin{aligned} \nu_0(\rho, \theta) &\equiv \arg\min_{u \in U}\{\hat{c}(\rho, u, \theta)\} \\ \nu_{t+1}(\rho, \theta) &\equiv \arg\min_{u \in U}\{\hat{c}(\rho, u, \theta) + \sum_{k=0}^{1} D(k, \rho, u, \theta) h_t(T(k, \rho, u, \theta), \theta)\}, \end{aligned} \tag{10}$$

for $t = 0, 1, 2, \ldots$.

Observe that since U is finite, there is no problem guaranteeing the existence of the functions $\{\nu_t(\rho, \theta)\}_{t \geq 0}$. The policy $\overline{g}_{\mathrm{NVI}} \equiv \{\nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t)\}_{t \geq 0}$ (also denoted $\overline{g}_{\mathrm{NVI}}(\rho, \theta)$), with $\{\nu_t\}_{t \geq 0}$ given by (10), $\{\hat{\rho}_{t/t}\}_{t \geq 0}$ a sequence of estimates of the conditional probability that the machine is in the bad state at time t given past observations and actions, and $\{\hat{\theta}_t\}_{t\geq 1}$ a sequence of estimates converging w.p.1 to (the true value of the parameter) θ_0, is called an NVI adaptive policy.

An "iteration" of the NVI adaptive policy, for the replacement problem of Section 2, can be described as follows. Let $\hat{\theta}_k$ and $\hat{\rho}_{t_k/t_k}$, be given from the previous iteration; ($\hat{\rho}_{t_k/t_k}$ is an estimate of the conditional probability that the machine is in the bad state at time t_k, given past observations and actions, computed using the parameter estimates $\{\hat{\theta}_k\}_{k\geq 1}$ as if they were, for each k, the true value of the parameter; see equation (11) below). Using observation y_{t_k+1} one updates $\hat{\rho}_{t_k+1/t_k+1}$ by means of:

$$\hat{\rho}_{t_k+1/t_k+1} = T(0, \hat{\rho}_{t_k/t_k}, u_{t_k}, \hat{\theta}_k)\cdot(1-y_{t_k+1}) + T(1, \hat{\rho}_{t_k/t_k}, u_{t_k}, \hat{\theta}_k)\cdot y_{t_k+1} \quad (11)$$

(this is equation (3) with $\hat{\theta}_k$ taken as the true value of the parameter). Next, one computes $h_{t_k+1}(\rho,\theta)$ and $\nu_{t_k+1}(\rho,\theta)$ using equations (9) and (10) respectively. The control action at time t_k+1 is $u_{t_k+1} = \nu_{t_k+1}(\rho,\theta)$. If $u_{t_k+1} = 0$, the time counter is increased and the iterative process is repeated with the same value of the parameter estimate. If $u_{t_k+1} = 1$, one uses the parameter estimation algorithm to update the estimate, and after the time counter is increased, the iterative procedure is repeated, now with $\hat{\theta}_{k+1}$ as the true value of the parameter.

REMARK 4.1 In order for the procedure described above to work, one has to guarantee that the process will replace infinitely often (i.o.) for otherwise, if the parameter estimate reaches a value for which the solution of the optimization problem is to "produce for all values of ρ", then convergence will not be obtained. Since for this replacement problem it is known (see [2, Lemma 4.1]) that under adaptive control policies (parametrized by θ), regeneration (i.e., replacement of failed machines) occurs i.o., we assume the following to hold throughout this work:

ASSUMPTION. Replacement occurs i.o., and after each replacement the machine is in the good state.

This assumption is satisfied if the parameter space is given by

$$\Theta = \left[\delta, \min\{1-\delta, \frac{C}{R-C} - \delta\}\right],$$

$\delta > 0$ (cf. [2, Theorem 3.2]). This remark also applies in the case of the certainty equivalent adaptive policy (see [2]).

In order to prove the average cost optimality of the NVI adaptive policy, we need to consider the sequence of functions $\{e_t(\rho,\theta)\}_{t\geq 0}$, $\rho \in [0,1]$, $\theta \in \Theta$, defined by:

$$e_t(\rho,\theta) \equiv h_t(\rho,\theta) - t \cdot V(0) - h(\rho,\theta), \quad t = 0, 1, \ldots, \tag{12}$$

where $V(\cdot)$ and $h(\cdot,\cdot)$ are, respectively, the constant and the mapping satisfying the average cost optimality equation for the replacement problem of Section 2 (see equation (13) below), and the functions $h_t(\cdot,\cdot)$, $t = 0, 1, \ldots$, are those computed using equation (9).

For each $\theta \in \Theta$, the sequence $\{e_t(\rho,\theta)\}_{t\geq 0}$, $\rho \in [0,1]$ converges to a unique, finite limit point, which is independent of ρ. This follows from the work in [12], because for the replacement problem considered here, despite the fact that the number of possible states is uncountably infinite, *the set of possible next states is finite* since Y and U are both finite sets. Thus, the results in [12], obtained for the denumerable state space problem, apply to our replacement problem as well. We now show this explicitly.

