ANALYSIS OF AN ADAPTIVE CONTROL SCHEME FOR A PARTIALLY OBSERVED CONTROLLED MARKOV CHAIN*

Emmanuel Fernández-Gaucherand†, Aristotle Arapostathis‡, and Steven I. Marcus§.

Revised: December 1991

Corresponding Author: A. Arapostathis

Key Words: Stochastic adaptive control, Markov chains, partial observations.

Abstract

We consider an adaptive finite state controlled Markov chain with partial state information, motivated by a class of replacement problems. We present parameter estimation techniques based on the information available after actions that reset the state to a known value are taken. We prove that the parameter estimates converge w.p.1 to the true (unknown) parameter, under the feedback structure induced by a certainty equivalent adaptive policy. We also show that the adaptive policy is self-optimizing, in a long-run average sense, for any (measurable) sequence of parameter estimates converging w.p.1 to the true parameter.

* This work was supported in part by the Texas Advanced Technology Program under Grant No. 003658-093, in part by the Air Force Office of Scientific Research under Grant AFOSR-91-0033 and in part by the National Science Foundation under Grant CDR-8803012.

† Systems and Industrial Engineering Department, The University of Arizona, Tucson, Arizona 85721.

‡ Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78712-1084.

§ Department of Electrical Engineering and Systems Research Center, The University of Maryland, College Park, Maryland 20742.

I. Introduction

In recent years, there has been a considerable amount of work in stochastic adaptive control [10]–[11]. However, aside from results for linear systems, little progress has been made on problems with incomplete or noisy state observations. An initial step in this direction was taken in [1], where the adaptive estimation of the state of a finite state Markov chain, with incomplete state information, and with the state transition probabilities depending on unknown parameters, is studied. This adaptive estimation problem is that of computing recursive estimates of the conditional probability vector of the state at time t, given all the past observations, when the transition matrix P is not completely known, i.e., it depends on a vector of unknown parameters θ — this dependence is expressed as $P(\theta)$. In [1] we use the previously derived recursive filter for the conditional probabilities, and simultaneously recursively estimate the parameters, using the most recent parameter estimates to update the filter. This adaptive estimation algorithm is then analyzed via the Ordinary Differential Equation (ODE) Method [12]–[13]. The convergence of the recursive parameter estimates is established, and optimality of the adaptive state estimator is proved, in a long-run average sense.

In [7]–[8], we began to investigate the application of similar techniques to the control of adaptive finite state Markov chains with incomplete observations. One interesting set of problems for which some results are available when the parameters are *known* are those involving quality control, replacement, and repair of a unit in a manufacturing system or communication network [9], [15], [18]. We formulated the adaptive version of a problem of this type in the above references; however, the presence of feedback makes this problem much more difficult than that of [1]. Discontinuities in the optimal control strategies lead to averaged ODE's with discontinuous right-hand sides that cannot be handled by currently available methods.

In this paper we present parameter estimation techniques based on the information available after actions that reset the state to a known value are taken. At these times, the (augmented) state process *regenerates,* its future evolution becoming independent of the past. We prove (by means of the ODE method) w.p.1 convergence of the parameter estimates to the true (unknown) parameter θ_0, for a parameter estimation scheme of this type. Then, given *any* sequence of parameter estimates which converges w.p.1 to θ_0, and which is measurable with respect to the filtration generated by the observations, we show that a certainty equivalent adaptive policy is *self-optimizing.* The latter is obtained by an analysis which uses the known (threshold) structure of optimal policies for problems with *known* parameters. Our analysis is of particular interest since the nice formalism recently presented in [17] cannot be directly applied in the present situation: here the state is only partially observed and the optimal policy is not a continuous function of θ.

The methodology exposed in the analysis relies largely on the w.p.1 convergence to θ_0 of the parameter estimates, and the continuity in the parameterization of quantities in the model, like $P(\theta)$ and the solutions to the corresponding optimality equations. Hence, this methodology is also applicable to a more general situation than the one presented here; see [6]. In addition, we note that the feedback structure induced by our adaptive policy obviates the need for, e.g. forced choice schemes, c.f. [11].

II. A Partially Observed Binary Replacement Problem

Consider a situation in which a system, such as a machine, production process, or computer communications network can fail. The (core) state X_t of the system can either be *good* (0), or *failed* (1); let $\mathbf{X} := \{0,1\}$. The available control actions (or decisions) are to *operate* the system in its current condition (0), or to *reset/replace* the system to an *as new* condition (1); let $\mathbf{U} := \{0,1\}$. Assume for the moment that there is an underlying probability space $(\Omega, \mathcal{B}, \mathcal{P})$. The process $\{X_t\}_{t\in\mathbb{N}_0}$ is modeled as a controlled finite state Markov chain, where we have that

$$\begin{aligned}\mathcal{P}\bigl\{X_{t+1} = j \bigm| X_t = i, X_{t-1}, \ldots, X_0;\, U_t = u, U_{t-1}, \ldots, U_0\bigr\}\\ = [P(u)]_{i,j}\,; \qquad t \in \mathbb{N}_0 := \{0,1,2,\ldots\},\end{aligned} \tag{2.1}$$

and the state transition probability matrices are given as

$$P(0) = \begin{bmatrix} 1-\theta & \theta \\ 0 & 1 \end{bmatrix}; \qquad P(1) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}. \tag{2.2}$$

Here $\theta \in [0,1]$ gives the *failure rate* of the system. Only imperfect observations of $\{X_t\}_{t\in\mathbb{N}_0}$ are available in the form of a random process $\{Y_t\}_{t\in\mathbb{N}}$; Y_t gives a correct observation of X_t with probability q, when $U_{t-1} = 0$, whereas if $U_{t-1} = 1$ then $Y_t = X_t = 0$. More precisely, $Y_t \in \mathbf{Y} = \{0,1\}$ and

$$\begin{aligned}\mathcal{P}\bigl\{Y_{t+1} = i \bigm| Y_t, \ldots, Y_1;\, X_{t+1} = i, X_t, \ldots, X_0;\, U_t = 0, U_{t-1}, \ldots, U_0\bigr\}\\ = \mathcal{P}\bigl\{Y_{t+1} = i \bigm| X_{t+1} = i;\, U_t = 0\bigr\} =: q\,, \qquad t \in \mathbb{N}_0\,.\end{aligned} \tag{2.3}$$

