Proceeding of the 2004 American Control Conference
Boston, Massachusetts June 30 - July 2, 2004

A Note on Some New Results on the Ergodic Control of Partially Observed Markov Chains

Shun-Pin Hsu
National Chi-Nan University
Electrical Engineering
Nantou, Taiwan 545

Ari Arapostathis
The University of Texas at Austin
Electrical and Computer Engineering
Austin, TX 78712-0240

***Abstract*— In an early paper [3], the ergodic control problem of partially observed Markov chains was studied. A major assumption was proposed to justify the existence of the optimal policy characterized by a dynamic programming equation. In this note we show that the major assumption, though quite general and easily verifiable, is not satisfied by an important class of machine maintenance problems. A modified version of the assumption is thus presented to improve this shortage and an example is analyzed to show our work.**

I. INTRODUCTION

Deriving the dynamic programming equations (DPEs) for the ergodic control problems of partially observed Markov chains is aged yet far from been solved with satisfaction [1]. Classical sufficient conditions for the existence of solutions to the DPEs include Ross's *renewability* condition [5], Platzman's *reachability-detectability* condition [4], and Stettner's *positivity* condition [6]. Recently, Chung and Arapostathis [3] proposed another condition which is considered more concise and practical in comparison with those proposed before. However, a major drawback of the condition is that it is not satisfied by any system with a reachable state that is completely observable. This issue is addressed in the note and a modified version of the condition is proposed to improve its generality. In Section 2 we review all the required technical background and the major result in [3]. Section 3 provides this paper's main results, including a detailed analysis of an example by which we compare several important assumptions.

II. PRELIMINARIES

A partially observed controlled Markov chain, also known as partially observed Markov decision process (POMDP), is governed by a five-tuple ($\boldsymbol{S}$,$\boldsymbol{U}$,$\mathcal{U}$, Q, c) with the following meanings: $\boldsymbol{S} = \boldsymbol{X} \times \boldsymbol{Y}$ is the process's state space where $\boldsymbol{X}$, $\boldsymbol{Y}$ is the finite system space, finite observable space with cardinality N_x, N_y respectively. $\boldsymbol{U}$ is the compact action space. Let $\mathcal{B}(\boldsymbol{V})$ denote the σ-algebra for a given topological space $\boldsymbol{V}$, then $\mathcal{U}$: $\boldsymbol{X} \to \mathcal{B}(\boldsymbol{U})$ is a set-valued map with compact non-empty value and $\mathcal{U}(x)$ is the set of feasible actions when the system is in state $x \in \boldsymbol{X}$. Q is the Markovian transition kernel of the process and $c : \boldsymbol{X} \times \boldsymbol{U} \to \mathbb{R}^n$ is the cost function assumed continuous and bounded. Specifically, when the system state at time t is X_t and a control U_t is taken, a cost $c(X_t, U_t)$ is incurred and the system moves to next state X_{t+1} with observation Y_{t+1} according to the transition kernel Q defined by

$$[Q_y^u]_{ij} := \text{Prob}(X_{t+1} = j, Y_{t+1} = y | X_t = i, U_t = u)$$

for all $t \in \mathbb{N}_0$ (the set of nonnegative integers), $j \in \boldsymbol{X}$, and $y \in \boldsymbol{Y}$. Also, the mapping $u \to Q_y^u$ is assumed continuous. It is well know that we can transform the partially observed process into the completely observed process by constructing the information state ψ_t , recursively defined by

$$\psi_{t+1} := \sum_{y \in \boldsymbol{Y}} \frac{\psi_t Q_y^u}{\psi_t Q_y^u \mathbf{1}} \cdot \mathbf{1}_{\{Y_{t+1}=y\}} ,$$

for $\psi_t \in \boldsymbol{\Psi} = \boldsymbol{\mathcal{P}}(\boldsymbol{X})$, the probability (row) vector space on $\boldsymbol{X}$, $u \in \boldsymbol{U}$ and $t \in \mathbb{N}_0$ where $\mathbf{1}_{\{\cdot\}}$ is the indicator function and $\mathbf{1}$ the column vector of 1's of appropriate size. Let $V(\psi, y, u) := \psi Q_y^u \mathbf{1}$ and $T(\psi, y, u) := \psi Q_y^u / V(\psi, y, u)$ for $V(\psi, y, u) \neq 0$ then the transition kernel for the information state is given by