The remainder of this section is organized as follows. First, we prove that for each $\theta \in \Theta$ the sequence defined in equation (12) is bounded for all $\rho \in [0,1]$ and $t \geq 0$. Next, we show that for each $\theta \in \Theta$ the sequence $\{e_t(\rho,\theta)\}_{t\geq 0}$, $\rho \in [0,1]$, has a finite limit point, which is independent of $\rho \in [0,1]$. In order to prove this last statement, we will recall some of the results obtained in [6] for PO Markov decision models. Finally, the above results, together with some additional results obtained in [11] and [2] for the replacement problem considered in this work, will be used to show the optimality of the NVI adaptive policy.

Observe that:

(C1) For the replacement problem considered in this work, there exist (for each $\theta \in \Theta$) a bounded constant $V(\theta)$ and a bounded (measurable) function $h(\rho,\theta)$, $\rho \in [0,1]$, satisfying:

$$V(\theta) + h(\rho,\theta) = \min_{u\in U}\{\hat{c}(\rho,u,\theta) + \sum_{k=0}^{1} D(k,\rho,u,\theta)h(T(k,\rho,u,\theta),\theta)\}, \tag{13}$$

(cf. [5, Theorem 3.1], [6, Theorem 4.2]). Furthermore,

(C2) Since for each $\theta \in \Theta$ the immediate costs $c(x_t, u_t)$ are uniformly bounded for $x_t \in X$, $u_t \in U$ and $t \geq 0$, then $h_0(\cdot,\cdot)$, defined in equation (9), is finite.

Thus we have the following result.

LEMMA 4.1 *Consider the sequence* $\{e_t(\rho,\theta)\}_{t\geq 0}$, $\rho \in [0,1]$, $\theta \in \Theta$, *defined in equation (12). Then there exists a constant* $M(\theta) < \infty$, *such that for each* $\theta \in \Theta$, $|e_t(\rho,\theta)| \leq M(\theta)$, *for all* $\rho \in [0,1]$ *and* $t \geq 0$.

PROOF: The proof of this lemma closely follows that of [12, Lemma 1]. Because of results (C1) and (C2) mentioned above, for each $\theta \in \Theta$ there is a finite number $M(\theta)$ such that for all $\rho \in [0,1]$, $e_0(\rho,\theta)$ is bounded by $M(\theta)$. Assume that $|e_t(\rho,\theta)| \leq M(\theta)$ for all $\rho \in [0,1]$. Also, let $\overline{g}^\theta$ be an average cost optimal policy (i.e., one that minimizes the right hand side of equation (13)), and let $\overline{g}_t^\theta$ be a policy minimizing the right hand side of equation (9). Then, from (9):

$$h_{t+1}(\rho,\theta) \leq \hat{c}(\rho,\overline{g}^\theta(\rho),\theta)+\sum_{k=0}^{1} D(k,\rho,\overline{g}^\theta(\rho),\theta)h_t(T(k,\rho,\overline{g}^\theta(\rho),\theta),\theta). \quad (14)$$

Now, subtract $h(\rho,\theta)$ and $(t+1)V(\theta)$ from both sides of (14), to obtain:

$$\begin{aligned} e_{t+1}(\rho,\theta) &\leq \hat{c}(\rho,\overline{g}^\theta(\rho),\theta) + \sum_{k=0}^{1} D(k,\rho,\overline{g}^\theta(\rho),\theta)h_t(T(k,\rho,\overline{g}^\theta(\rho),\theta),\theta) \\ &\quad -tV(\theta) - V(\theta) - h(\rho,\theta) \\ &= \sum_{k=0}^{1} D(k,\rho,\overline{g}^\theta(\rho),\theta)h_t(T(k,\rho,\overline{g}^\theta(\rho),\theta),\theta) - tV(\theta) \\ &\quad -\sum_{k=0}^{1} D(k,\rho,\overline{g}^\theta(\rho),\theta)h(T(k,\rho,\overline{g}^\theta(\rho),\theta),\theta), \end{aligned} \quad (15)$$

where the equality follows because $\overline{g}^\theta$ satisfies (13). That is, we obtain:

$$e_{t+1}(\rho,\theta) \leq \sum_{k=0}^{1} D(k,\rho,\overline{g}^\theta(\rho),\theta)e_t(T(k,\rho,\overline{g}^\theta(\rho),\theta),\theta). \quad (16)$$

Similarly, from (9):

$$h_{t+1}(\rho,\theta) = \hat{c}(\rho,\overline{g}_t^\theta(\rho),\theta)+\sum_{k=0}^{1} D(k,\rho,\overline{g}_t^\theta(\rho),\theta)h_t(T(k,\rho,\overline{g}_t^\theta(\rho),\theta),\theta), \quad (17)$$

since $\overline{g}_t^\theta$ minimizes the right hand side of (9) at the $(t+1)^{\text{st}}$ iteration. Following the same procedure we just did to obtain (16), one obtains from (17) that:

$$e_{t+1}(\rho,\theta) = \sum_{k=0}^{1} D(k,\rho,\overline{g}_t^\theta(\rho),\theta)e_t(T(k,\rho,\overline{g}_t^\theta(\rho),\theta),\theta). \quad (18)$$

As in [12], from equations (16), (18) and the induction hypothesis, one obtains that for each $\theta \in \Theta$, $|e_t(\rho, \theta)| \leq M(\theta) < \infty$, for all $\rho \in [0,1]$, completing the proof. ■

We note that the proof of Lemma 4.1 requires fewer assumptions than that of ([12, Lemma 1]) since the summations in equations (14) through (18) are finite, due to the finiteness of the set of possible next states.

