equivalent or lift of the multirate plant, translating the given multirate costs and noise covariances to the lifted space, and then solving a constrained shift-invariant LQG problem. Our main result also gives a procedure for rms noise power cost translation in a multirate QDES package.

References

[1] M. Athans and P. L. Falb, *Optimal Control*. New York: McGraw-Hill, 1966.

[2] H. Al-Rahmani and G. F. Franklin, "Linear periodic systems: Eigenvalue assignment using discrete periodic feedback," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 99–103, 1989.

[3] ——, "A new optimal multirate control of linear periodic and time-invariant systems," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 406–415, 1990.

[4] ——, "Multirate control: A new approach," preprint, Info. Syst. Lab., Dept. Elec. Eng., Stanford Univ., Stanford, CA.

[5] W. F. Arnold and A. J. Laub, "Generalized eigenvalue problem algorithms and software for algebraic Riccati equations," in *Proc. IEEE*, vol. 72, pp. 1746–1754, 1984.

[6] B. D. O. Anderson and J. B. Moore, *Linear Optimal Control*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[7] N. Amit, "Optimal control of multirate digital control systems," Ph.D. dissertation, Dept. Aeronautics, Astronautics, Stanford Univ., Stanford, CA, July 1980.

[8] N. Amit and J. D. Powell, "Optimal control of multirate systems," in *Proc. AIAA Guidance Contr. Conf.*, Albuquerque, NM, Aug. 1981, paper 81-1797.

[9] M. Araki and K. Yamamoto, "Multivariable multirate sampled-data systems: State-space description, transfer characteristics, and Nyquist criterion," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 145–154, Feb. 1986.

[10] M. C. Berg, N. Amit, and J. Powell, "Multirate digital control system design," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 1139–1150, 1988.

[11] S. P. Boyd, V. Balakrishnan, C. H. Barratt, N. M. Khraishi, X. M. Li, D. G. Meyer, and S. A. Norman, "A new CAD method and associated architectures for linear controllers," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 268–284, Mar. 1988.

[12] R. E. Bellman, J. Bentsman, and S. M. Meerkov, "Vibrational control of nonlinear systems: Vibrational stabilizability," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 710–717, 1986.

[13] T. Chen and B. A. Francis, "Linear time-varying H_2 optimal control of sampled data systems," Dept. Electrical Eng., Univ. of Toronto, Canada, Sys. Contr. Group Rep. 9005, 1990.

[14] A. B. Chammas and C. Leondes, "Pole-placement by piecewise constant output feedback," *Int. J. Contr.*, vol. 29, pp. 31–38, 1979.

[15] P. R. Colaneri, R. Scattolini, and N. Sciavoni, "The LQG problem for multirate sampled-data systems," in *Proc. 28th Conf. Decision Contr.*, Tampa, FL, 1989, pp. 469–474.

[16] ——, "Stabilization of multirate sampled-data liner systems," *Automatica*, vol. 26, no. 2, pp. 377–380, 1990.

[17] B. A. Francis and T. T. Georgiou, "Stability theory for linear time-invariant plants with periodic digital controllers," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 820–832, 1988.

[18] D. P. Glasson, "A new technique for multirate digital control design and sample rate selection," *AIAA J. Guid. Contr.*, vol. 5, pp. 379–382, 1982.

[19] T. T. Georgiou, A. M. Pascoal, and P. P. Khargonekar, "On the robust stabilizability of uncertain liner time-invariant plants using nonlinear time-varying controllers," *Automatica*, vol. 23, pp. 617–625, Sept. 1987.

[20] E. I. Jury and F. J. Mullin, "The analysis of sampled-data control systems with a periodically time-varying sampling rate," *IRE Trans. Automat. Contr.*, vol. AC-24, pp. 15–21, 1959.

[21] R. E. Kalman and J. E. Bertram, "A unified approach to the theory of sampling systems," *J. Franklin Inst.*, no. 267, pp. 405–436, 1959.

[22] P. P. Khargonekar, K. Poola, and A. Tannenbaum, "Robust control of linear time-invariant plants using periodic compensation," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 1088–1097, 1985.

[23] G. M. Kranc, "Input–output analysis of multirate feedback systems," *IRE Trans. Automat. Contr.*, vol. AC-3, pp. 21–28, 1957.

[24] B. Lennartson, "On the design of stochastic control systems with multirate sampling," Ph.D. dissertation, Chalmers Univ. Technol., Göteborg, Sweden, Tech. Rep. 161, 1986.

[25] S. Lee, S. M. Meerkov, and T. Runolfsson, "Vibrational feedback control: Zero placement capabilities," *IEEE Trans. Automat. Contr.*, vol. AC-32, pp. 604–611, 1987.

[26] D. G. Meyer, "A new class of shift-varying operators, their shift-invariant equivalents, and multirate digital systems," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 429–433, 1990.

[27] J. Sklansky and J. R. Ragazzini, "Analysis of errors in sampled-data feedback systems—Part II," *AIEE Trans.*, vol. 74, pp. 65–71, 1955.

[28] J. C. Willems, "Least squares stationary optimal control and the algebraic Riccati equation," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 621–634, 1971.

Perturbation and Stability Theory for Markov Control Problems

Mohammed Abbad and Jerzy A. Filar

***Abstract*—We propose a unified approach to the asymptotic analysis of a Markov decision process disturbed by an ϵ-additive perturbation. Irrespective of whether the perturbation is regular or singular, the underlying control problem that needs to be understood is the limit Markov control problem. The properties of this problem are the subject of this study.**

I. Introduction

Finite state and action Markov decision processes (MDP's) are dynamic, stochastic, systems controlled by one or more controllers, sometimes referred to as decision makers. These models have been extensively studied since the 1950's by applied probabilists, operations researchers, and by engineers who often refer to them as Markov control problems. The case of the single controller constitutes the now classical MDP models initially studied by Howard [13] and Blackwell [4] and, following the latter, is often referred to as discrete dynamic programming.

During the 1960's and 1970's the theory of classical MDP's evolved to the extent that there is now a complete existence theory, and a number of good algorithms for computing optimal policies, with respect to criteria such as maximization of limiting average expected reward, or the discounted expected reward. These models were applied in a variety of contexts, ranging from water-resource models, through communication networks, to inventory and maintenance models.

