

A METHODOLOGY FOR THE ADAPTIVE CONTROL
OF MARKOV CHAINS UNDER PARTIAL
STATE INFORMATION

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

SUMMARY

We consider a stochastic adaptive control problem where complete state information is *not available* to the controller. The system is modelled as a *finite stochastic automaton* (FSA) [PAZ], [DOB]. These models are a slight generalization of the more common partially observable controlled Markov chain models as presented in, e.g. [BE], [KV]. A controlled FSA is described by the quintuplet $\langle \mathbf{X}, \mathbf{Y}, \mathbf{U}, \{P(y | u) : (y, u) \in \mathbf{Y} \times \mathbf{U}\}, c \rangle$; here $\mathbf{X} = \{1, 2, \dots, N_{\mathbf{X}}\}$ is the finite set of internal states, $\mathbf{Y} = \{1, 2, \dots, N_{\mathbf{Y}}\}$ is the set of observations (or messages), $\mathbf{U} = \{1, 2, \dots, N_{\mathbf{U}}\}$ is the set of decisions (or controls), and $c(\cdot, \cdot)$ is the one-stage cost function. For each pair $(y, u) \in \mathbf{Y} \times \mathbf{U}$, we have that $P(y | u) := [p_{i,j}(y | u)]$ is a $N_{\mathbf{X}} \times N_{\mathbf{X}}$ matrix, such that

$$p_{i,j}(y | u) \geq 0, \quad \sum_{j=1}^{N_{\mathbf{X}}} p_{i,j}(y | u) = 1, \quad \forall i \in \mathbf{X}, u \in \mathbf{U}.$$

If at time t the automaton is in state $X_t = i$ and decision $U_t = u$ is made, then by the beginning of the next decision time the automaton would have evolved to state $X_{t+1} = j$, and output a message $Y_{t+1} = y$ with probability $p_{i,j}(y | u)$. The cost incurred in this process is $c(i, j)$. We refer to [ABFGM], [DOB], [PAZ] for more details.

At decision time t , the information available to the decision-maker is

$$I_t := \{p_0, U_0, Y_1, U_1, \dots, U_{t-1}, Y_t\} = (I_{t-1}, U_{t-1}, Y_t),$$

where $p_0 \in S_{N_{\mathbf{X}}} := \{p \in \mathbf{R}^{N_{\mathbf{X}}} \mid p^{(i)} \geq 0, \sum_i p^{(i)} = 1\}$ is the initial state distribution. It is well known that the partially observable optimal control problem for a FSA, under several optimality criteria, can be transformed into an *equivalent completely observable problem*, in terms of an *information state process* [ABFGM], [BE], [KV], as follows. Given $p_0 \in S_{N_{\mathbf{X}}}$, compute recursively

$$p_{t+1} = T(Y_{t+1}, p_t, U_t), \quad t \in \mathbf{N},$$

where,

$$T(y, p, u) := \frac{P(y | u)p}{\mathbf{1}'P(y | u)p};$$

here we have $\mathbf{1} = (1, 1, \dots, 1)'$. The process $\{p_t\}$ is a controlled Markov chain and equals the conditional distribution of the internal state X_t given I_t [ABFGM], [BE]. The "new" state is then taken as the process $\{p_t\}$.

† Systems and Industrial Engineering Department,
The University of Arizona, Tucson, Arizona 85721
(emmanuel@sie.arizona.edu).

‡ Department of Electrical and Computer Engineering,
The University of Texas at Austin, Austin, Texas 78712-1084
(ari@emx.utexas.edu).

§ Systems Research Center & Electrical Engineering
Department, The University of Maryland, College Park,
Maryland 20742 (marcus@src.umd.edu).

A (stationary separated) policy π is a rule for making decisions, based on $\{p_t\}$, i.e. $\pi : S_{N_{\mathbf{X}}} \rightarrow \mathbf{U}$, and $U_t = \pi(p_t)$. The stochastic optimal control problem of interest to us is that of finding a policy π^* , optimal with respect to the long-run expected average cost (AC) performance criterion, which for a given policy π and initial state distribution $p_0 \in S_{N_{\mathbf{X}}}$ is given as

$$J(p_0; \pi) := \lim_{N \rightarrow \infty} \frac{1}{N} \mathbf{E}_{p_0}^{\pi} \left\{ \sum_{t=0}^{N-1} \bar{c}(p_t, U_t) \right\},$$