$$\begin{aligned} \mathcal{K}(B|\psi, u) :&= \text{Prob}\{\psi_{t+1} \in B | \psi_t = \psi, U_t = u\}, \quad \text{(II.1)} \\ &= \sum_{Y_{t+1} \in \boldsymbol{Y}} V(\psi, Y_{t+1}, u) \cdot \mathbf{1}_{\{T(\psi, Y_{t+1}, u) \in B\}} \end{aligned}$$

for all $B \in \mathcal{B}(\boldsymbol{\Psi})$, $\psi \in \boldsymbol{\Psi}$, and $u \in \boldsymbol{U}$. So we can write the transformed five-tuple as ($\boldsymbol{\Psi}$, $\boldsymbol{U}$, $\tilde{\mathcal{U}}$, $\mathcal{K}$, $\tilde{c}$) where $\tilde{\mathcal{U}}$: $\boldsymbol{\Psi} \to \mathcal{B}(\boldsymbol{U})$ and $\tilde{c}(\psi, u) := \sum_{x \in \boldsymbol{X}} c(x, u)\psi(x)$ for all $\psi \in \boldsymbol{\Psi}$ and $u \in \boldsymbol{U}$. For the original history space of the partially observed process $\boldsymbol{H}_0 := \boldsymbol{\Psi}$, $\boldsymbol{H}_t := \boldsymbol{H}_{t-1} \times \boldsymbol{U} \times \boldsymbol{Y}$ for all $t \in \mathbb{N}$ (the set of positive integers) we obtain a corresponding completely observed history space: $\hat{\boldsymbol{H}}_0 := \boldsymbol{\Psi}$, $\hat{\boldsymbol{H}}_t := \hat{\boldsymbol{H}}_{t-1} \times \boldsymbol{U} \times \boldsymbol{\Psi}$ for all $t \in \mathbb{N}$.

An *admissible strategy* or *admissible policy* π is a sequence $\{\pi_t\}_{t=0}^{\infty}$ of Borel measurable stochastic kernels π_t on $\boldsymbol{U}$ given $\hat{\boldsymbol{H}}_t$ satisfying $\pi_t(\tilde{\mathcal{U}}(\psi_t)|h_t) = 1$ for all $\psi_t \in \boldsymbol{\Psi}$, $h_t \in \hat{\boldsymbol{H}}_t$ and $t \in \mathbb{N}_0$. An admissible policy is called *deterministic* if there exists a function $f : \boldsymbol{\Psi} \to \boldsymbol{U}$ such that $\pi_t(f(\psi_t)|h_t) = 1$ for all $\psi_t \in \boldsymbol{\Psi}$, $h_t \in \hat{\boldsymbol{H}}_t$ and $t \in \mathbb{N}_0$.

It is shown that for an initial distribution $\psi_0 \in \boldsymbol{\Psi}$ and admissible strategy π, there exists an unique probability measure $\mathbb{P}_{\psi_0}^{\pi}$ induced on the sample path $(\boldsymbol{\Psi} \times \boldsymbol{U})^{\infty}$. Denote the corresponding expectation operator as $\mathbb{E}_{\psi_0}^{\pi}$ and the incurred β-discounted cost, $\beta \in (0, 1)$, as

$$J_\beta(\psi_0, \pi) = \lim_{T \to \infty} \mathbb{E}_{\psi_0}^{\pi} \left[\sum_{t=0}^{T-1} \beta^t \tilde{c}(\psi_t, U_t) \right] .$$

0-7803-8335-4/04/$17.00 ©2004 AACC

If $h_\beta(\psi) = \inf_{\pi\in\Pi} J_\beta(\psi,\pi)$ for $\psi \in \boldsymbol{\Psi}$, then it is well know that $h_\beta(\psi)$ is concave on $\boldsymbol{\Psi}$ and is the unique solution in $\boldsymbol{C}(\boldsymbol{\Psi})$ (the space of continuous functions on $\boldsymbol{\Psi}$) for Bellman's β-discounted optimality equation

$$h_\beta(\psi) = \min_{u\in\boldsymbol{U}} \left\{ \tilde{c}(\psi,u) + \beta \int_{\boldsymbol{Y}} h_\beta(\psi')\mathcal{K}(d\psi'|\psi,u) \right\}. \tag{II.2}$$

Other properties regarding $h_\beta(\psi)$ can be seen in [4].