The recent graduate-level texts and monographs by Ross [16], Denardo [6], Federgruen [10], Kallenberg [14], Tijms [19], Kumar and Varaiya [15], and Hernandez-Lerma [12] indicate the contin-

Manuscript received August 24, 1990; revised April 11, 1991. Paper recommended by Associate Editor, A. Shwartz.

M. Abbad was with the Department of Mathematics and Statistics, University of Maryland at Baltimore, Baltimore, MD 21228. He is now with the Université Mohammed V, Faculté des Sciences, Département de Mathematiques, B.P. 1014, Rabat, Morocco.

J. A. Filar is with the Department of Mathematics and Statistics, University of Maryland at Baltimore, Baltimore, MD 21228.

IEEE Log Number 9201324

ued research interest in these topics. Implicit in many of these works (with the notable exception of [12] and [15]) are the assumptions of complete information of the model data and parameters. However, in recent years a new generation of challenging problems in MDP's began to be addressed. One class of these problems focused around the following question: In view of the fact that in most applications the data of the problem are known at best, only approximately, how are optimal controls from the complete information model affected by perturbations (typically small) of the problem data?

From the practical point of view the above question is of vital importance, however, it leads to challenging mathematical problems. Much of the complexity arises from the fact that if the perturbation of a Markov chain alters the ergodic structure of that chain, then the stationary distribution of the perturbed process has a discontinuity at the zero value of the disturbance parameter. This phenomenon was illustrated by Schweitzer [17] with the following example. Let

$$P_\epsilon = \begin{pmatrix} 1-\epsilon/2 & \epsilon/2 \\ \epsilon/2 & 1-\epsilon/2 \end{pmatrix}$$

be the perturbed Markov chain whose stationary distribution matrix is

$$P_\epsilon^* = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

for all $\epsilon \in (0, 2]$. Thus, we have

$$\lim_{\epsilon \downarrow 0} P_\epsilon^* = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \neq P_0^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

where P_0^* is the stationary distribution matrix of the unperturbed Markov chain P_0. The above difficulty has led researchers to differentiate between the case that avoids the above-mentioned discontinuity, and the cases that permit it. Somewhat imprecisely, perhaps, the former is often referred to as a regular perturbation, and the latter as a singular perturbation. Of course, it is possible to study the properties of perturbed MDP's without performing the asymptotic analysis (as the perturbation tends to zero), and in such a case the distinction between the regular and singular perturbations is not essential (see, for instance, [9] and [8]).

In this note we propose a unified approach to the asymptotic analysis of an MDP with an ϵ-additive perturbation. Irrespective of whether the perturbation is regular or singular, the underlying control problem that needs to be understood is the limit Markov control problem (see Section II). The properties of this problem are the subject of this study. The note is organized as follows.

In Section II, we give some definitions and formulate the limit Markov control problem. In Section III, we present some theoretical results. In particular, we show that an optimal solution to the perturbed MDP can be approximated by an optimal solution of the limit Markov control problem for sufficiently small perturbations. In Section IV, we investigate the discounted case, and we show that an optimal solution to the perturbed MDP can be approximated by an optimal solution of the original MDP for sufficiently small perturbations. In Section V, we discuss a more general additive perturbation, and we show that the same conclusion as in Section IV can be derived for the unichain, the communicating, and the discounted cases. In Section VI, we present an application concerning the approximating models for the communicating and unichain cases.

II. Definitions and Preliminaries

A discrete Markovian decision process (MDP) is observed at time points $t = 0, 1, 2, \cdots,$. The state space is denoted by $S = \{1, 2, \cdots, N\}$. With each state $s \in S$ we associate a finite action set $A(s) = \{1, 2, \cdots, m_s\}$. At any time point t the system is in one of the states s, and the controller chooses an action $a \in A(s)$; as a result the following occur: i) an immediate reward $r(s, a)$ is accrued; and ii) the process moves to a state $s' \in S$ with transition probability $p(s'|s, a)$, where $p(s'|s, a) \geq 0$ and $\sum_{s' \in S} p(s'|s, a) = 1$. Henceforth, such an MDP will be synonymous with the four-tuple $\Gamma = \langle S, \{A(s): s \in S\}, \{r(s, a): s \in S, a \in A(s)\}, \{p(s'|s, a): s, s' \in S, a \in A(s)\}\rangle$.

A decision rule π^t at time t is a function which assigns a probability to the event that any particular action is taken at time t. In general π^t may depend on all realized states, and on all realized actions up to time t. Let $h_t = (s_0, a_0, s_1, \cdots, a_{t-1}, s_t)$ be the history up to time t where $a_0 \in A(s_0), \cdots, a_{t-1} \in A(s_{t-1})$, then $\pi^t(h_t, \cdot)$ is a probability distribution on $A(s_t)$, that is, $\pi^t(h_t, a_t)$ is the probability of selecting the action a_t at time t, given the history h_t. A strategy π is a sequence of decision rules $\pi = (\pi^0, \pi^1, \cdots, \pi^t, \cdots)$. A Markov strategy is one in which π^t depends only on the current state at time t. A stationary strategy is a Markov strategy with identical decision rules. A deterministic strategy is a stationary strategy whose single decision rule is nonrandomized.

Let C, $C(S)$, and $C(D)$ denote the sets of all strategies, all stationary strategies, and all deterministic strategies, respectively.

Let R_t and $E_\pi(R_t, s)$ denote, respectively, the random variable representing the immediate reward at time t, and its expectation, when the process begins in state s and the controller follows the strategy π.

The overall reward criterion in the limiting average MDP is defined by

$$J(s, \pi) := \liminf_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} E_\pi(R_t, s), \qquad s \in S, \pi \in C.$$

A strategy π^0 is called optimal if $J(s, \pi^0) = \max_{\pi \in C} J(s, \pi)$ for all $s \in S$.

It is well known that there always exists an optimal deterministic strategy and there are a number of finite algorithms for its computation (e.g., [7], [6], [14]).

We shall now consider the situation where the transition probabilities of Γ are perturbed slightly. Towards this goal we shall define the disturbance law as the set, $D = \{d(s'|s, a)|s, s' \in S, a \in A(s)\}$, where the elements of D satisfy: i) $\sum_{s' \in S} d(s'|s, a) = 0$ for all $s \in S$, $a \in A(s)$; and ii) there exists $\epsilon_0 > 0$ such that for all $\epsilon \in [0, \epsilon_0]$, $p(s'|s, a) + \epsilon d(s'|s, a) \geq 0$, for all $s, s' \in S$, $a \in A(s)$.