We recall from [6] the following definitions and results, to be used in the proof of the optimality of the NVI adaptive policy:

- ([6, Definition 4.1]) For each $\rho \in [0,1]$, the sets of *ancestors, descendents* and *relatives* of ρ are defined recursively as follows:

$$\begin{aligned} A_\rho &\equiv \{s \in [0,1] \colon \exists n \in \mathbb{N} \cup \{0\},\ y^{n+1} \equiv \{y_1, \ldots, y_{n+1}\} \subseteq Y, \\ &\qquad u^n \equiv \{u_0, \ldots, u_n\} \subseteq U, \text{ for which } \rho = T(y^{n+1}, s, u^n)\}, \\ D_\rho &\equiv \{s \in [0,1] \colon \exists n \in \mathbb{N} \cup \{0\},\ y^{n+1} \equiv \{y_1, \ldots, y_{n+1}\} \subseteq Y, \\ &\qquad u^n \equiv \{u_0, \ldots, u_n\} \subseteq U, \text{ for which } s = T(y^{n+1}, \rho, u^n)\}, \\ R_\rho^{(1)} &\equiv A_\rho \cup \{\rho\} \cup D_\rho, \end{aligned}$$

  where the maps used in these definitions are obtained by using the maps given in (4), and are computed recursively as follows:

$$\begin{aligned} T(y^1, \cdot, u^0) &\equiv T(y_1, \cdot, u_0), \\ T(y^{n+1}, \cdot, u^n) &\equiv T(y_{n+1}, T(y^n, \cdot, u^{n-1}), u_n). \end{aligned}$$

- ([6, Definition 4.2]) For $\rho \in [0,1]$, define its *genealogical tree* GT_ρ as $GT_\rho \equiv \cup_{n \in \mathbb{N}} R_\rho^{(n)}$, where the sets $R_\rho^{(n)}$ are defned recursively by $R_\rho^{(n+1)} \equiv \cup_{s \in R_\rho^{(n)}} R_s^{(1)}$, $n \in \mathbb{N}$.

- ([6, Section 6]) For the replacement problem considered in this work, the maps $T(y, \cdot, 0)$ are injective for $y \in Y$, and since $T(y, \cdot, 1) = \theta$, then for all $\rho \in [0,1]$, GT_ρ is a countable set. Also (see [6, Section 4.1]), GT_ρ is the smallest invariant set containing ρ (a set B contained in the Borel σ-algebra generated by $[0,1]$, $\mathcal{B}([0,1])$, is called *invariant* if $D_\rho \subseteq B$ and $A_\rho \subseteq B$ for all $\rho \in B$; see [6, Definition 4.3]). Furthermore ([6, Section 4]), the sequence of conditional probabilities $\{\rho_{t/t}\}_{t \geq 0}$, defined by equation (3), remains in the particular subset GT_ρ containing the given initial distribution ρ.

- Finally, recall result (C1) mentioned above, namely, that under the uniform boundedness condition there is, for each $\theta \in \Theta$, a bounded solution to the average cost optimality equation (13) on each invariant set GT_ρ, for all $\rho \in [0,1]$. Furthermore, the same average cost is attained on each invariant set GT_ρ, for all $\rho \in [0,1]$.

These results suggest that the proof that the sequence $\{e_t(\rho,\theta)\}_{t\geq 0}$, $\rho \in [0,1]$, defined in (12), has a finite limit point independent of ρ, for each $\theta \in \Theta$, follows from the results obtained in [12] for the denumerable state space case. This is so, provided that the following conditions are satisfied:

(C3) For each $\theta \in \Theta$, and for any stationary policy, the Markov chain restricted to GT_ρ, for all $\rho \in [0,1]$, is non–dissipative.

(C4) For each $\theta \in \Theta$, and for any average cost optimal policy it holds that each state in GT_ρ, for all $\rho \in [0,1]$, which is positive recurrent under this average cost optimal policy, is also aperiodic.

(C5) For each $\theta \in \Theta$, and for any average cost optimal stationary policy the associated Markov chain in GT_ρ, for all $\rho \in [0,1]$, has no two disjoint closed sets.

(conditions (C3), (C4) and (C5) correspond to assumptions 3, 4 and 5 respectively, in [12], assumptions 1 and 2 in [12] correspond to results (C1) and (C2), which we already mentioned, hold for the problem considered in this work).