It suffices to consider only $0.5 \le q \le 1$. The cases $q = 0.5$ and $q = 1$ correspond to the *completely unobserved* and *completely observed* situations, respectively; we restrict our analysis to the situation of strict partial observability, i.e., $q < 1$. The one-step cost $c(x,u)$ is defined as $c(0,0) = 0$, $c(1,0) = C$, $c(x,1) = R$, where $0 < C < R$. Probability distribution vectors on $\mathbf{X}$ are elements of $\mathbf{\Delta} := \bigl\{p \in \mathbb{R}^2 : p = [1-\rho\,,\ \rho]\,,\ 0 \le \rho \le 1\bigr\}$. Thus, each $p \in \mathbf{\Delta}$ can be uniquely identified with a scalar $\rho \in [0,1]$, as indicated. Initially, there is a given probability $0 \le \rho_0 \le 1$ that the system is failed, an action is taken, and the state evolves according to (2.2); a first observation is received, another action is taken; and so on.

An (admissible) control *law, policy,* or *strategy* π is a rule for selecting the actions U_t, based on $h_t = (\rho_0, U_0, Y_1, \ldots, Y_{t-1}, U_{t-1}, U_t)$, where h_t is the available information at time t. The canonical sample path space is $\mathbf{\Omega} = \mathbf{X} \times \mathbf{U} \times (\mathbf{X} \times \mathbf{Y} \times \mathbf{U})^\infty$, and $\mathcal{B}$ denotes the Borel σ-algebra obtained by endowing $\mathbf{\Omega}$ with the discrete topology. Then to each admissible strategy π and $0 \le \rho_0 \le 1$, we associate the average cost

$$J(\pi, \rho_0) := \limsup_{n\to\infty} \; E^\pi_{\rho_0}\left[\frac{1}{n}\sum_{t=0}^{n-1} c(X_t, U_t)\right], \tag{AC}$$

where $E^\pi_{\rho_0}$ is the expectation with respect to an appropriate marginal of the (unique) probability measure $\mathcal{P}^\pi_{\rho_0}$ on $\mathcal{B}$ induced by ρ_0 and the strategy π; see [2], [10]. The *optimal* (AC) *control (or decision) problem* is that of selecting a strategy such that the average cost is minimized, over all admissible strategies. The optimal (AC) cost function is defined as $\Gamma(\rho_0) := \inf_\pi \big\{ J(\pi, \rho_0) : \pi \text{ is an admissible strategy} \big\}$, for $0 \le \rho_0 \le 1$.

A. Information States

It is well known that the conditional probability distribution process, whose i^{th} component is given by

$$p^{(i)}_t := \mathcal{P}^\pi_{\rho_0}\big\{X_t = i \;\big|\; Y_t, \ldots, Y_1; U_{t-1}, \ldots, U_0\big\}, \qquad t \in \mathbb{N}, \qquad p_0 := [1-\rho_0, \rho_0],$$

constitutes an *information state* (or statistic sufficient for control) [2], [4], [5], [10], [11]; for this problem, it can be written as $p_t = [1-\rho_t, \rho_t]$, where ρ_t is the conditional probability of the process being in the failed state.

A *separated* strategy is a sequence of maps $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$, where $\pi_t : [0,1] \to \mathbf{U}$. When $\pi_t(\cdot) = \pi(\cdot)$ for all values of t, then the policy is said to be stationary. Then the partially observed, average cost problem is equivalent (i.e., equal minimum costs for each ρ_0) to the *completely observed* problem, with state ρ_t and state space $[0,1]$, of finding a separated admissible strategy which minimizes

$$\overline{J}(\pi, \rho_0) := \limsup_{n\to\infty} \; E^\pi_{\rho_0}\left[\frac{1}{n}\sum_{t=0}^{n-1} \overline{c}(\rho_t, U_t)\right],$$

where $\overline{c}(\rho, u) = (1-\rho)c(0,u) + \rho c(1,u)$. Note that $\overline{c}(\rho, 0) = \rho C$ and $\overline{c}(\rho, 1) = R$. Using Bayes' rule, it is easily shown that ρ_t can be computed recursively, as follows:

$$\rho_{t+1} = T(1, \rho_t, U_t) Y_{t+1} + T(0, \rho_t, U_t)(1 - Y_{t+1}), \tag{2.4}$$

where

$$V(1, \rho, 0) = (1-q)(1-\rho)(1-\theta) + q[\rho(1-\theta) + \theta] = 1 - V(0, \rho, 0), \tag{2.5}$$

$$V(1, \rho, 1) = 0, \qquad V(0, \rho, 1) = 1, \tag{2.6}$$

$$T(0,\rho,0) = \frac{(1-q)[\rho(1-\theta)+\theta]}{V(0,\rho,0)}\,, \qquad T(1,\rho,0) = \frac{q[\rho(1-\theta)+\theta]}{V(1,\rho,0)}\,, \tag{2.7}$$

$$T(y,\rho,1) = 0\,; \qquad y = 0,1, \quad \rho \in [0,1]\,. \tag{2.8}$$

Here $V(y,\rho,u)$ is interpreted as the (one-step ahead) conditional probability of the observation being y given the decision u and an *a priori* probability ρ of the state being failed. Likewise, $T(y,\rho,u)$ is interpreted as the *a posteriori* conditional probability of the unit being failed given that decision u was made, observation y obtained, and an *a priori* probability ρ. Let $I[A]$ denote the indicator function of the event A. A well known property of the process $\{\rho_t\}_{t=0}^{\infty}$ is the following [4].