Note that D is more general than the perturbation permitted by Delebecque [5] where it is assumed that $d(s'|s, a) \geq 0$ whenever $s' \neq s$.

We now have a family of perturbed finite Markovian decision processes Γ_ϵ for all $\epsilon \in [0, \epsilon_0]$ that differ from the original MDP Γ only in the transition law, namely, in Γ_ϵ we have $p_\epsilon(s'|s, a) := p(s'|s, a) + \epsilon d(s'|s, a)$ for all $s, s' \in S$, $a \in A(s)$.

The limiting average Markov decision problem corresponding to Γ_ϵ is the optimization problem

$$J_\epsilon(s) := \max_{\pi \in C} J_\epsilon(s, \pi), \qquad s \in S \qquad (L_\epsilon)$$

where $J_\epsilon(s, \pi)$ is defined in Γ_ϵ, in the same way as $J(s, \pi)$ was defined in Γ.

For every strategy $\pi \in C$, we define $J^0(s,\pi) := \liminf_{\epsilon\to 0} J_\epsilon(s,\pi)$, $s \in S$. The optimization problem (L): $J^0(s) := \max_{\pi\in C} J^0(s,\pi)$, $s \in S$, is called the limit Markov control problem. Note that the optimization problem (L) is the natural problem to attempt to solve in the case of a slightly perturbed Markov decision process. Our objective in the next section is twofold. First we want to show that the optimization problem (L) has an optimal deterministic strategy. Next we want to demonstrate the validity of the so-called limit control principle in the present framework, that is, we want to show that an optimal strategy in (L) is δ-optimal in Γ_ϵ for any $\delta > 0$ and ϵ sufficiently small.

III. Asymptotics of the Perturbed MDP and the Limit Control Principle

For every $\pi \in C(S)$ we define: the Markov matrix $P(\pi) = (p_{ss'}(\pi))_{s,s'=1}^N$ where $p_{ss'}(\pi) := \sum_{a\in A(s)} p(s'|s,a)\pi(s,a)$, for all $s, s' \in S$; the Markov matrix $P_\epsilon(\pi) = (p_{ss'}^\epsilon(\pi))_{s,s'=1}^N$ where $p_{ss'}^\epsilon(\pi) := \sum_{a\in A(s)} p_\epsilon(s'|s,a)\pi(s,a)$ for all $s, s' \in S$; the perturbation generator matrix $D(\pi) = (d_{ss'}(\pi))_{s,s'=1}^N$ where $d_{ss'}(\pi) := \sum_{a\in A(s)} d(s'|s,a)\pi(s,a)$; and the Cesaro-limit matrix of $P_\epsilon(\pi)$, $P_\epsilon^*(\pi) = (p_{ss'}^{*\epsilon}(\pi))_{s,s'=1}^N := \lim_{t\to\infty} (1/t+1)\sum_{k=0}^t P_\epsilon^k(\pi)$ where $P_\epsilon^0(\pi) := I_N$, an $N \times N$ identity matrix. The Cesaro-limit matrix $P^*(\pi)$ of $P(\pi)$ is defined similarly. Note that for every $\pi \in C(S)$, $P_\epsilon(\pi) = P(\pi) + \epsilon D(\pi)$. Now we shall show that for every $\pi \in C(S)$, $P_\epsilon^*(\pi)$ has a limit as ϵ goes to 0. Our proof uses the following two lemmas.

Lemma 3.1: Let $A = (a_{ss'})_{s,s'=1}^N$ and $B = (b_{ss'})_{s,s'=1}^N$ be two stochastic matrices satisfying: $a_{ss'} = 0 \Leftrightarrow b_{ss'} = 0$, then A and B have the same ergodic classes and the same transient class.

Proof: We refer the reader to [1]; for related results also see [18]. □

Lemma 3.2: For any $\pi \in C(S)$, there exists $\bar\epsilon \in (0, \epsilon_0]$ such that for any $\epsilon \in (0, \bar\epsilon)$: $[P_{\bar\epsilon}(\pi)]_{ss'} = 0 \Leftrightarrow [P_\epsilon(\pi)]_{ss'} = 0$.

Proof: $P_\epsilon(\pi)$ is linear in ϵ and nonnegative for any $\epsilon \in [0, \epsilon_0]$. □

Theorem 3.1: For any stationary strategy $\pi \in C(S)$, the limit stationary matrix $\hat P(\pi) := \lim_{\epsilon\to 0} P_\epsilon^*(\pi)$ exists.[1]

Proof: Let $\pi \in C(S)$. From Lemmas 3.1 and 3.2 it follows that the family of Markov matrices $\{P_\epsilon(\pi) | \epsilon \in (0, \bar\epsilon)\}$ has the same ergodic classes: say $E_1, \cdots, E_m$ and the same transient class: say T.

Consider a standard algorithm (e.g., see [14, p. 28]) for the computation of $P_\epsilon^*(\pi)$. Determine for $k = 1, \cdots, m$:

i) the unique solution $\{x_s^k | s \in E_k\}$ of the linear system

$$\sum_{s\in E_k} (\delta_{ss'} - [P_\epsilon(\pi)]_{ss'})x_s^k = 0, \; s' \in E_k; \qquad \sum_{s\in E_k} x_s^k = 1 \quad (3.1)$$

ii) the unique solution $\{a_{s'}^k | s' \in T\}$ of the linear system

$$\sum_{s'\in T} (\delta_{ss'} - [P_\epsilon(\pi)]_{ss'})a_{s'}^k = \sum_{s'\in E_k} [P_\epsilon(\pi)]_{ss'}; \qquad s \in T. \quad (3.2)$$

Then $P^*_\epsilon(\pi)$ is obtained by

$$[P_\epsilon^*(\pi)]_{ss'} = \begin{cases} x_{s'}^k, & s, s' \in E_k, k = 1, \cdots, m \\ a_s^k x_{s'}^k, & s \in T, s' \in E_k, k = 1, \cdots, m \\ 0, & \text{elsewhere.} \end{cases} \quad (3.3)$$

[1] The same result was proved by Delebecque [5], but for a more restrictive disturbance law, and by a more complicated technique. However, Delebecque derives an explicit expression for $\hat P^*(\pi)$.