For each $\theta \in \Theta$, let $m(\rho,\theta) \equiv \lim\inf_{t\to\infty} e_t(\rho,\theta)$ and $M(\rho,\theta) \equiv \lim\sup_{t\to\infty} e_t(\rho,\theta)$. By Lemma 4.1 the functions $m(\cdot,\cdot)$ and $M(\cdot,\cdot)$ are bounded. Consider the following condition, weaker than condition (C3):

(C3') For each $\theta \in \Theta$, and for any stationary policy which minimizes the right hand side of equation (9), or the right hand side of

$$\begin{aligned} m(\rho,\theta) \quad \geq \quad & \min_{v\in U}\{\hat{c}(\rho,v,\theta) - V(\theta) + \sum_{k=0}^{1} D(k,\rho,v,\theta)h(T(k,\rho,v,\theta),\theta) \\ & -h(\rho,\theta) + \sum_{k=0}^{1} D(k,\rho,v,\theta)m(T(k,\rho,v,\theta),\theta)\}, \end{aligned}$$

or the right hand side of

$$M(\rho,\theta) \quad \leq \quad \min_{v\in U}\{\hat{c}(\rho,v,\theta) - V(\theta) + \sum_{k=0}^{1} D(k,\rho,v,\theta)h(T(k,\rho,v,\theta),\theta)$$

$$-h(\rho,\theta)+\sum_{k=0}^{1} D(k,\rho,v,\theta)M(T(k,\rho,v,\theta),\theta)\},$$

the Markov chain specified by GT_ρ, for all $\rho \in [0,1]$, is non–dissipative.

From the proofs in [12] it follows that if the Markov chain has a non–empty set of recurrent states under the policies specified in (C3'), then (C3') (as opposed to (C3)) suffices for obtaining the results in [12]. This can be concluded by simply following the proofs in [12]: cf., [12, p. 300, proof of Lemma 1] and [12, p. 304, proof of Theorem 2]; in all the other cases in which the non–dissipativeness assumption is used in [12], it is associated with average cost optimal policies: e.g., [12, p. 301, proof of Theorem 1] , [12, p. 502, proof of Lemma 3].

In our case, (C3') and a non–empty set of recurrent states under the policies specified in (C3') is all that is needed because, as remarked above, the adaptive policy will only work under the assumption that the process replaces i.o. (cf. Remark 4.1), and so we do not need to consider arbitrary stationary policies.

We now verify that conditions (C3'), (C4) and (C5) are satisfied for the PO replacement problem considered in this work.

First, observe that due to the regenerative structure of the replacement process being considered, the zero state in GT_ρ, for all $\rho \in [0,1]$, is recurrent. Now, the replacement process starts in a state that may not be reached from the zero state (recall that the set of next states is finite due to the finiteness of Y and U). However, *after the first replacement*, and for each $\theta \in \Theta$, the Markov chain restricted to GT_0 specifies a recurrent class. This is the case because zero is a recurrent state and recurrence is an equivalence relation. Also, note that from the results in [6] we have that after the first replacement the Markov chain always lives in GT_0, which as mentioned before, is a countable set. In addition, since we are considering the long–run average cost problem, the early behaviour of the process (i.e., that up to the first replacement), does not affect the optimal expected average cost.

From the above discussion we have that GT_0 does not contain two disjoint closed subsets, because: (a) GT_0 specifies a recurrent class, and so it does not contain transient states, and (b) GT_0 is the smallest invariant set containing 0 (the zero state).

Furthermore, if we denote by P_{ij} the probabilities associated with the Markov chain restricted to GT_0, with $i,j \in GT_0$, then $\sum_{j=1}^{N} P_{ij} \to 1$ uniformly in i as $N \to \infty$ (simply because GT_0 specifies a single class). This in turn implies, by [7, Theorem 3], that for each $\theta \in \Theta$ the Markov chain restricted to GT_0 is non–dissipative.

Finally, from [15, Chapter 6, page 145], recall that if for a denumerable recurrent Markov chain with period d, one defines the relation R on the states of the Markov chain by: "iRj if and only if starting at i the process can reach j in md steps, for some $m \in \mathbb{N}$", then R is an equivalence relation, and it partitions the states into *cyclic* (or *periodic*) subclasses. The important point here is that if one observes the chain after every d steps, the resulting process is again a Markov chain, with *noncyclic* (or *aperiodic*, or periodic with period 1) behaviour. Thus a recurrent Markov chain with period $d > 1$ is really d separate recurrent aperiodic classes. It follows for the replacement problem we are studying that since for each $\theta \in \Theta$, GT_0 is the smallest invariant set, then GT_0 specifies a recurrent aperiodic chain. In fact, since the Markov chain is non-dissipative, it follows from [15, Definition 6.2, page 131] that this chain is ergodic. This fact, together with the recurrent and aperiodic behaviour (and denoting again by P_{ij} the probabilities associated with the countable state space specified by GT_0, with $i, j \in GT_0$) implies that $\lim_{n\to\infty} P_{ii}^{(n)}$ exists and is strictly positive for all i in the class GT_0. Thus for each $\theta \in \Theta$ GT_0 specifies a *positive recurrent* (or *strongly ergodic*) Markov chain (see [14, page 85]).

The previous discussion implies that assumptions 1, 2, 4 and 5 in [12], and the weaker condition (C3') given above, are satisfied for the replacement problem considered in this work. This, together with Lemma 4.1, enables the conclusion of [12, Theorem 2]. We state this result precisely in the following lemma.