Lemma 2.1. *$\{\rho_t\}_{t=0}^{\infty}$ is a controlled Markov process, and its state transition probabilities are given by*

$$\begin{aligned}&\mathcal{P}^{\pi}_{\rho_0}\{\rho_{t+1} \in B \mid \rho_t = \rho;\, U_t = u\} \\ &\qquad = \sum_{y\in\mathbf{Y}} V(y,\rho,u)\, I\big[T(y,\rho,u) \in B\big] =: \mathcal{K}(B \mid \rho, u)\,,\end{aligned} \tag{2.9}$$

for all (Borel) subsets B of $[0,1]$.

III. The Structure of Optimal Policies.

Consider the optimal control problem corresponding to each parameter value $\theta \in [0,1]$. Then, the existence of solutions to the corresponding (average cost) optimality equation follows from the existence of a reset/repair action [6], [9], [15]–[16]. We summarize these results as follows; dependence on θ is made explicit.

Theorem 3.1. *Assume $q \in [0.5, 1)$, $\theta \in [0,1]$.*

(i) There exist a constant $0 \le \Gamma^*_\theta \le R$ and a concave, nondecreasing map $h_\theta : [0,1] \to [0,R]$, with $h_\theta(0) = 0$, such that

$$\Gamma^*_\theta + h_\theta(\rho) = \min\Big\{ f_\theta(\rho)\,;\; R \Big\}\,, \tag{3.1}$$

where

$$f_\theta(\rho) := \rho C + \sum_{y=0}^{1} V(y,\rho,0;\theta)\, h_\theta(T(y,\rho,0;\theta)). \tag{3.2}$$

(ii) Any stationary separated policy that achieves the minimum in (3.1) is average cost optimal; the minimum cost is Γ^*_θ, for any value of ρ_0.

The following will be used in the sequel, and its proof is given in the Appendix.

Corollary 3.1. *Assume $q \in [0.5, 1)$ and $\theta \in [0, 1]$.*

(i) Any concave and nondecreasing solution $h_\theta(\cdot)$ of (3.1) is continuous on $[0, 1]$;

(ii) furthermore, there is only one such solution satisfying $h_\theta(0) = 0$.

Henceforth, the dependence of $\mathcal{P}^\pi_{\rho_0}$ on ρ_0 will be omitted, in view of Theorem 3.1. Equation (3.1) can then be used to determine the structure of the optimal policies [6], [9], [15].

Theorem 3.2. *Assume $q \in [0.5, 1)$ and $\theta \in (0, 1]$.*

(i) If

$$\frac{C(1+\theta)}{\theta} \le R \iff \frac{C}{R-C} \le \theta\,,$$

then the policy "operate ($U_t = 0$) for all $\rho_t \in [0, 1]$" is average cost optimal.

(ii) If

$$R < \frac{C(1+\theta)}{\theta} \iff \theta < \frac{C}{R-C}\,,$$

then there exists a threshold policy which is average cost optimal; i.e., there exists $\alpha(\theta) \in (0, 1)$ such that it is optimal to operate ($U_t = 0$) for $\rho_t \in [0, \alpha(\theta))$, and to repair ($U_t = 1$) for $\rho_t \in [\alpha(\theta), 1]$.

IV. The Adaptive Binary Replacement Problem

If the parameter θ is unknown, we cannot compute ρ_t, nor can we directly solve the optimal control problem. The *enforced certainty equivalence* approach which we will adopt involves simultaneously computing recursive estimates $\hat{\theta}_t$ of the unknown parameter, and $\hat{\rho}_t$ of the information state, and using the latest available parameter estimate in the filtering equation (2.4) to compute the next estimate $\hat{\rho}_{t+1}$; the decision U_t is made taking $\hat{\theta}_t$ and $\hat{\rho}_t$ as if they were the true (correct) values. Let $\Theta_\delta := [\delta, \delta']$ be the parameter set in which $\hat{\theta}_t$ is allowed to take its values, where δ is an arbitrarily small positive number and $\delta' = \min\big\{1, \frac{C}{R-C} - \delta\big\}$. For decision-making, we define the set $\mathcal{OP} = \big\{\pi(\cdot\,;\theta)\big\}_{\theta\in\Theta_\delta}$ of optimal threshold policies described above, parameterized by θ. Thus, we conclude from Theorem 3.2 (ii) that $0 < \alpha(\theta) < 1$, for each $\theta \in \Theta_\delta$, where $\alpha(\theta)$ denotes the dependence of the threshold on θ. We also let θ_0 denote the (unknown) true value of the parameter, which we assume to be constant and an element of the interior of Θ_δ. The following result, on the continuity in θ of the optimal cost, the value function and the threshold, is proved in [3, Theorem A.1].

Theorem 4.1. *Assume $q \in [0.5, 1)$. Let $0 < \delta < 1$. Then for $\theta \in \Theta_\delta$, we have that:*

(i) the pair $(\Gamma^*_\theta, h_\theta)$ is continuous in θ;

(ii) there exists a unique $\alpha(\theta) \in (0, 1)$ such that $f_\theta(\alpha(\theta)) = R$;

(iii) $\alpha(\cdot)$ *is continuous on* Θ_δ.

Observe that by Theorem 4.1 (iii) and since Θ_δ is compact, there is a number $\alpha^* < 1$ such that, for all $\theta \in \Theta_\delta$, $0 < \alpha(\theta) \le \alpha^*$.

A. Adaptive Policy.

Given a sequence of estimates $\{\hat{\theta}_t\}_{t=0}^{\infty}$ of θ_0, compute the control action at each time $t \in \mathbb{N}_0$ by

$$U_t = \pi(\hat{\rho}_t; \hat{\theta}_t), \qquad \pi(\cdot\,;\,\cdot) \in \mathcal{OP}, \tag{4.1}$$

where the conditional probability estimate is computed recursively via

$$\begin{aligned}\hat{\rho}_{t+1} =& T(1, \hat{\rho}_t, \pi(\hat{\rho}_t; \hat{\theta}_t); \hat{\theta}_{t+1}) \cdot Y_{t+1} \\ &+ T(0, \hat{\rho}_t, \pi(\hat{\rho}_t; \hat{\theta}_t); \hat{\theta}_{t+1}) \cdot (1 - Y_{t+1}), \qquad \hat{\rho}_0 = \rho_0.\end{aligned} \tag{4.2}$$

We will denote by π^a the policy given by (4.1) and (4.2).