Since (3.1) and (3.2) are linear systems with affine linear coefficients, their solutions x_s^k and $a_{s'}^k$ are rational functions of ϵ. It follows from (3.3) that the entries of $P_\epsilon^*(\pi)$ are also rational functions of ϵ. Therefore, $P_\epsilon^*(\pi)$ has a limit when ϵ tends to 0 because its entries are bounded. □

Remark 3.1: Theorem 3.1 can be extended easily (following an analogous proof) to the case where the disturbance law D is of the form $D_0 + \epsilon D_1 + \epsilon^2 D_2 + \cdots$, depending analytically and locally on the parameter ϵ.

With every $\pi \in C(S)$ we associate the vector of single stage expected rewards $r(\pi) = (r_1(\pi), \cdots, r_N(\pi))^T$ in which $r_s(\pi) := \sum_{a\in A(s)} r(s,a)\pi(s,a)$ for each $s \in S$. It is well known that for each stationary strategy $\pi \in C(S)$

$$J_\epsilon(s,\pi) = [P_\epsilon^*(\pi)r(\pi)]_s, \qquad s \in S. \quad (3.4)$$

The following proposition shows that the limit Markov control problem (L) can be restricted to the class $C(D)$ of deterministic strategies.

Proposition 3.1: For any strategy $\pi \in C$, there exists a deterministic strategy $f \in C(D)$ such that $J^0(s,\pi) \le J^0(s,f)$ for each $s \in S$.

Proof: Let $\pi \in C$. Let $\{\epsilon_n\}_{n=1}^\infty$ be any sequence in $(0, \epsilon_0]$ which converges to 0. From Markov decision theory we have that for any n there exists $f_{\epsilon_n} \in C(D)$ such that $J_{\epsilon_n}(s,\pi) \le J_{\epsilon_n}(s, f_{\epsilon_n})$ for each $s \in S$. Since $C(D)$ is finite, there must exist a deterministic strategy $f \in C(D)$ and a subsequence $\{\epsilon_{n_k}\}_{k=1}^\infty$ of the sequence $\{\epsilon_n\}_{n=1}^\infty$ such that $J_{\epsilon_{n_k}}(s,\pi) \le J_{\epsilon_{n_k}}(s,f)$ for each k and s. It follows that

$$J^0(s,\pi) := \liminf_{\epsilon\to 0} J_\epsilon(s,\pi) \le \liminf_{k\to\infty} J_{\epsilon_{n_k}}(s,\pi) \le \liminf_{k\to\infty} J_{\epsilon_{n_k}}(s,f).$$

From (3.4) and Theorem 3.1, it follows that

$$\liminf_{k\to\infty} J_{\epsilon_{n_k}}(s,f) = \liminf_{k\to\infty} \left[P^*_{\epsilon_{n_k}}(f)\mathbf{r}(f)\right]_s = \lim_{\epsilon\to 0} [P_\epsilon^*(f)\mathbf{r}(f)]_s = J^0(s,f). \qquad \square$$

Remark 3.2: From the results above, it follows that the problem (L) can be restricted to the following optimization problem (L'): $\max_{\pi\in C(D)} [\hat P^*(\pi)r(\pi)]_s$, $s \in S$. Any maximizing strategy for (L') is also a maximizing strategy for (L).

The next theorem shows in particular that the problem (L) has an optimal deterministic strategy.

Theorem 3.2: There exist a deterministic strategy $f^0 \in C(D)$ and a positive number δ, such that for any $\epsilon \in (0, \delta)$ f^0 is a maximizer in (L_ϵ). Moreover, f^0 is a maximizer in (L).

Proof: From Markov decision theory, for any $\epsilon \in (0, \epsilon_0)$ there exists an optimal deterministic strategy $f_\epsilon^0 \in C(D)$ for the problem (L_ϵ). Since the class $C(D)$ is finite, there exist a deterministic strategy f^0 and a sequence $\{\epsilon_n\}_{n=1}^\infty$ in $(0, \epsilon_0)$ which converges to 0 such that f^0 is an optimal strategy in (L_{ϵ_n}) for all n. Thus, $[P^*_{\epsilon_n}(f^0)r(f^0)]_s \ge [P^*_{\epsilon_n}(g)r(g)]_s$ for all $n, s \in S, g \in C(D)$. From the proof of Theorem 3.1, it can be seen that $[P^*_{\epsilon_n}(f^0)r(f^0)]_s$ and $[P^*_{\epsilon_n}(g)r(g)]_s$ are rational functions of ϵ_n for n large. Therefore, there exists $\epsilon(s,g) \in (0, \epsilon_0)$ such that $[P_\epsilon^*(f^0)r(f^0)]_s \ge [P_\epsilon^*(g)r(g)]_s$ for any $\epsilon \in (0, \epsilon(s,g))$. Define $\delta := \min\{\epsilon(s,g) | s \in S, g \in C(D)\}$. Now we have $[P_\epsilon^*(f^0)r(f^0)]_s \ge [P_\epsilon^*(g)r(g)]_s$ for all $\epsilon \in (0, \delta), s \in S, g \in C(D)$. This proves the first part of the theorem. For the second part let $\epsilon \to 0$, then

Theorem 3.1 implies that $[\hat{P}^*(f^0)r(f^0)]_s \geq [\hat{P}^*(g)r(g)]_s$ for all $s \in S, g \in C(D)$. □

Corollary 3.1 (Limit Control Principle): Let $\pi^0 \in C(D)$ be any maximizer in (L). Then for all $\beta > 0$ there exists $\epsilon_\beta > 0$ such that for all $\epsilon \in (0, \epsilon_\beta)$, $|J_\epsilon(s, \pi^0) - J_\epsilon(s)| < \beta$ for all $s \in S$.

Proof: Let $\epsilon \in (0, \delta)$. By Theorem 3.2, for all $s \in S$, we have $|J_\epsilon(s, \pi^0) - J_\epsilon(s)| = |J_\epsilon(s, \pi^0) - J^0(s, \pi^0) + J^0(s, f^0) - J_\epsilon(s, f^0)| \leq |J_\epsilon(s, \pi^0) - J^0(s, \pi^0)| + |J^0(s, f^0) - J_\epsilon(s, f^0)|$, where f^0 is as in Theorem 3.2. In view of Theorem 3.1 and (3.4) we conclude that for all $\beta > 0$ there exists $\epsilon_\beta > 0$ such that for all $\epsilon \in (0, \epsilon_\beta)$, $|J_\epsilon(s, \pi^0) - J^0(s, \pi^0)| < (\beta/2)$ and $|J^0(s, f^0) - J_\epsilon(s, f^0)| < (\beta/2)$ for all $s \in S$. This proves the corollary. □

Remark 3.3: A problem of interest is to find an optimal deterministic strategy for the limit Markov control problem (L) (which exists by Theorem 3.2). In the case of completely decomposable Markov control problems, in [2] we give two methods for the computation of such an optimal strategy.