The classical vanishing discount limit method extends the result of β-discounted cost model to the long-run average cost model with incurred costs

$$J(\psi_0,\pi) := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi_{\psi_0}\left[\sum_{t=0}^{T-1} \tilde{c}(\psi_t,U_t)\right] \tag{II.3}$$

by letting $\beta \to 1$ on the modified form of (II.2). (See, e.g., [1]). In order to justify the method, the following assumption was proposed in [3], along with its implication on the existence of the optimal policy.

Assumption 2.1: There exist constants $\varepsilon > 0$, $N_b \in \mathbb{N}$ and $\beta_0 < 1$ such that for each $\beta \in [\beta_0, 1)$ we have

$$\max_{1\le k\le N_b} \mathbb{P}^{\pi_\beta}_{\psi_*}\{\psi_k \in \boldsymbol{\Psi}_\varepsilon\} \ge \varepsilon\,,$$

$\boldsymbol{\Psi}_\varepsilon := \{\psi \in \boldsymbol{\Psi} \,|\psi(i) \ge \varepsilon, i \in \boldsymbol{X}\}$, $\psi_* := \text{argmin}_{\psi\in\boldsymbol{\Psi}} h_\beta(\psi)$.

Theorem 2.1: If Assumption 2.1 holds, then there exist a bounded, concave and continuous function h: $\boldsymbol{\Psi} \to \mathbb{R}$ and an optimal ergodic cost ρ such that $(\rho, h(\cdot))$ is a solution of the dynamic programming equation:

$$\rho + h(\psi) = \min_{u\in\boldsymbol{U}} \left\{ \tilde{c}(\psi,u) + \int_{\boldsymbol{Y}} h(\psi')\mathcal{K}(d\psi'|\psi,a) \right\}. \tag{II.4}$$

Any measurable selector π of the minimizer in (II.4) is an optimal policy in the sense of the long-run average cost.

III. Main Result

An immediate example failing to meet Assumption 2.1 is the completely observable system, which is a special case of the partially observable system and is well known on the existence of its optimal control policy under mild conditions. To deal with this issue, Assumption 2.1 is modified as follows.

Assumption 3.1: There exist constants $\varepsilon > 0$, $N_b \in \mathbb{N}$ and $\beta_0 < 1$ such that $\forall \beta \in [\beta_0, 1)$ we have

$$\max_{1\le k\le N_b} \mathbb{P}^{\pi_\beta}_{\psi_*}\{T(\psi_*,Y^k,U^{k-1}) \ge \varepsilon T(\psi^*,Y^k,U^{k-1})\,, \\ V(\psi^*,Y^k,U^{k-1}) \ge \varepsilon V(\psi_*,Y^k,U^{k-1})\} \ge \varepsilon\,,$$

where $\psi^* := \text{argmax}_{\psi\in\boldsymbol{\Psi}} h_\beta(\psi)$. $V(\psi,y^k,u^{k-1})$ and $T(\psi,y^k,u^{k-1})$ are defined similar to $V(\psi,y,u)$ and $T(\psi,y,u)$ in Section II except for the multi-step transition kernel $Q^{u^{k-1}}_{y^k} := Q^{u_0}_{y_1}\cdots Q^{u_{k-1}}_{y_k}$.

Theorem 3.1: In Theorem 2.1, if Assumption 2.1 is replaced by Assumption 3.1 then its result still holds.

Now we study an example to compare several assumptions proposed in the literature.