B. Parameter Estimation.

There are a number of ways to compute the estimates $\hat{\theta}_t$; we consider here only recursive schemes. One method, discussed in [7]–[8], updates the parameter estimate $\hat{\theta}_t$ at each time step t, and is similar to that used for adaptive estimation in [1]. However, the analysis of convergence is very difficult, due to the complex feedback structure induced. We concentrate here on algorithms which update $\hat{\theta}_t$ after each repair. The advantage of this approach is that when a repair event occurs, the state of the system is reset to the "as new" state, and thus the processes of interest are identically distributed between these events. On the other hand, the convergence rate may be too slow, and thus some forcing may be needed to accelerate the convergence. Algorithms that take advantage of analogous *regenerative* behavior in some queueing problems, by updating after each busy period, have been presented in [13]. The next result is a direct consequence of [3, Theorem A.2].

Theorem 4.2. *Under the adaptive policy* π^a*, regeneration occurs infinitely often (i.o.), i.e.,*

$$\mathcal{P}^{\pi^a}\big\{U_t = 1, \text{ i.o.}\big\} = 1\,.$$

Let τ_k be the k^{th} repair time under π^a (i.e., the k^{th} time such that $U_t = 1$). Since $U_{\tau_k} = 1$, then $X_{\tau_k+1} = 0$, $U_{\tau_k+1} = 0$, and Y_{τ_k+2} is observed. Hence, the state is known perfectly at $\tau_k + 1$ and the observations $\big\{Y_{\tau_k+2} : \ k = 1, 2, \ldots\big\}$ form an independent identically distributed (i.i.d.) sequence of Bernoulli random variables, with $\mathcal{P}^{\pi^a}\{Y_{\tau_k+2} = j\} = \lambda_j(\theta_0)$, $j = 0, 1$, where

$$\lambda_1(\theta) := (1-\theta)(1-q) + \theta q = 1 - \lambda_0(\theta)\,. \tag{4.3}$$

This sequence provides information about the transition from $X_{\tau_k+1} = 0$ to X_{τ_k+2}, and thus can be used to estimate θ_0. Define $\overline{Y}_k := Y_{\tau_k+2}$. The sequence $\left\{\overline{Y}_k\right\}_{k=0}^{\infty}$ is i.i.d., its distribution depending only on the true parameter θ_0 and the reliability of the measuring device (q).

Note that by Theorem 4.2 and the strong law of large numbers we have that

$$\frac{1}{n}\sum_{k=1}^{n}\overline{Y}_k \underset{n\to\infty}{\longrightarrow} \lambda_1(\theta_0), \qquad \mathcal{P}^{\pi^a}\text{–a.s.}. \tag{4.4}$$

Let $\hat{\overline{\theta}}_n := \hat{\theta}_{\tau_n+2}$. Then, setting

$$\lambda_1(\hat{\overline{\theta}}_n) = \frac{1}{n}\sum_{k=1}^{n}\overline{Y}_k\,,$$

where $\lambda_1(\cdot)$ is defined in (4.3), we obtain a sequence of strongly consistent parameter estimates $\left\{\hat{\overline{\theta}}_n\right\}$. Also, a *prediction error-based* algorithm can be formulated. Since the observations take only the values $\{0,1\}$, then the prediction error in this case is

$$\epsilon_n(\theta) = \overline{Y}_n - \lambda_1(\theta)\,. \tag{4.5}$$

However, in order to have $\hat{\overline{\theta}}_n \in \Theta_\delta$, a projection mechanism is required. A stochastic approximation-type recursive algorithm which is designed to minimize $E^{\pi^a}[\frac{1}{2}\epsilon_n(\theta)^2]$ is then

$$\hat{\overline{\theta}}_{n+1} = \Pi_{\Theta_\delta}\left(\hat{\overline{\theta}}_n + \tfrac{1}{n+1}\,R_{n+1}^{-1}\psi_n\epsilon_n(\hat{\overline{\theta}}_n)\right), \qquad \hat{\overline{\theta}}_0 \in \Theta_\delta, \tag{4.6a}$$

where the map Π_{Θ_δ} is a projection into the interior of Θ_δ. Also, R_n can be computed in different ways, e.g. if $R_n = (2q-1)^2$, then we obtain a recursive (and projected) version of the scheme obtained from (4.4) above. We choose to use

$$\begin{aligned} R_{n+1} &= R_n + \tfrac{1}{n+1}\,\left(\psi_n^2 - R_n\right), \qquad\qquad R_1 = 1,\\ \psi_n &= -\tfrac{\partial}{\partial\theta}\,\epsilon_n(\theta)\Big|_{\theta=\hat{\overline{\theta}}_n} = \tfrac{\partial}{\partial\theta}\,\lambda_1(\theta) = 2q-1. \end{aligned} \tag{4.6b}$$

The following can then be shown using the techniques in [12], [13].

Theorem 4.3. *Consider the algorithm (4.6). The sequence* $\left\{\hat{\overline{\theta}}_n\right\}_{n=0}^{\infty}$ *converges* $\mathcal{P}^{\pi^a}$–a.s., *as* $n\to\infty$, *to the set of limit points of the ODE*

$$\begin{aligned} \dot{\theta}(t) &= -R^{-1}(t)(2q-1)^2(\theta(t)-\theta_0)\,,\\ \dot{R}(t) &= (2q-1)^2 - R(t)\,. \end{aligned} \tag{4.7}$$

Since θ_0 is assumed to lie in the interior of Θ_δ, all solutions of the ODE (4.7) leave the interior of Θ_δ invariant and thus the projection operator Π_{Θ_δ} need not be considered in the averaged equations. It is straightforward to show that (4.7) is globally asymptotically stable with unique limit point θ_0. In the natural way, we define $\hat{\theta}_t$ to be constant between updates: $\hat{\theta}_t := \hat{\overline{\theta}}_n$, $t \in \left\{\tau_n+2,\, \tau_n+3, \ldots,\, \tau_{n+1}+1\right\}$. We thus have the following result, which is a direct consequence of Theorem 4.2.