Let $\{\epsilon_n\}_{n=1}^\infty$ be any sequence in $(0, \epsilon_0]$ converging to 0. We define the following sequence of strategies:

i) choose f_0 arbitrary;

ii) for $n \geq 1$, we define f_n as an optimal strategy in the perturbed MDP (L_{ϵ_n}) obtained by the policy improvement algorithm with f_{n-1} as the starting strategy.

Proposition 3.2: There exists n^* such that for any $n \geq n^*$, $f_n = f_{n^*}$ and f_{n^*} is optimal in the limit Markov control problem (L).

Proof: Since the sequence $\{f_n\}_{n=1}^\infty$ is in $C(D)$ which is finite, then this sequence has a limit point f^*. By the same argument as in the proof of Theorem 3.2, it follows that there exists $\delta \in (0, \epsilon_0)$ such that f^* is optimal in (L_ϵ) for any $\epsilon \in (0, \delta)$. Let $n(\delta)$ be such that $\epsilon_n \in (0, \delta)$ for all $n \geq n(\delta)$. Since f^* is a limit point, then there exists $n^* \geq n(\delta)$ such that $f^* = f_{n^*}$. Now, by definition of the sequence $\{f_n\}_{n=1}^\infty$, we have $f_n = f_{n+1}$ for all $n \geq n^*$. □

Remark 3.4: The procedure for generating the sequence $\{f_n\}_{n=1}^\infty$ gives a heuristic for finding an optimal strategy for (L).

IV. Discounted Case

In this section we shall show that the perturbation in the discounted case can be analyzed by solving the original problem.

The discounted Markov decision problem corresponding to Γ_ϵ is defined by

$$V_\epsilon(s) := \max_{\pi \in C} V_\epsilon(s, \pi), \qquad s \in S \qquad (\mathrm{DP}_\epsilon)$$

where $V_\epsilon(s, \pi) := \sum_{t=0}^\infty \alpha^t E_\pi(R_t, s)$, $\alpha \in (0, 1)$ is the discount factor.

It is well known that for any $\epsilon \in [0, \epsilon_0]$, there always exists an optimal deterministic strategy for the problem (DP_ϵ); and there are a number of finite algorithms for its computation (e.g., [7], [6], [14]).

If $\pi \in C(S)$, it is well known that

$$V_\epsilon(s, \pi) = \left([I_N - \alpha P_\epsilon(\pi)]^{-1} r(\pi)\right)_s, \qquad s \in S. \quad (4.1)$$

Lemma 4.1: Let $\{M_\epsilon | \epsilon > 0\}$ be a family of nonsingular matrices. If $M_\epsilon \to M$ as $\epsilon \to 0$ and M is nonsingular, then $M_\epsilon^{-1} \to M^{-1}$ as $\epsilon \to 0$.

Proof: For any $\epsilon > 0$, $M_\epsilon^{-1} = (1/\det M_\epsilon)\operatorname{adj}(M_\epsilon)$. Since $\det(\cdot)$ is a continuous function, $\det(M_\epsilon) \to \det(M)$ and $\operatorname{adj}(M_\epsilon) \to \operatorname{adj}(M)$ as $\epsilon \to 0$. This proves the proposition since M is nonsingular. □

Theorem 4.1: There exist a deterministic strategy f^0 and a positive number δ such that for any $\epsilon \in [0, \delta)$, f^0 is a maximizer in (DP_ϵ).

Proof: The proof is along analogous lines to that of Theorem 3.2. The key observations are that by (4.1) $V_\epsilon(s, \pi)$ is a rational function of ϵ and that the correct limit is obtained by Lemma 4.1. For details we refer the reader to [1]. □

The next result is the limit control principle for the discounted case.

Corollary 4.1: Let $\pi^0 \in C(D)$ be any maximizer in the original problem (DP_0). Then for all $\beta > 0$, there exists $\epsilon_\beta > 0$ such that $|V_\epsilon(s, \pi^0) - V_\epsilon(s)| < \beta$ for all $\epsilon \in (0, \epsilon_\beta)$ and $s \in S$.

Proof: Let $\epsilon \in (0, \delta)$. By Theorem 4.1, for all $s \in S$ we have $|V_\epsilon(s, \pi^0) - V_\epsilon(s)| = |V_\epsilon(s, \pi^0) - V_0(s, \pi^0) + V_0(s, f^0) - V_\epsilon(s, f^0)| \leq |V_\epsilon(s, \pi^0) - V_0(s, \pi^0)| + |V_0(s, f^0) - V_\epsilon(s, f^0)|$, where f^0 is as in Theorem 4.1. In view of (4.1) and Lemma 4.1, the result follows. □

Remark 4.1: All the results in Sections III and IV can be extended easily (following analogous proofs) to the case where the disturbance law D is of the form $D_0 + \epsilon D_1 + \cdots + \epsilon^n D_n$ and n is any natural number or, more generally, if $D := D(\epsilon)$ is a rational function of ϵ.

V. The Generalized Disturbance Law

In this section, we consider the general perturbation

$$p_d(s'|s, a) := p(s'|s, a) + d(s'|s, a), \qquad s, s' \in S; a \in A(s). \quad (5.1)$$

Define $\|d\| := \max\{|d(s'|s, a)| \, | s, s' \in S; a \in A(s)\}$.

We assume that there exists $\epsilon_0 > 0$ such that for any d satisfying $\|d\| \leq \epsilon_0$, p_d is a transition probability, that is, for any $s, s' \in S$ and $a \in A(s)$, $p_d(s'|s, a) \geq 0$ and $\sum_{s' \in S} p_d(s'|s, a) = 1$.

In general, $P_d^*(\pi)$ may not have a limit when $\|d\|$ tends to 0 as is illustrated by the following example.