LEMMA 4.2 *For each $\theta \in \Theta$, the sequence $\{e_t(\rho, \theta)\}_{t\geq 0}$, $\rho \in [0, 1]$, defined in equation (12), converges to a finite limit point as $t \to \infty$, for all $\rho \in [0, 1]$. Furthermore, this limit is independent of ρ.*

REMARK 4.2 In [1] and [9], the proofs of the optimality of the NVI policy use the fact that sequences analogous to that defined in (12), can be bounded for each $t = 0, 1, 2, \ldots$, by $\overline{M} \cdot (1 - \mu)^t$, with $\overline{M} < \infty$ and $0 < \mu < 1$. This is not the case for the replacement problem of Section 2.

Recall from [11, Lemma 3.1 and Example in page 238], that for the PO replacement problem considered in this work we have that:

$$\sup_{\rho,\mu} |\hat{c}(\rho, u, \theta_0) - \hat{c}(\rho, u, \hat{\theta}_t)| \to 0 \quad \text{as} \quad t \to \infty, \tag{19}$$

and

$$\sup_{\rho,\mu} \|D(\cdot,\rho,u,\theta_0) - D(\cdot,\rho,u,\hat{\theta}_t)\|_v \to 0 \quad \text{as} \quad t \to \infty \tag{20}$$

(with $\|\cdot\|_v$ the total variation norm for signed measures; cf. [11, p. 237]). With these results we can state the following lemma, the proof of which follows from [1, Lemma 3.0].

LEMMA 4.3 *For all $\rho \in [0,1]$, the functions $h_t(\rho,\cdot)\colon \Theta \to \mathbb{R}$, $t \geq 0$, defined in equation (9), are continuous.*

We can now state the main result of this work.

THEOREM *Consider the replacement problem described in Section 2, with $U = \{produce,\ replace\}$, and $q_v \in (0.5, 1.0)$, $v = 0,1$. Let $\{\hat{\theta}_t\}_{t\geq 1}$ be a sequence of estimates that converges w.p.1 (with respect to the unique probability measure induced by the NVI adaptive policy, say $\mathcal{P}^{\overline{g}_{\text{NVI}}}_{\rho_0,\theta_0}$, since ρ_0 completely characterizes $p_{0/0}$) to θ_0. Then, the NVI adaptive policy $\overline{g}_{\text{NVI}}$ is average cost optimal.*

PROOF. The reasoning of this proof parallels that in [1]. Define for each $\rho \in [0,1]$ and $u \in U$, the function:

$$\begin{aligned} \Phi(\rho,u,\theta_0) \equiv\ & \hat{c}(\rho,u,\theta_0) + \\ & + \sum_{k=0}^{1} D(k,\rho,u,\theta_0) h(T(k,\rho,u,\theta_0),\theta_0) - h(\rho,\theta_0) - V(\theta_0) \end{aligned} \tag{21}$$

with θ_0 the true (unknown) value of the parameter.

Note that since $h(\cdot,\cdot)$ and $V(\cdot)$ are bounded, so is $\Phi(\cdot,\cdot,\cdot)$. Furthermore, it satisfies (see e.g., [10, Theorem 3.2] or [1, Theorem 3.2]).

$$\begin{aligned} \Phi(\rho_{t/t},u_t,\theta_0) = E^{\overline{g}_{\text{NVI}}}_{\rho_0,\theta_0} \big[& \hat{c}(\rho_{t/t},u_t,\theta_0) + h(\rho_{t+1/t+1},\theta_0) \\ & - h(\rho_{t/t},\theta_0) | \rho_{t/t},u_t \big] - V(\theta_0) \end{aligned} \tag{22}$$

for each $t = 0,1,2,\ldots$. Next: (i) sum both sides of (22) from $t=0$ to $t=l$; (ii) take expectation (with respect to $\mathcal{P}^{\overline{g}_{\text{NVI}}}_{\rho_0,\theta_0}$) on both sides of the resulting expression (this gets rid of the conditional expectation); and (iii) multiply both sides of the resulting expression by $1/(l+1)$, to obtain:

$$\begin{aligned} \frac{1}{l+1} E^{\overline{g}_{\text{NVI}}}_{\rho_0,\theta_0} \left[\sum_{t=0}^{l} \Phi(\rho_{t/t},u_t,\theta_0) \right] = & \frac{1}{l+1} E^{\overline{g}_{\text{NVI}}}_{\rho_0,\theta_0} \left[\sum_{t=0}^{l} \hat{c}(\rho_{t/t},u_t,\theta_0) \right] \\ & - V(\theta_0) + \frac{1}{l+1} E^{\overline{g}_{\text{NVI}}}_{\rho_0,\theta_0} [h(\rho_{l/l},\theta_0) - h(\rho_0,\theta_0)]. \end{aligned} \tag{23}$$

Since $h(\cdot,\cdot)$ is bounded (cf. [5, Theorem 3.1]), the last term in (23) vanishes as $l \to \infty$. Therefore, for the NVI adaptive policy $\overline{g}_{\text{NVI}}$, we will have that:

$$V(\theta_0) = \limsup_{n\to\infty} \frac{1}{l+1} E^{\overline{g}_{\text{NVI}}}_{\rho_0,\theta_0}\left[\sum_{t=0}^{l} \hat{c}(\rho_{t/t}, u_t, \theta_0)\right], \tag{24}$$

meaning that $\overline{g}_{\text{NVI}}$ is average cost optimal, if we can show that, under $\overline{g}_{\text{NVI}}$, the left hand side of (23) goes to zero as $l \to \infty$.