Corollary 4.1. *Assume $q \in (0.5, 1)$. Then the sequence $\{\hat{\theta}_t\}_{t=0}^{\infty}$ converges to θ_0, as $t \to \infty$,* $\mathcal{P}^{\pi^a}$–a.s..

Remark 4.1: Let π be any separated policy satisfying $\pi_t(0) = 0$, for all $t \in \mathbb{N}_0$, and $\mathcal{P}^{\pi}\{U_t = 1, \text{ i.o.}\} = 1$. Then the results above will also hold if π is used instead of π^a.

V. Average Cost Optimality of the Adaptive Policy

We examine next the long-run average performance of the adaptive policy π^a given by (4.1) and (4.2). Let $\mathcal{F}_t$ be the σ-algebra generated by the observations up to time t, i.e., $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$. Note that $\{\hat{\theta}_t\}$ of Corollary 4.1 satisfies the following conditions:

(E1) $\hat{\theta}_t$ is $\mathcal{F}_t$-measurable, and $\hat{\theta}_t \in \Theta_\delta$, for all $t \in \mathbb{N}_0$;

(E2) $\hat{\theta}_t \to \theta_0$, $\mathcal{P}^{\pi^a}$–a.s..

Consider also the weaker condition:

(E2′) $\hat{\theta}_t \to \theta_0$, in probability under $\mathcal{P}^{\pi^a}$.

Let $\{\hat{\theta}_t\}_{t=0}^{\infty}$ be *any* sequence of parameter estimates satisfying (E1) and (E2′); we will show that the corresponding adaptive policy π^a is *self-optimizing,* i.e., $\overline{J}(\pi^a, \rho_0) = \Gamma^*_{\theta_0}$, for all $0 \leq \rho_0 \leq 1$. In the case where $\{\hat{\theta}_t\}_{t=0}^{\infty}$ satisfies (E2), we will show the stronger sample path result

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \overline{c}(\rho_t, U_t) = \Gamma^*_{\theta_0}, \quad \mathcal{P}^{\pi^a}\text{–a.s.}\,. \tag{5.1}$$

The method we use to verify these self-optimizing properties of π^a is motivated by techniques in [14] and [17]. However, the verification here does not fit in the same framework, due to (a) discontinuity of $\pi(\cdot\,;\cdot) \in \mathcal{OP}$ in both its arguments and (b) the fact that the cost $\overline{c}(\rho, u)$ is an explicit function of u. We have that $T(y, \rho, u; \theta)$ is continuous in θ. Using this and the fact that regeneration occurs infinitely often, the following is shown in the Appendix.

Lemma 5.1. *If $\hat{\theta}_t \to \theta_0$, as $t \to \infty$, in probability under $\mathcal{P}^{\pi^a}$ ($\mathcal{P}^{\pi^a}$–*a.s.*), then $|\hat{\rho}_t - \rho_t| \to 0$, as $t \to \infty$, in probability under $\mathcal{P}^{\pi^a}$ ($\mathcal{P}^{\pi^a}$–*a.s.*).*

Then, we have the following.

Theorem 5.1. *Assume $q \in (0.5, 1)$.*

(i) If $\{\hat{\theta}_t\}_{t=0}^{\infty}$ satisfies (E1) and (E2′), then π^a is self-optimizing.

(ii) If in addition $\{\hat{\theta}_t\}_{t=0}^{\infty}$ satisfies (E2), then π^a is self-optimizing in a sample-path sense, i.e., (5.1) holds.

PROOF: (i) Let $\Phi_\theta(\cdot\,,\cdot)$ denote Mandl's discrepancy function, corresponding to the parameter value $\theta \in \Theta_\delta$, i.e., for $\rho \in [0,1]$ and $u \in \mathbf{U}$

$$\Phi_\theta(\rho, u) := \overline{c}(\rho, u) + \sum_{y\in\mathbf{Y}} V(y,\rho,u;\theta) h_\theta(T(y,\rho,u;\theta)) - \Gamma^*_\theta - h_\theta(\rho).$$

Then by (2.5)–(2.8), Corollary 3.1 and Theorem 4.1, $\Phi_\theta(\rho, u)$ is continuous in both $\rho \in [0,1]$ and $\theta \in \Theta_\delta$. Furthermore, since Θ_δ is compact, then $\Phi_\theta(\rho, u)$ is uniformly continuous and bounded in $(\rho,\theta) \in [0,1] \times \Theta_\delta$; thus, $\Phi_{\hat{\theta}_t}(\hat{\rho}_t, u)$ is uniformly integrable, for each $u \in \mathbf{U}$. Therefore, for each $u \in \mathbf{U}$, we have

$$\left|\Phi_{\hat{\theta}_t}(\hat{\rho}_t, u) - \Phi_{\theta_0}(\rho_t, u)\right| \underset{t\to\infty}{\longrightarrow} 0\,, \qquad L_1(\mathcal{P}^{\pi^a}),$$

and since $\mathbf{U}$ is finite,

$$E^{\pi^a}\left\{\Phi_{\theta_0}(\rho_t, \pi(\hat{\rho}_t;\hat{\theta}_t))\right\} \underset{t\to\infty}{\longrightarrow} 0, \tag{5.2}$$

where we used the fact that $\Phi_{\hat{\theta}_t}(\hat{\rho}_t, \pi(\hat{\rho}_t;\hat{\theta}_t)) = 0$, since $\pi(\cdot\,;\theta) \in \mathcal{OP}$ minimizes the optimality equation (3.1), for the parameter value $\theta \in \Theta_\delta$. The result then follows from (5.2); see [2], [10], [14], [17].