Example: Let $d = (d_1, d_2)$; $d_1, d_2 \in (0, 1)$. The perturbed transition matrix is given by

$$P_d = \begin{pmatrix} 1 - d_1 & d_1 \\ d_2 & 1 - d_2 \end{pmatrix}.$$

The stationary distribution of P_d is

$$P_d^* = \begin{pmatrix} \dfrac{d_2}{d_1 + d_2} & \dfrac{d_1}{d_1 + d_2} \\ \dfrac{d_2}{d_1 + d_2} & \dfrac{d_1}{d_1 + d_2} \end{pmatrix}$$

but $(d_1/d_1 + d_2)$ has no limit as $\|d\|$ tends to 0.

However, if $\pi \in C(S)$ is unichain in Γ_0, that is the corresponding Markov matrix $P(\pi)$ has one ergodic class plus (perhaps empty) class of transient states, then $P_d^*(\pi)$ has a limit when $\|d\|$ tends to 0 as is shown in the following proposition.

Proposition 5.1: Let $\pi \in C(S)$ be unichain in Γ_0, then $\lim_{\|d\| \to 0} P_d^*(\pi) = P^*(\pi)$.

Proof: Let $\{d_n\}_{n=1}^\infty$ be any sequence satisfying: i) $\|d_n\| \leq \epsilon_0$ for all n; ii) $\|d\|$ converges to 0 and; iii) $\lim_{n \to \infty} P_{d_n}^*(\pi)$ exists. Since $P_{d_n}^*(\pi) P_{d_n}(\pi) = P_{d_n}^*(\pi)$ for all n, then

$(\lim_{n\to\infty} P^*_{d_n}(\pi))P(\pi) = \lim_{n\to\infty} P^*_{d_n}(\pi)$. It follows that $\lim_{n\to\infty} P^*_{d_n}(\pi) = P^*(\pi)$ because π is unichain. Now assume that $P^*_d(\pi)$ does not go to $P^*(\pi)$ as $\|d\| \to 0$. Then there exists $\epsilon > 0$ such that for any n there exists d_n satisfying $\|d_n\| \le (\epsilon_0/n)$ and $\|P^*_{d_n}(\pi) - P^*(\pi)\| > \epsilon$. Since the sequence $\{\|P^*_{d_n}(\pi)\|\}_{n=1}^{\infty}$ is bounded, there exists a subsequence $\{d_{n_k}\}_{k=1}^{\infty}$ of the sequence $\{d_n\}_{n=1}^{\infty}$ satisfying i)–iii). Thus, $\lim_{k\to\infty} P^*_{d_{n_k}}(\pi) = P^*(\pi)$. This is a contradiction to $\|P^*_{d_{n_k}}(\pi) - P^*(\pi)\| > \epsilon$ for all k. □

The MDP Γ is called unichain if every stationary strategy in Γ is unichain.

The MDP Γ is called communicating if for any $(s, s') \in S \times S$, there exist a deterministic strategy $f \in C(D)$ and a natural number n such that $[P^n(f)]_{ss'} > 0$. Let Γ_d be the MDP defined by Γ except for the transition law which is defined by (5.1).

In the remainder of this section we shall prove that in the unichain case, the communicating case, and the discounted case, the limit control problem (L) is equivalent to the original problem (L_0) in the sense that any optimal deterministic strategy in (L_0) is δ-optimal in Γ_d for any $\delta > 0$ when $\|d\|$ is sufficiently small.

The limiting average Markov decision problem corresponding to Γ_d is defined by

$$J_d(s) := \max_{\pi \in C(S)} [P^*_d(\pi)r(\pi)]_s, \qquad s \in S. \tag{L_d}$$

The generalized limit control principle for the unichain case is stated as follows.

Theorem 5.1: Let $\pi^0 \in C(S)$ be any maximizer in the original problem (L_0). Then for all $\beta > 0$, there exists $\epsilon_\beta > 0$ such that for all d satisfying $\|d\| < \epsilon_\beta$, $\|P^*_d(\pi^0)r(\pi^0) - J_d\| < \beta$.

Proof: Let d be such that $\|d\| \le \epsilon_0$. We have

$$\begin{aligned}
&\|P^*_d(\pi^0)r(\pi^0) - J_d\| \\
&\quad = \|P^*_d(\pi^0)r(\pi^0) - P^*(\pi^0)r(\pi^0) \\
&\qquad + P^*(\pi^0)r(\pi^0) - J_d\| \\
&\quad \le \|P^*_d(\pi^0)r(\pi^0) - P^*(\pi^0)r(\pi^0)\| \\
&\qquad + \|P^*(\pi^0)r(\pi^0) - J_d\|
\end{aligned}$$

and

$$\begin{aligned}
&\|P^*(\pi^0)r(\pi^0) - J_d\| \\
&\quad = \|\max_{\pi\in C(S)} P^*(\pi)r(\pi) - \max_{\pi\in C(S)} P^*_d(\pi)r(\pi)\| \\
&\quad = \|\max_{\pi\in C(D)} P^*(\pi)r(\pi) - \max_{\pi\in C(D)} P^*_d(\pi)r(\pi)\| \\
&\quad \le \max_{\pi\in C(D)} \|P^*(\pi)r(\pi) - P^*_d(\pi)r(\pi)\|.
\end{aligned}$$

Now since the class $C(D)$ is finite, the theorem follows from Proposition 5.1. □

Lemma 5.1: If Γ is communicating then there exists $\bar{\epsilon} > 0$ such that for any d satisfying $\|d\| < \bar{\epsilon}$, Γ_d is communicating.

Proof: Define $\delta := \min\{p(s'|s, a) | p(s'|s, a) > 0,\ s, s' \in S, a \in A(s)\}$, and $\bar{\epsilon} := \min\{\delta, \epsilon_0\}$, then Γ_d must be communicating when $\|d\| < \bar{\epsilon}$. □

Remark 5.1: Recall that any communicating MDP possesses a unichain deterministic strategy which is optimal with respect to the limiting average criterion (e.g., see [13]).

The generalized limit control principle for the communicating case is stated as follows.

Theorem 5.2: Let $\pi^0 \in C(D)$ be any maximizer unichain strategy in the original problem (L_0). Then for all $\beta > 0$, there exists $\epsilon_\beta > 0$ such that for all d satisfying $\|d\| < \epsilon_\beta$, $\|P^*_d(\pi^0)r(\pi^0) - J_d\| < \beta$.

Proof: The proof is by contradiction. Assume that there exists $\beta > 0$ such that for all n, there exists d_n satisfying

$$\|d_n\| < \frac{\bar{\epsilon}}{n} \text{ and } \|P^*_{d_n}(\pi^0)r(\pi^0) - J_{d_n}\| \ge \beta \tag{5.2}$$

where $\bar{\epsilon}$ is as in Lemma 5.1.