Example 3.1: Consider a machine with its state space $\boldsymbol{X}=\{0,1,2\}$ where 0,1 and 2 represents *good*, *need maintenance*, and *down*, respectively; and action space $\boldsymbol{U}=\{0,1\}$ where 0 and 1 means to *continue* and to *replace*, respectively. Assume that the relation between the costs $c(x,u)$ for various kinds of combinations of $x \in \boldsymbol{X}$ and $u \in \boldsymbol{U}$ is $0 \le c(0,0) < c(1,0) < c(2,0) < c(x,1) < \infty$. Suppose action '0' and '1' influence the evolution of the machine state according to the transition matrix P_0 and P_1, respectively, where

$$P_0 = \begin{bmatrix} \theta_1 & \theta_2 & 1-\theta_1-\theta_2 \\ 0 & \theta_3 & 1-\theta_3 \\ 0 & 0 & 1 \end{bmatrix}, P_1 = \begin{bmatrix} 1 & 0 & 0 \\ \theta_4 & 1-\theta_4 & 0 \\ \theta_5 & 1-\theta_5 & 0 \end{bmatrix},$$

and there exists a probability of erroneous observation between state 0,1 and 2, then the process is partially observable with observation space $\boldsymbol{Y}$=$\boldsymbol{X}$ and the transition kernel $Q^u_y = P_u O_y$, with $O_y = diag[q_{1y}, q_{2y}, q_{3y}]$, the diagonal matrix with $[O_y]_{ii} = q_{iy} \ge 0$ and $\sum_{y\in\boldsymbol{Y}} q_{iy} = 1$ for $i \in \boldsymbol{X}$. Assume all of the θ_i's are lower-bounded by a positive constant, then we are able to show that $\pi_\beta(\psi_*)$=0 where $\psi_* = [\,1\;0\;0\,]$ or $[\,0\;1\;0\,]$, depending on the parameters in the transition kernels and the cost function. If $\psi_* = [\,1\;0\;0\,]$, we consider *Case 1*: there exists an observation $y \in \boldsymbol{Y}$ such that $q_{1y}, q_{2y},$ and q_{3y} are all lower-bounded by a positive number; and *Case 2*: the wrong observation could just happen between state 0 and 1 in the sense that we can express $O_1 = diag[q, 1-q, 0], O_2 = diag[1-q, q, 0]$, and $O_3 = diag[0,0,1]$ with $q \in$(.5, 1).

In Case 1 Assumption 2.1 is satisfied but renewability condition in [4] fails. In Case 2 we have a mixed observation possibility since both partial and complete observation can occur. In this case information state $[\,0\;0\;1\,]$ serves as a recurrent state for the overall process so the renewability condition in [4] is satisfied. However, the information state ψ_t will never enter the interior of simplex $\boldsymbol{\Psi}$ for $t \ge 1$, so Assumption 2.1 fails. Finally, we note that in either case the positivity assumption in [6] or detectability condition in [4] is not satisfied due to the appearance of zeros in the transition kernels, but Assumption 3.1 can be checked to hold in either case.

References

[1] A. Arapostathis, V. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM Journal on Control & Optimization*, **31** (1993): pp. 282–344.

[2] D.-M. Chuang. *Risk-sensitive control of discrete-time partially observed Markov decision processes.* Ph.D. Dissertation, The University of Texas at Austin, 1999.

[3] D.-M. Chuang and A. Arapostathis. Some new results on the ergodic control of partially observed Markov chains. *Proceedings of the 38th IEEE conference on decision and control*, pp. 1908-1909, 1999

[4] L. K. Platzman. Optimal infinite-horizon undiscounted control of finite probabilistic systems. *SIAM Journal on Control & Optimization* , **18** (1980): pp. 362–380.

[5] S. M. Ross. Arbitrary state Markovian decision processes. *Annals of Mathematical Statistics*, **6** (1968): pp. 2118–2122.

[6] W. J. Runggaldier and L. Stettner. *Approximations of discrete time partially observed control problems*, Applied Mathematics Monographs **6** (1994), Giardini Editori E Stampatori in Pisa, Italy .