In order to prove this last statement, let, as before, $\{\hat{\theta}\}_{t\geq 1}$ be a sequence of estimates that converges w.p.1 to θ_0, and let $\hat{\rho}_{t/t}$ be an estimate of the conditional probability that the machine is in the bad state at time t, computed using the estimate θ_t. Then (cf. [1])

$$\begin{aligned}
\Phi(\hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \theta_0) &= \Phi(\hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \theta_0) \\
&\quad - \Big(h_t(\hat{\rho}_{t/t}, \hat{\theta}_t) - h_t(\rho_0, \hat{\theta}_t)\Big) + \Big(h_t(\hat{\rho}_{t/t}, \hat{\theta}_t) - h_t(\rho_0, \hat{\theta}_t)\Big) \qquad (25) \\
= \;& \hat{c}(\hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \theta_0) - h_t(\hat{\rho}_{t/t}, \theta_0) - V(\theta_0) \\
&\quad + \sum_{k=0}^{1} D(k, \hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \theta_0) h(T(k, \hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \theta_0), \theta_0) \\
&\quad - \Big[\hat{c}(\hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \hat{\theta}_t) - h_t(\rho_0, \hat{\theta}_t) \\
&\quad + \sum_{k=0}^{1} D(k, \hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \hat{\theta}_t) h_{t-1}(T(k, \hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \hat{\theta}_t), \hat{\theta}_t)\Big] \\
&\quad + \Big(h_t(\hat{\rho}_{t/t}, \hat{\theta}_t) - h_t(\rho_0, \hat{\theta}_t)\Big), \qquad (26)
\end{aligned}$$

where: the first four terms in (26) correspond to the definition of $\Phi(\cdot,\cdot,\cdot)$ in (21); the terms in square brackets come from the definition of $h_t(\cdot,\cdot)$ in (9) and the third term in (25); and the terms in parentheses are those in (25). Now, we add and subtract $h(\rho_0, \theta_0)$ and $h_{t-1}(\rho_0, \hat{\theta}_t)$ in the right hand side of equation (26), and rewrite it as follows:

$$\begin{aligned}
&= \Big[\hat{c}(\hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \theta_0) - \hat{c}(\hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \hat{\theta}_t)\Big] \\
&\quad + \left[\sum_{k=0}^{1} D(k, \hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \theta_0)\right. \\
&\qquad \cdot \Big(h(T(k, \hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \theta_0), \theta_0) - h(\rho_0, \theta_0)\Big) \\
&\quad - \sum_{k=0}^{1} D(k, \hat{\rho}_{t/t}, \nu_t(\hat{\rho}_{t/t}, \hat{\theta}_t), \hat{\theta}_t)
\end{aligned}$$

$$
\begin{aligned}
&\cdot\Big(h_{t-1}(T(k,\hat\rho_{t/t},\nu_t(\hat\rho_{t/t},\hat\theta_t),\hat\theta_t),\hat\theta_t)-h_{t-1}(\rho_0,\hat\theta_t)\Big)\Big]\\
&+\Big[h_t(\hat\rho_{t/t},\hat\theta_t)-h_t(\rho_0,\hat\theta_t)+h(\rho_0,\theta_0)-h_t(\hat\rho_{t/t},\theta_0)\Big]\\
&+\Big[h_t(\rho_0,\hat\theta_t)-h_{t-1}(\rho_0,\hat\theta_t)-V(\theta_0)\Big]\,. \qquad (27)
\end{aligned}
$$

We want to show that each term in square brackets in equation (27) goes to 0 as $t \to \infty$. This will complete the proof of the theorem.

The first term in brackets in equation (27) vanishes as $t \to \infty$ because of (19). For simplicity in the presentation, we consider first the last two terms in brackets in (27). Rewrite the third term in brackets in (27) as:

$$
\begin{aligned}
&h_t(\hat\rho_{t/t},\hat\theta_t)-h_t(\rho_0,\hat\theta_t)+h_t(\rho_0,\theta_0)-h_t(\hat\rho_{t/t},\theta_0) \qquad (28)\\
&=\Big(h_t(\hat\rho_{t/t},\hat\theta_t)-t\cdot V(\theta_0)-h_t(\hat\rho_{t/t},\theta_0)\Big)-\Big(h_t(\rho_0,\hat\theta_t)-t\cdot V(\theta_0)-h_t(\rho_0,\theta_0)\Big),
\end{aligned}
$$

and similarly, rewrite the fourth term in brackets in (27) as:

$$
\begin{aligned}
&h_t(\rho_0,\hat\theta_t)-h_{t-1}(\rho_0,\hat\theta_t)-V(\theta_0)=\Big(h_t(\rho_0,\hat\theta_t)-t\cdot V(\theta_0)-h(\rho_0,\theta_0)\Big)\\
&\quad-\Big(h_{t-1}(\rho_0,\hat\theta_t)-(t-1)\cdot V(\theta_0)-h(\rho_0,\theta_0)\Big)\,. \qquad (29)
\end{aligned}
$$

Since: (i) $\hat\theta_t \to \theta_0$ w.p.1 as $t \to \infty$; (ii) $h_t(\rho,\cdot)$ are continuous functions (cf. Lemma 4.3); and (iii) for each $\theta \in \Theta$, the sequences $\{e_t(\rho,\theta)\}_{t\geq 0}$, $\rho \in [0,1]$, converge to the same limit independent of ρ, we obtain that equations (28) and (29) converge to zero as $t \to \infty$.