From Lemma 5.1, it follows that the MDP Γ_{d_n} is communicating for all n, and hence (from Remark 5.1) for all n, Γ_{d_n} has an optimal unichain deterministic strategy g_n. Since the class $C(D)$ is finite, there exist a subsequence $\{(1/n_k)\}_{k=1}^{\infty}$ of the sequence $\{(1/n)\}_{n=1}^{\infty}$, and a deterministic strategy g which is unichain and optimal in $\Gamma_{d_{n_k}}$ for all k. Thus

$$\left[P^*_{d_{n_k}}(g)r(g)\right]_s \ge \left[P^*_{d_{n_k}}(f)r(f)\right]_s$$

$$\text{for all } k, s \in S, f \in C(D). \tag{5.3}$$

Since the sequence $\{P^*_{d_{n_k}}\}_{k=1}^{\infty}$ is bounded, there exists a subsequence $\{n_{k_l}\}_{l=1}^{\infty}$ of the sequence $\{n_k\}_{k=1}^{\infty}$ such that $\lim_{l\to\infty} P^*_{d_{n_{k_l}}}(g) := \hat{P}^*(g)$ exists. From (5.3), we have

$$\left[P^*_{d_{n_{k_l}}}(g)r(g)\right]_s \ge \left[P^*_{d_{n_{k_l}}}(f)r(f)\right]_s$$

$$\text{for all } l, s \in S, f \in C(S). \tag{5.4}$$

Let $l \to \infty$ in (5.4), then from Proposition 5.1 we have

$$[\hat{P}^*(g)r(g)]_s \ge [P^*(f)r(f)]_s,$$

$$\text{for any unichain strategy } f \text{ and any } s \in S. \tag{5.5}$$

Note that g need not be unichain in the original MDP Γ_0. Hence, let $S^1, \cdots, S^K$ be the ergodic sets with respect to $P(g)$, and for each $k \in \{1, \cdots, K\}$, let $\tilde{q}^k(g)$ be the unique stationary distribution of the restriction of $P(g)$ to S^k, and define $q^k_s(g) := \tilde{q}^k_s(g)$ for $s \in S^k$ and $q^k_s(g) := 0$ for $s \in S \setminus S^k$. Note that since g is unichain in $\Gamma_{d_{n_{k_l}}}$ for all l, the rows of $P^*_{d_{n_{k_l}}}(g)$ are identical. Let $p^*_{d_{n_{k_l}}}(g)$ be a row of $P^*_{d_{n_{k_l}}}(g)$. We have $p^*_{d_{n_{k_l}}}(g)P_{d_{n_{k_l}}}(g) = p^*_{d_{n_{k_l}}}(g)$ for all l. When l tends to infinity, we get $\hat{p}^*(g)P(g) = \hat{p}^*(g)$, where $\hat{p}^*(g) := \lim_{l\to\infty} p^*_{d_{n_{k_l}}}(g)$. Therefore $\hat{p}^*(g)$ is a fixed probability vector of $P(g)$ and hence there exist $\mu_1, \cdots, \mu_K$ satisfying $\sum_{k=1}^K \mu_k = 1$ and $\mu_k \ge 0$ for all $k \in \{1, \cdots, K\}$ such that $\hat{p}^*(g) = \sum_{k=1}^K \mu_k q^k(g)$. Define $\bar{k} := \arg\max\{q^k(g)r(g) | k = 1, \cdots, K\}$. Since the MDP Γ is communicating, we can easily construct a deterministic strategy $\bar{g}$ with the single ergodic set $S^{\bar{k}}$, which coincides with the strategy g in $S^{\bar{k}}$. Note that $p^*(\bar{g})r(\bar{g}) = q^{\bar{k}}(g)r(g)$ and $\hat{p}^*(g)r(g) = \sum_{k=1}^K \mu_k q^k(g)r(g) \le q^{\bar{k}}(g)r(g) = p^*(\bar{g})r(\bar{g})$. Now it follows from (5.5) that $\hat{p}^*(g)r(g) = p^*(\bar{g})r(\bar{g})$, that is, $\lim_{l\to\infty} P^*_{d_{n_{k_l}}}(g)r(g) = P^*(\bar{g})r(\bar{g})$, and by (5.5) $\bar{g}$ is the maximizer in (L_0).

Finally, we have for all l,

$$\begin{aligned}\|P^*_{d_{n_{kl}}}(\pi^0)r(\pi^0) - J_{d_{n_{kl}}}\| &= \|P^*_{d_{n_{kl}}}(\pi^0)r(\pi^0) - P^*(\pi^0)r(\pi^0) \\ &\quad + P^*(\bar{g})r(\bar{g}) - P^*_{d_{n_{kl}}}(g)r(g)\| \\ &\le \|P^*_{d_{n_{kl}}}(\pi^0)r(\pi^0) - P^*(\pi^0)r(\pi^0)\| \\ &\quad + \|P^*(\bar{g})r(\bar{g}) - P^*_{d_{n_{kl}}}(g)r(g)\|\end{aligned}$$

which goes to 0 as l goes to infinity. This is a contradiction with (5.2). □

The discounted Markov decision problem corresponding to Γ_d is defined by

$$V_d(s) := \max_{\pi \in C} V_d(s, \pi) \qquad s \in S \qquad (\mathrm{DP}_d)$$

where $V_d(s, \pi) := \sum_{t=0}^{\infty} \alpha^t E_\pi(R_t, s)$.

We define $R := \max\{|r(s, a)| \mid s \in S, a \in A(s)\}$. Let $\|a(\cdot)\|$ denote any vector norm of $a = (a(1), \cdots, a(N))^T$. The following theorem shows that the limit control principle for the discounted case is also valid if we consider the general perturbation.

Theorem 5.3: Let $\pi^0 \in C(S)$ be any maximizer in the original (DP_0). Then for all $\beta > 0$, there exists $\epsilon_\beta > 0$ such that for all d satisfying $\|d\| < \epsilon_\beta, \|V_d(\cdot, \pi^0) - V_d(\cdot)\| < \beta$.