where $\bar{c}(p, u) := p'(c(1, u), \dots, c(N_{\mathbf{X}}, u))'$. The infinite horizon optimal control problem for FSA under an AC criterion has been studied by the authors and others [ABFGM]. Since the state space $S_{N_{\mathbf{X}}}$ is a general (Borel) space, then this problem can be thought of as falling into the realm of completely observable *controlled Markov processes* (CMP), with general (Borel) state space, c.f. [ABFGM]. However, the problem with partial information has a very rich structure which is not fully utilized by following the latter approach [ABFGM], [BE], [FAM1], [FG]. Furthermore, many of the assumptions used in the literature on the AC control problem for general state space CMP require some form of strong ergodicity for the controlled process $\{p_t\}$, *under all stationary control policies*. This is not satisfied for many applications of much interest, c.f. [FAM1]. Hence the idea of viewing the (equivalent) FSA problem as a general state space completely observable CMP very often is not advantageous at all, for many purposes. This is especially true for the case of *parametric adaptive control* of FSA. In this situation, the model depends on some unknown parameter θ_0 , which we denote as $P_{\theta_0}(y | u)$; the parameter takes values in some (Borel) parameter space Θ . Hence, the *true* conditional probability depends on this parameter, i.e.:

$$p_{t+1} = T(Y_{t+1}, p_t, U_t; \theta_0).$$

Therefore, since the true value of the parameter is unknown to the controller, $\{p_t\}$ *cannot be computed* and thus the equivalent problem is *not completely observable* anymore.

Although a very interesting problem with much potential for applications [BDO], [BTE], [FAM2], [WAK] there is very little available in the literature concerning the adaptive control of FSA. Recently, the above adaptive control problem has been studied by the authors [FAM2], [FG]. In [FAM2], a complete analysis for a particular case study has been reported; the methodology used has been generalized in [FG] as follows: we adopt an "enforced certainty equivalence" approach which involves recursively computing estimates $\{\hat{\theta}_t\}$ of the unknown parameter, and using at each decision time the latest available estimate to compute

$$\hat{p}_{t+1} = T(Y_{t+1}, \hat{p}_t, U_t; \hat{\theta}_t), \quad \hat{p}_0 = p_0. \quad (1)$$

We assume that the solution to the stochastic optimal control problem is known, for each $\theta \in \Theta$, which is expressed as follows; see [ABFGM], [BE], [KV].

Assumption A.1: For each θ , there is a bounded solution $(\rho_\theta^*, h_\theta)$, with $\rho_\theta^* \in \mathbb{R}$, to the corresponding *average cost optimality equation* (ACOE)

$$\rho_\theta^* + h_\theta(p) = \min_{u \in \mathbf{U}} \left\{ \bar{c}(p, u) + \sum_{y \in \mathbf{Y}} \mathbf{1}' P(y | u) p h_\theta(T(y, p, u)) \right\}.$$

Under the above assumption, there exists a set of optimal policies $\mathcal{OP} = \{\pi^*(\cdot; \theta)\}_{\theta \in \Theta}$, see [ABFGM]. The certainty equivalent adaptive policy is given as follows.

- **Adaptive Policy:** Given a sequence of estimates $\{\hat{\theta}_t\}_{t=0}^\infty$ of θ_0 , compute the control action at each time t by

$$U_t = \pi^*(\hat{p}_t; \hat{\theta}_t),$$

where \hat{p}_t is computed recursively using (1).

The above adaptive policy will be denoted by π^a . Under a set of assumptions, it was shown in [FG] that the adaptive policy π^a is *self-optimizing* with respect to the AC criterion, i.e. it achieves the same asymptotic average performance as the optimal policy $\pi^*(\cdot; \theta_0) \in \mathcal{OP}$ corresponding to the true parameter. The other assumptions used in [FG] are the following.

Assumption A.2: The parameter set Θ is compact; $(\rho_\theta^*, h_\theta)$ are continuous and bounded, both in p and in θ .

Assumption A.3: $P_\theta(y | u)$ is continuous in θ , for each $(y, u) \in \mathbf{Y} \times \mathbf{U}$.

Assumption A.5: We have that $\hat{p}_t \xrightarrow[t \rightarrow \infty]{} p_t$, in probability, for all p_0 .

We can now prove our main result.

Theorem: Under Assumptions A.1-A.5, π^a is self-optimizing with respect to the AC criterion.