For the second term in brackets in equation (27) we write, after "adding zero" to the second summation, the following:

$$
\begin{aligned}
&\sum_{k=0}^{1} D(k,\hat\rho_{t/t},\nu_t(\hat\rho_{t/t},\hat\theta_t),\theta_0)\Big(h(T(k,\hat\rho_{t/t},\nu_t(\hat\rho_{t/t},\hat\theta_t),\theta_0),\theta_0)-h(\rho_0,\theta_0)\Big)\\
&\quad-\sum_{k=0}^{1} D(k,\hat\rho_{t/t},\nu_t(\hat\rho_{t/t},\hat\theta_t),\hat\theta_t)\Big\{h_{t-1}(T(k,\hat\rho_{t/t},\nu_t(\hat\rho_{t/t},\hat\theta_t),\hat\theta_t),\hat\theta_t)\\
&\quad-h_{t-1}(\rho_0,\hat\theta_t)+h(T(k,\hat\rho_{t/t},\nu_t(\hat\rho_{t/t},\hat\theta_t),\hat\theta_t),\theta_0)\\
&\quad-h(T(k,\hat\rho_{t/t},\nu_t(\hat\rho_{t/t},\hat\theta_t),\hat\theta_t),\theta_0)+h(\rho_0,\theta_0)-h(\rho_0,\theta_0)\\
&\quad+(t-1)\cdot V(\theta_0)-(t-1)\cdot V(\theta_0)\}\,. \qquad (30)
\end{aligned}
$$

The expression in (30) can be rewritten as follows:

$$
\left(-\sum_{k=0}^{1} D(k,\hat\rho_{t/t},\nu_t(\hat\rho_{t/t},\hat\theta_t),\theta_0)+\sum_{k=0}^{1} D(k,\hat\rho_{t/t},\nu_t(\hat\rho_{t/t},\hat\theta_t),\hat\theta_t)\right)h(\rho_0,\theta_0)
$$

$$
\begin{aligned}
&+\Bigg(-\sum_{k=0}^{1} D(k,\hat{\rho}_{t/t},\nu_t(\hat{\rho}_{t/t},\hat{\theta}_t),\hat{\theta}_t)\\
&\cdot\Big[\Big\{h_{t-1}(T(k,\hat{\rho}_{t/t},\nu_t(\hat{\rho}_{t/t},\hat{\theta}_t),\hat{\theta}_t),\hat{\theta}_t)-(t-1)\cdot V(\theta_0)\\
&-h(T(k,\hat{\rho}_{t/t},\nu_t(\hat{\rho}_{t/t},\hat{\theta}_t),\hat{\theta}_t),\theta_0)\Big\}\\
&-\Big\{h_{t-1}(\rho_0,\hat{\theta}_t)-(t-1)\cdot V(\theta_0)+h(\rho_0,\theta_0)\Big\}\Big]\Bigg)\\
&+\Bigg(\sum_{k=0}^{1} D(k,\hat{\rho}_{t/t},\nu_t(\hat{\rho}_{t/t},\hat{\theta}_t),\theta_0)h(T(k,\hat{\rho}_{t/t},\nu_t(\hat{\rho}_{t/t},\hat{\theta}_t),\theta_0),\theta_0)\\
&-\sum_{k=0}^{1} D(k,\hat{\rho}_{t/t},\nu_t(\hat{\rho}_{t/t},\hat{\theta}_t),\hat{\theta}_t)h(T(k,\hat{\rho}_{t/t},\nu_t(\hat{\rho}_{t/t},\hat{\theta}_t),\hat{\theta}_t),\theta_0)\Bigg). \qquad (31)
\end{aligned}
$$

Thus, we have that: (i) the first term in parentheses in (31) vanishes as $t \to \infty$ because of equation (20), since $h(\cdot,\cdot)$ is bounded; (ii) the second term in parentheses in (31) goes to zero as $t \to \infty$ because the summation is finite, and the terms in brackets cancel each other, where the reasoning is identical to that used to cancel the third and fourth terms in equation (27); and (iii) since $h(\rho,\cdot)$ converges uniformly (cf. [2, Theorem A.1]), the last term in parentheses in (31) goes to zero as $t \to \infty$ because of equation (20).

Therefore, the left hand side of equation (23) goes to zero as $t \to \infty$. Thus, the NVI adaptive policy $\overline{g}_{\text{NVI}}$ is average cost optimal. ■

5. Conclusions

In this work we gave sufficient conditions for the optimality of the NVI adaptive policy for a particular production/replacement problem, but the main ideas can be readily put to use in many other applications, and with more general models (see [20] for examples).

Future areas of research should include the effect that "better" (for example in the sense of [21]) parameter estimation algorithms have on the adaptive control policies. Also, from the point of view of implementation, it is important to investigate the behaviour of the adaptive policies for finite time horizons: the results in [19, Chapter 7] tell us that the performance of the adaptive policies differ significantly for finite time.