(ii) If the convergence is in the stronger $\mathcal{P}^{\pi^a}$–a.s. sense, then similarly as above, we obtain that

$$\Phi_{\theta_0}(\rho_t, \pi(\hat{\rho}_t;\hat{\theta}_t)) \underset{t\to\infty}{\longrightarrow} 0, \qquad \mathcal{P}^{\pi^a}\text{–a.s.}\,,$$

from which the result follows. □

References

[1] A. Arapostathis and S. I. Marcus, "Analysis of an Identification Algorithm Arising in the Adaptive Estimation of Markov Chains," *Mathematics of Control, Signals and Systems,* vol. 3, 1990, pp. 1–29.

[2] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey," submitted for publication.

[3] A. Arapostathis, E. Fernández-Gaucherand, and S. I. Marcus, "Analysis of an Adaptive Control Scheme for a Partially Observed Controlled Markov Chain," *Proc. 29th IEEE Conf. Decision and Control,* Honolulu, HI, 1990, pp. 1438–1444.

[4] K. J. Åström, "Optimal Control of Markov Processes with Incomplete State Information," *J. Math. Anal. Appl.,* vol. 10, 1965, pp. 174–205.

[5] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models,* Prentice-Hall, Englewood Cliffs, NJ, 1987.

[6] E. Fernández-Gaucherand, "Controlled Markov Processes on the Infinite Planning Horizon: Optimal & Adaptive Control," Ph.D. Dissertation, The University of Texas at Austin, August 1991.

[7] E. Fernández-Gaucherand, A. Arapostathis and S.I. Marcus, "On the Adaptive Control of Partially Observable Markov Decision Processes," *Proc. 27th IEEE Conf. Decision and Control,* Austin, TX, 1988, pp. 1204–1210.

[8] E. Fernández-Gaucherand, A. Arapostathis and S. I. Marcus, "On the Adaptive Control of a Partially Observable Binary Markov Decision Process," in *Advances in Computing and Control,* W. A. Porter, S. C. Kak, J. L. Aravena, eds., Lecture Notes in Control and Information Sciences, vol. 130, Springer-Verlag, Berlin, 1989, pp. 217–228.

[9] E. Fernández-Gaucherand, A. Arapostathis and S. I. Marcus, "On the Average Cost Optimality Equation and the Structure of Optimal Policies for Partially Observable Markov Decision Processes," *Annals of Operations Research,* vol. 29, 1991, pp. 439–470.

[10] O. Hernández-Lerma, *Adaptive Markov Control Processes,* Springer Verlag, New York, 1989.

[11] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control,* Prentice-Hall, Englewood Cliffs, NJ, 1986.

[12] H. J. Kushner, "An Averaging Method for Stochastic Approximations with Discontinuous Dynamics, Constraints, and State Dependent Noise," in *Recent Advances in Statistics,* Rizvi, Rustagi and Siegmund, Eds., Academic Press, New York, 1983, pp. 211–235.

[13] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems,* Springer-Verlag, New York, 1978.

[14] P. Mandl, "Estimation and Control in Markov Chains," *Adv. Appl. Prob.,* vol. 6, 1974, pp. 40–60.

[15] M. Ohnishi, H. Mine and H. Kawai, "An Optimal Inspection and Replacement Policy Under Incomplete State Information: Average Cost Criterion," in *Stochastic Models in Reliability Theory,* S. Osaki and Y. Hatoyama, eds., Lecture Notes in Econ. and Math. Systems No. 235, Springer-Verlag, Berlin, 1984, pp. 187–197.

[16] L. K. Platzman, "Optimal Infinite-Horizon Undiscounted Control of Finite Probabilistic Systems," *SIAM J. Control Optim.,* Vol. 18, 1980, pp. 362–380.

[17] A. Shwartz and A. M. Makowski, "Comparing Policies in Markov Decision Processes: Mandl's Lemma Revisited," *Math. Oper. Res.,* vol. 15, 1990, pp. 155–174.

[18] C. C. White, "A Markov Quality Control Process Subject to Partial Observation," *Mang. Sci.,* Vol. 23, 1977, pp. 843–852.

Appendix

For ease of notation, we will write

$$V^y(\rho;\theta) := V(y,\rho,0;\theta), \qquad T^y(\rho;\theta) := T(y,\rho,0;\theta)\,, \tag{A.1}$$

We quote the following useful result from [3].

Lemma A.1. *Let $q \in [0.5, 1)$, $\rho \in [0,1]$ and $\theta \in (0,1]$. Then*

- *(i) $T^1(\rho,\theta) \geq T^0(\rho,\theta)$ and the inequality is strict for $q \in (0.5,1)$, $\rho \in [0,1)$ and $\theta \in (0,1)$.*
- *(ii) $T^y(\rho,\theta)$ is monotone nondecreasing with respect to both ρ and θ.*
- *(iii) $\big|V^y(\rho,\theta) - V^y(\rho',\theta')\big| \leq |\rho-\rho'| + |\theta-\theta'|$.*
- *(iv) $\big|T^y(\rho,\theta) - T^y(\rho',\theta')\big| \leq \left(\frac{q}{1-q}\right)\big\{(1-\theta)|\rho-\rho'| + (1-\rho')|\theta-\theta'|\big\}$.*
- *(v) The iterates of $T^1(\cdot\,,\theta)$ converge uniformly and monotonically to 1.*

Proof of Corollary 3.1: (i) Let $\theta \in (0,1]$ be fixed. Continuity of $h_\theta(\cdot)$ on $(0,1]$ is immediate since $h_\theta(\cdot)$ is concave and nondecreasing. To show that it is continuous at 0 observe that $T^y(0;\theta) > 0$, for $y = 0,1$, and thus $T^y(\cdot;\theta)$ maps a neighborhood of 0 into $(0,1]$. Thus, the continuity of $T^y(\cdot;\theta)$ and $V^y(\cdot;\theta)$ on $[0,1]$ (see Lemma A.1) along with that of $h_\theta(\cdot)$ on $(0,1]$ imply the continuity of $f_\theta(\cdot)$ on $[0,1]$, which in turn implies, in view of (3.1), that $h_\theta(\cdot)$ is continuous on $[0,1]$.