Proof: From Markov decision theory, for any $\pi \in C(S)$: $V_d(s, \pi) = [r(\pi)]_s + \alpha \sum_{s' \in S}[P_d(\pi)]_{ss'} V_d(s', \pi), s \in S$. It follows that for any $s \in S$,

$$\begin{aligned}|V_d(s, \pi) - V_0(s, \pi)| &= \alpha \Big| \sum_{s' \in S} [P_d(\pi)]_{ss'} V_d(s', \pi) - \sum_{s' \in S} [P(\pi)]_{ss'} V_0(s', \pi) \Big| \\ &= \alpha \Big| \sum_{s' \in S} [D(\pi)]_{ss'} V_d(s', \pi) \\ &\quad + \sum_{s' \in S} [P(\pi)]_{ss'} (V_d(s', \pi) - V_0(s', \pi)) \Big| \\ &\le \alpha((NR\|d\|/1 - \alpha) + \|V_d(\cdot, \pi) - V_0(\cdot, \pi)\|).\end{aligned}$$

Now, for any $s \in S$,

$$\begin{aligned}|V_0(s, \pi^0) - V_d(s)| &= \Big| \max_{\pi \in C(S)} V_0(s, \pi) - \max_{\pi \in C(S)} V_d(s, \pi) \Big| \\ &\le \max_{\pi \in C(S)} |V_0(s, \pi) - V_d(s, \pi)| \\ &\le (\alpha N \|d\| R/(1 - \alpha)^2).\end{aligned}$$

Hence $\|V_0(\cdot, \pi^0) - V_d(\cdot)\| \le (\alpha N\|d\|R/(1 - \alpha)^2)$. Finally,

$$\begin{aligned}\|V_d(\cdot, \pi^0) - V_d(\cdot)\| &\le \|V_d(\cdot, \pi^0) - V_0(\cdot, \pi^0)\| \\ &\quad + \|V_0(\cdot, \pi^0) - V_d(\cdot)\| \le (2\alpha N\|d\|R/(1 - \alpha)^2).\end{aligned}$$

This proves the theorem. □

VI. Application: Approximating Models

Let $\Gamma_t = \langle S, A, q_t, r \rangle$, where $t = 1, 2, \cdots$, be a sequence of MDP's. Also let $\Gamma = \langle S, A, q, r \rangle$ be the limit, where $q := \lim_{t \to \infty} q_t$.

For the limiting average (discounted) overall reward criterion we shall define the optimal reward from state s to be $J(s)$ $(v(s))$.

The problem of approximating models is: when does the optimal reward of the MDP Γ_t converge to the optimal reward of the limit MDP Γ?

Hernandez-Lerma [12] solved this problem for the discounted case, and for the limiting average case restricted by a rather strong ergodicity assumption. Below we provide an answer for the case of general communicating limiting average MDP.

Let $J_t(\cdot)$ and $J(\cdot)$ denote the vectors of average-optimal rewards in the MDP's Γ_t and Γ, respectively.

Theorem 6.1: If the limit MDP Γ is communicating then $\lim_{t \to \infty} J_t(\cdot) = J(\cdot)$.

Proof: From Theorem 5.2, it follows that $\lim_{\|d\| \to 0} J_d = \lim_{\|d\| \to 0} P^*_d(\pi^0)r(\pi^0) = P^*(\pi^0)r(\pi^0) := J_0$. The second equality follows from Proposition 5.1 since π^0 is unichain in Γ_0. Set $d_t := q_t - q$, hence, we have $J_t = J_{d_t}$ and $J = J_0$. This proves the theorem. □

Remark 6.1: An analog to Theorem 6.1 for the unichain case can be derived from Theorem 5.1 in the same way as Theorem 6.1 is derived from Theorem 5.2.

Acknowledgment

The authors are indebted to A. Shwartz for his comments and suggestions of improvement.

References

[1] M. Abbad, "Perturbation and stability theory for Markov control problems," Ph.D. dissertation, Univ. Maryland at Baltimore County, 1991.
[2] M. Abbad, T. Bielecki, and J. A. Filar, "Algorithms for singularly perturbed limiting average Markov control problems," *IEEE Trans. Automat. Contr.*, to be published.
[3] T. Bielecki and J. A. Filar, "Singular perturbations of Markov decision chains," *Annals Operations Research*, to be published.
[4] D. Blackwell, "Discrete dynamic programming," *Annals Math. Statist.*, vol. 33, pp. 719–726, 1962.
[5] F. Delebecque, "A reduction process for perturbed Markov chains," *SIAM J. Appl. Math.*, vol. 48, pp. 325–350, 1983.
[6] E. V. Denardo, *Dynamic Programming*. Englewood Cliffs, NJ: Prentice-Hall, 1982.
[7] C. Derman, *Finite State Markovian Decision Process*. New York: Academic, 1970.
[8] N. V. Dijk, "Perturbation theory for unbounded Markov reward processes with applications to queueing," *Adv. Appl. Prob.*, vol. 20, pp. 99–111, 1988.
[9] N. V. Dijk and M. Puterman, "Perturbation theory for Markov reward processes with applications to queueing systems," *Adv. Appl. Prob.*, vol. 20, pp. 79–98, 1988.
[10] A. Federgruen, "Markovian control problems," in *Mathematical Centre Tracts Vol. 97*. Amsterdam, The Netherlands: Centre for Mathematics and Computer Science, 1983.
[11] J. A. Filar and T. A. Schultz, "Communicating MDP's: Equivalence, and LP-properties," *Operations Res. Lett.*, vol. 7, pp. 303–307, 1988.
[12] O. Hernandez-Lerma, "Adaptive Markov control processes," in *Applied Mathematical Sciences Vol. 79*. New York: Springer-Verlag, 1989.
[13] R. A. Howard, *Dynamic Programming and Markov Processes*. Cambridge, MA: M.I.T. Press, 1960.
[14] L. C. M. Kallenberg, "Linear programming and finite Markovian control problems," in *Mathematical Centre Tracts Vol. 148*. Amsterdam: The Netherlands: Centre for Mathematics and Computer Science, 1983.
[15] P. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. Englewood Cliffs, NJ: Prentice-Hall, 1986.
[16] S. M. Ross, *Introduction to Stochastic Dynamic Programming*. New York: Academic, 1983.
[17] P. J. Schweitzer, "Perturbation theory and finite Markov chains," *J. Appl. Probability*, vol. 5, pp. 401–413, 1968.
[18] E. Seneta, *Non-Negative Matrices and Markov Chains*. New York: Wiley, 1981.
[19] H. C. Tijms, *Stochastic Modelling and Analysis*. New York: Wiley, 1986.