References

[1] Acosta-Abreu R.S., Hernández-Lerma O., Iterative adaptive control of denumerable state average-cost markov systems, *Control & Cybernetics, 1985,* **14**, 313-322.

[2] Arapostathis A. , Fernandez-Gaucherand E., Marcus S.I., Analysis of an adaptive control scheme for a partially observable markov decision process, In *Proceedings of the 29th IEEE Conference on Decision and Control*, 1990, pages 1438-1444, Honolulu, Hawaii.

[3] Benveniste A., Ruget G., A measure of the tracking capability of recursive stochastic algorithms with constant gains, *IEEE Transactions on Automatic Control*, 1982, **27**, 639-649.

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

[5] Fernandez-Gaucherand E., Arapostathis A., Marcus S.I., On partially observable markov decision processes with an average cost criterion. In *Proceedings of the 28th IEEE Conference on Decision and Control*, 1989, pages 1267-1272, Tampa, Florida.

[6] Fernandez-Gaucherand E., Arapostathis A., Marcus S.I., On the average cost optimality equation and the structure of optimal policies for partially observable markov decision processes. In O. Hernández–Lerma and J.B. Lasserre, editors, *Annals of Operations Research, Special Volume on Markov Decision Processes*, **29**, 1991, pages 439–470.

[7] Foster F.G., Markoff chains with an enumerable number of states and a class of cascade processes, *Proceedings of the Cambridge Philosopkical Society*, 1951, **47**, 77-85.

[8] Hernández–Lerma O., Approximation and adaptive policies in discounted dynamnic programming, *Boletín de la Sociedad Matemática Mexicana*, 1985, **30**, 25-35.

[9] Hernández–Lerma O., Adaptive Markov Control Processes. Spinger Verlag, New York, 1989.

[10] Hernández–Lerma O., Marcus S.I., Adaptive control of discounted markov chains, *Journal of Optimization Theory and Applications*, 1985, **46**, 227–235.

[11] Hernández–Lerma O., Marcus S.I., Adaptive control of markov processes with incomplete state information and unknown parameters, *Journal of Optimization Theory and Applications*, 1987, **52**, 227-241.

[12] Hordijk A., Schweitzer P.J., Tijms H., The asymptotic behaviour of the minimal total expected cost for the denumerable state markov decision model, *Journal of Applied Probability*, 1975, **12**, 298-305.

[13] HÜBNER G., A unified approach to adaptive control of average reward markov decision processes, *OR Spektrum*, 1988, **10**, 161-166.

[14] KARLIN S., TAYLOR H.M., A First Course in Stochastic Processes, Academic Press, New York, 1975.

[15] KEMENY J.G., SNELL J.L., KNAPP A.W., Denumerable Markov Chains, D. van Nostrand, Princeton, New Jersey, 1966.

[16] KUSHNER H.J., CLARK D.S., Stochastic Approximation Methods for Constrained and Unconstrained Systems, Springer Verlag, New York, 1978, Applied Mathematical Sciences, Vol. 26.

[17] MCLEISH D.L., Functional and random central limit theorems for the robbins-monro process, *Journal of Applied Probability*, 1976, **13**, 148-154.

[18] ROSS S.M., Quality control under markovian deterioration, *Management Science*, 1971, **17**, 587-596.

[19] SERNIK E.L., On the Optimal Control of Partially Observed Markov Decision Models, PhD thesis, The University of Texas at Austin, 1991, Electrical and Computer Engineering.

[20] SERNIK E.L.,. MARCUS S.I., On the computation of the optimal cost function for discrete time markov models with partial observations. In O. Hernández-Lerma and J. B. Lasserre, editors, *Annals of Operations Research, Special Volume on Markov Decision Processes*,1991, **29**, 471-512.

[21] SERNIK E.L.,. MARCUS S.I., A parameter estimation procedure for a markovian replacement problem, In *Proceedings of the American Control Conference*, 1991, 2923-2929, Boston, Massachussetts.

[22] WHITE C.C., A markov quality control process subject to partial observation. *Management Science*, 1977, **23**, 843-852.

Optymalność adaptacyjnych polityk z niestacjonarnymi iteracyjnymi wartościami dla częściowo obserwowanych decyzyjnych modeli Markowa

W artykule podano warunki wystarczające optymalności asymptotycznej adaptacyjnych polityk niestacjonarnych iteracji wartości dla częściowo obserwowanego markowowskiego modelu wymiany przy funkcji jakości odzwierciedlającej średni koszt w dłuższym okresie czasu. Użyto podejścia zaproponowanego w [1] i rozszerzono wyniki z tej pracy na przypadek rozważany w artykule posługując się pewnymi rezultatami uzyskanymi w [6] i [2] dla zagadnienia wymiany.

Оптимальность адаптивной политики с нестационарными итерациями значений для частично наблюдаемых марковских моделей принятия решений

В статье даются достаточные условия асимптотической оптимальности адаптивной политики нестационарных итераций значений для частично наблюдаемой марковской модели обмена при функции качества отражающей средние затраты за длительный период времени. Используется подход предложенный в паботе [1], результаты которой обобщены на случай рассматрибаемый в данной статье на основие работ [2] и [6] для задачи обмена.