(ii) Now suppose $h_\theta^{(1)}$ and $h_\theta^{(2)}$ are any two solutions of (3.1), satisfying $h_\theta^{(1)}(0) = h_\theta^{(2)}(0) = 0$, and let $\tilde\rho \in [0,1]$ satisfy

$$h_\theta^{(1)}(\tilde\rho) - h_\theta^{(2)}(\tilde\rho) = \sup_{\rho\in[0,1]} \big\{h_\theta^{(1)}(\rho) - h_\theta^{(2)}(\rho)\big\}\,. \qquad (A.2)$$

We distinguish two cases.

First, suppose that $h_\theta^{(2)}(\tilde\rho) \neq R - \Gamma_\theta^*$. With $f_\theta^{(1)}(\cdot)$ and $f_\theta^{(2)}(\cdot)$ suitably defined, we obtain

$$\begin{aligned} h_\theta^{(1)}(\tilde\rho) - h_\theta^{(2)}(\tilde\rho) &= \min\big\{f_\theta^{(1)}(\tilde\rho)\,,\, R\big\} - f_\theta^{(2)}(\tilde\rho) \leq f_\theta^{(1)}(\tilde\rho) - f_\theta^{(2)}(\tilde\rho) \\ &= \sum_{y=0}^{1} V^y(\tilde\rho;\theta)\,\big\{h_\theta^{(1)}\big(T^y(\tilde\rho;\theta)\big) - h_\theta^{(2)}\big(T^y(\tilde\rho;\theta)\big)\big\}\,. \end{aligned}$$

Since $V^0(\tilde\rho;\theta) + V^1(\tilde\rho;\theta) = 1$ and $V^1(\tilde\rho;\theta) > 0$, we conclude that $h_\theta^{(1)}(\tilde\rho) - h_\theta^{(2)}(\tilde\rho) = h_\theta^{(1)}\big(T^1(\tilde\rho;\theta)\big) - h_\theta^{(2)}\big(T^1(\tilde\rho;\theta)\big)$, and thus (A.2) still holds if we replace $\tilde\rho$ with $T^1(\tilde\rho;\theta)$. By induction, for each $n \in \mathbb{N}$,

$$h_\theta^{(1)}\big((T^1)^n(\tilde\rho;\theta)\big) - h_\theta^{(2)}\big((T^1)^n(\tilde\rho;\theta)\big) = \sup_{\rho\in[0,1]} \big\{h_\theta^{(1)}(\rho) - h_\theta^{(2)}(\rho)\big\}\,. \qquad (A.3a)$$

Second, suppose that $h_\theta^{(2)}(\tilde\rho) = R - \Gamma_\theta^*$. Observe that $h_\theta^{(1)}(\tilde\rho) - h_\theta^{(2)}(\tilde\rho) \geq 0$ (since $h_\theta^{(1)}(0) = h_\theta^{(2)}(0)$) and therefore, necessarily, $h_\theta^{(1)}(\tilde\rho) = R - \Gamma_\theta^*$. Invoking the fact that $h_\theta(\cdot)$ is nondecreasing, we conclude that

$$h_\theta^{(1)}\big((T^1)^n(\tilde\rho;\theta)\big) - h_\theta^{(2)}\big((T^1)^n(\tilde\rho;\theta)\big) = 0\,, \qquad n \in \mathbb{N}\,. \qquad (A.3b)$$

From Lemma A.1 (v), $(T^1)^n(\tilde{\rho};\theta)$ converges to 1 as $n \to \infty$. Taking the limit, as $n \to \infty$, in (A.3a) and (A.3b), yields

$$h_\theta^{(1)}(1) - h_\theta^{(2)}(1) = \sup_{\rho\in[0,1]} \bigl\{h_\theta^{(1)}(\rho) - h_\theta^{(2)}(\rho)\bigr\} \geq 0\,. \tag{A.4}$$

Interchanging the roles of $h_\theta^{(1)}$ and $h_\theta^{(2)}$ in (A.4), we finally conclude that $h_\theta^{(1)} = h_\theta^{(2)}$. □

We need the following definition.

Definition A.1: For $n \in \mathbb{N}$, let D_n denote the set of multi-indices of length n, in $\{0,1\}$, i.e., $d_n \in D_n$ is of the form $d_n = \bigl(d(1),\dots,d(n)\bigr)$, $d(i) \in \{0,1\}$. If $k \leq n$ then $d_k \prec d_n$ denotes that d_k agrees with the first k coordinates of d_n, while $d_k \subset d_n$ denotes that d_n is the concatenation $d_\ell \cdot d_k \cdot d_m$, for some multi-indices d_ℓ and d_m, with $\ell + k + m = n$. Let Θ_δ^n, $n \geq 0$ denote the n-fold product of the parameter space. For a sequence $\bigl\{\hat{\theta}_t\bigr\}_{t=0}^\infty \subset \Theta_\delta$ and a positive integer $n \leq t$, we define

$$\hat{\theta}_t^n := \bigl(\hat{\theta}_{t-n+1},\dots,\hat{\theta}_t\bigr) \in \Theta_\delta^n\,.$$

For the map $T^y(\rho;\theta)$, defined in (A.1), and a multi-index $d_n \in D_n$, $T^{d_n}(\cdot;\hat{\theta}_t^n)$ denotes the n-fold composition $T^{d(n)}\bigl(T^{d(n-1)}\bigl(\cdots;\hat{\theta}_{t-1}\bigr);\hat{\theta}_t\bigr)$, and is the identity map if $n = 0$.