Proof: Let $\Phi_\theta(\cdot, \cdot)$ denote Mandl's discrepancy function, corresponding to the parameter value $\theta \in \Theta$, i.e. for $p \in S_{N_x}$ and $u \in \mathbf{U}$ (see [ABFGM], [FAM1])

$$\Phi_\theta(p, u) := \bar{c}(p, u) + \sum_{y \in \mathbf{Y}} \mathbf{1}' P_\theta(y | u) p h_\theta(T(y, p, u; \theta)) - \rho_\theta^* - h_\theta(p).$$

Then by the assumptions made, $\Phi_\theta(p, u)$ is continuous in both $p \in S_{N_x}$ and $\theta \in \Theta$. Furthermore, since Θ is compact, then $\Phi_\theta(p, u)$ is uniformly continuous and bounded in $(p, \theta) \in S_{N_x} \times \Theta$, and thus $\Phi_{\hat{\theta}_t}(\hat{p}_t, u)$ is uniformly integrable, for each $u \in \mathbf{U}$. Therefore, for each $u \in \mathbf{U}$, we have

$$\mathbf{E}_{p_0}^{\pi^a} \left\{ \left| \Phi_{\hat{\theta}_t}(\hat{p}_t, u) - \Phi_{\theta_0}(p_t, u) \right| \right\} \xrightarrow[t \rightarrow \infty]{} 0,$$

and since \mathbf{U} is finite, then

$$\mathbf{E}_{p_0}^{\pi^a} \left\{ \Phi_{\theta_0}(p_t, \pi(\hat{p}_t; \hat{\theta}_t)) \right\} \xrightarrow[t \rightarrow \infty]{} 0, \quad (2)$$

where we used the fact that $\Phi_{\theta_0}(\hat{p}_t, \pi(\hat{p}_t; \hat{\theta}_t)) = 0$, since $\pi(\cdot; \theta) \in \mathcal{OP}$ minimizes the corresponding ACOE, for the parameter value $\theta \in \Theta$. The result then follows from (2); see [ABFGM, Theorem 6.3]. \square

Let us briefly examine the assumptions used in deriving the result above. Verifiable conditions on the model specifications exist in the literature that imply Assumption A.1 holds [ABFGM], [FAM1]. Assumption A.3 is easy to verify, and holds trivially if the parameterization of the model is taken in terms of the entire matrices $P_\theta(y | u)$. Assumption A.4 depends on the parameter estimation scheme used, and is very problem-specific [FAM2]. The continuity required in Assumption A.2 depends to a large extent on the continuity required in Assumption A.3, and on some ergodic properties of the model [FAM2]. Finally, it is clear that even if Assumptions A.1-A.4 hold, it may be the case that Assumption A.5 does not. Under continuity with respect to the parameterization, and in the presence of converging parameter estimates, this last assumption will hold if there is some type of, e.g. *regenerative* behavior for the processes $\{p_t\}$ and $\{\hat{p}_t\}$, such that at some times both processes are reset to the same value. This type of behavior occurs naturally in some inventory, queueing and machine replacement problems [BE], [ABFGM], [FAM2].

ACKNOWLEDGEMENTS: This work was supported in part by the Texas Advanced Technology Program under Grants No. 003658-093 and No. 003658-186, in part by the Air Force Office of Scientific Research under Grants AFOSR-91-0033, F49620-92-J-0045, F49620-92-J-0083, and in part by the National Science Foundation under Grants CDR-8803012 and INT-9201430.

REFERENCES

- [ABFGM] 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, to appear in *SIAM Journal on Control & Optimization*.
- [BE] D.P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, Englewood Cliffs, 1987.
- [BDO] J.S. Baras and A.J. Dorsey, Stochastic Control of Two Partially Observed Competing Queues, *IEEE Transactions on Automatic Control*, **AC-26** (1981) 1106-1117.
- [BTE] F.J. Beutler and D. Teneketzis, Routing in Queueing Networks Under Imperfect Information: Stochastic Dominance and Thresholds, *Stochastics & Stochastics Reports*, **26** (1989) 81-100.
- [DOB] E.-E. Doberkat, *Stochastic Automata: Stability, Nondeterminism, and Prediction*, Springer-Verlag, Berlin, 1981.
- [FAM1] 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* **29** (1991) 439-470.
- [FAM2] E. Fernández-Gaucherand, A. Arapostathis and S.I. Marcus, Analysis of an Adaptive Control Scheme for a Partially Observed Controlled Markov Chain, to appear in *IEEE Transactions in Automatic Control*.
- [FG] 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.
- [KV] P.R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, 1986.
- [PAZ] A. Paz, *Introduction to Probabilistic Automata*, Academic Press, New York, 1971.
- [WAK] K. Wakuta, Optimal Control of an M/G/1 Queue with Imperfectly Observed Queue Length when the Input Source is Finite, *Journal of Applied Probability*, **28** (1991) 210-220.