Lemma A.2. *Let $\theta \in (0,1)$, $d_n \in D_n$, $0 < \delta_0 < 1$ and suppose that*

$$T^{d_i}(0;\theta) \leq 1 - \delta_0, \qquad \forall d_i \prec d_n, \quad i = 1,\dots,n\,. \tag{A.5}$$

Then

$$\frac{d}{d\theta}\,T^{d_n}(0;\theta) \leq \frac{q}{\delta_0(1-q)\theta}\,T^{d_n}(0;\theta)\,. \tag{A.6}$$

Proof: With $r := \frac{1-q}{q}$ and $|d_k| := \sum_{i=1}^k d(i)$, $d_k \prec d_n$, we inductively obtain

$$\frac{T^{d_k}(0;\theta)}{1 - T^{d_k}(0;\theta)} = \frac{\theta\,\sum_{i=1}^{k-1} r^{2|d_i|-i}(1-\theta)^i}{r^{2|d_k|-k}(1-\theta)^k}, \qquad k = 1,\dots,n\,. \tag{A.7}$$

Let $\rho_k := T^{d_k}(0;\theta)$, and $\beta_k := r^{2|d_k|-k}(1-\theta)^k$. The hypothesis in (A.5) is equivalent to

$$\theta\sum_{i=1}^{k-1}\beta_i \leq \frac{(1-\delta_0)}{\delta_0}\beta_k\,. \tag{A.8}$$

Differentiating (A.7) and using (A.8), we obtain,

$$\begin{aligned}\frac{d}{d\theta}\left(\frac{\rho_n}{1-\rho_n}\right) &= \frac{1}{\theta}\frac{\rho_n}{1-\rho_n} + \frac{\theta\,\sum_{k=1}^{n-1}(n-k)\beta_k}{(1-\theta)\beta_n}\\ &\leq \frac{1}{\theta}\frac{\rho_n}{1-\rho_n} + \frac{\bigl(1+\frac{1-\delta_0}{\delta_0\theta}\bigr)\rho_n}{(1-\theta)(1-\rho_n)}\\ &= \frac{\rho_n}{\delta_0\theta(1-\theta)(1-\rho_n)}.\end{aligned} \tag{A.9}$$

By Lemma A.1 (i) and (ii), we have the following estimate

$$1 - T^{d_n}(0;\theta) \le 1 - T^{0}(0;\theta) \le \frac{q(1-\theta)}{(1-q)}\,,$$

which, in conjunction with (A.9) yields (A.6). □

Proof of Lemma 5.1: Observe that

$$\hat\rho_t\, I\big[U_k = 1\ :\ k<t\big]\ \in\ \big\{T^{d_n}(0;\hat\theta^n_t)\ :\ d_n \in D_n\,,\ 0 \le n \le t-k-1\big\}\,, \tag{A.10}$$

and the analogous relation holds for ρ_t if we replace $\hat\theta^n_t$ by θ_0 in (A.10). By Lemma A.1 (ii), it is true in general that

$$T^{d_n}(0;\min\{\hat\theta^n_t\}) \le T^{d_n}(0;\hat\theta^n_t) \le T^{d_n}(0;\max\{\hat\theta^n_t\})\,. \tag{A.11}$$

Recalling that $0 < \delta \le \theta_0 < 1$, choose η such that $\big[\theta_0-\eta,\theta_0+\eta\big] \subset (\delta,1)\cap\Theta_\delta$. Since $\alpha(\theta) \le \alpha^* < 1$ on Θ_δ, using Lemma A.1 (i)–(ii), we have

$$\hat\rho_t\, I\big[U_k = 1\,,\ \hat\theta_i \in [\theta_0-\eta,\theta_0+\eta]\ :\ i = k+2,\dots,t\big] \le T^{1}(\alpha^*;\theta_0+\eta) < 1\,. \tag{A.12}$$

In view of (A.10)–(A.12), utilizing Lemma A.2 (with $\delta_0 := 1-T^1(\alpha^*;\theta_0+\eta)$), and the mean value theorem, we conclude that given $\varepsilon > 0$ there exists a neighborhood $V_\varepsilon \subset \big[\theta_0-\eta,\theta_0+\eta\big]$ of θ_0 such that

$$|\hat\rho_t - \rho_t|\, I\big[U_k = 1\,,\ \hat\theta_i \in V_\varepsilon\ :\ i = k+2,\dots,t\big] \le \frac{q\,\mathrm{diam}\,(V_\varepsilon)}{\delta_0(1-q)\delta} \le \varepsilon\,, \tag{A.13}$$

where $\mathrm{diam}\,(V_\varepsilon)$ denotes the diameter of V_ε. Now, let $\varepsilon > 0$ be chosen. If $\hat\theta_t \to \theta_0$, $\mathcal{P}^{\pi^a}$–a.s., then, outside a set of probability 0, $\hat\theta_t \to \theta_0$ and, by Theorem 4.2, $U_k = 1$ for infinitely many integers k, along every sample path. Consider an arbitrary such sample path and choose integers $n_0 \le n_1 \in \mathbb{N}$ such that $\hat\theta_t \in V_\varepsilon$ for all $t \ge n_0$ and $U_{n_1} = 1$. Then, by (A.13), $|\hat\rho_t - \rho_t| \le \varepsilon$, $\forall t > n_1$, along the sample path.

On the other hand, if $\hat\theta_t \to \theta_0$ in probability under $\mathcal{P}^{\pi^a}$ then defining

$$A^n_t := \big\{U_k = 0\ :\ t \le k \le t+n\big\}\,, \qquad t,n \in \mathbb{N}\,,$$

and applying [3, Theorem A.2], choose $n_0 \in \mathbb{N}$ such that $\mathcal{P}^{\pi^a}\big\{A^{n_0}_{t-n_0}\big\} < \frac{\varepsilon}{2}$, for all $t > n_0$. There exists an integer $n_1 \in \mathbb{N}$, $n_1 > n_0$, such that $\mathcal{P}^{\pi^a}\big\{\hat\theta^{n_0}_t \not\subset V_\varepsilon\big\} < \frac{\varepsilon}{2}$ for all $t > n_1$. Therefore, by (A.13),

$$\mathcal{P}^{\pi^a}\big\{|\hat\rho_t - \rho_t| > \varepsilon\big\} \le \mathcal{P}^{\pi^a}\big\{\hat\theta^{n_0}_t \not\subset V_\varepsilon\big\} + \mathcal{P}^{\pi^a}\big\{A^{n_0}_{t-n_0}\big\} < \varepsilon\,, \qquad \forall t > n_1\,,$$

and the proof is complete. □