[9] J. Delforge, "On local identifiability of linear systems," *Math. Biosci.*, vol. 70, pp. 1-37, 1984.
[10] R. E. Kalman, "Mathematical description of linear dynamical system," *SIAM Contr., Série A*, vol. 1, pp. 152-192, 1963.
[11] J. Delforge, "New results to the problem of identifiability of a linear system," *Math. Biosci.*, vol. 52, pp. 73-96, 1980.
[12] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Non Linear Equations in Several Variables*. New York: Academic, 1970.

A Lemma on the Multiarmed Bandit Problem

JOHN N. TSITSIKLIS

Abstract—**We prove a lemma on the optimal value function for the multiarmed bandit problem which provides a simple direct proof of optimality of writeoff policies. This, in turn, leads to a new proof of optimality of the index rule.**

I. The Main Result

We consider the multiarmed bandit problem which has been the subject of considerable interest recently [1]-[3]. We start by introducing the notation to be employed.

There are N projects; at any time, each project i has a state denoted by x^i. We use the vector x defined by $x = (x^1, \cdots, x^N)$ to denote the joint state of all projects. At each time t we have the options of retiring and receiving a retirement reward $\alpha^t M$ (where $\alpha \in (0, 1)$ is the discount rate) or work on one of the projects (say, project i). With this latter action, we receive a reward $\alpha^t R^i(x^i)$ (the functions R^i are assumed bounded); moreover the state of project i changes to $f^i(x^i, w^i)$, where f^i is a known function and w^i is a random disturbance whose probability distribution depends only on x^i. Concerning the projects other than i, their state does not change. The objective is to find a policy (which determines which project is to be worked on, given the current state of all projects) so that the expected infinite horizon discounted reward is maximized.

We denote by $V(x, M)$ this optimal reward, as a function of the joint initial state x and the retirement reward M. We also consider the optimal reward functions for two auxiliary armed bandit problems. Namely, we let $V^i(x^i, M)$ be the optimal reward for the problem in which the only available options are to retire or to work on project i. We also denote by $U^i(y^i, M)$ the optimal reward for a problem in which the available options are to retire or to work on any project other than project i and where y^i denotes the joint state of all projects other than project i.

Our main result is the following.

Lemma 1: $V(x, M) \le V^i(x^i, M) + U^i(y^i, M) - M$.

Proof: Without any loss of generality we assume that $i = 1$. Notice that $x = (x^1, y^1)$. We define recursively functions V_n, V_n^1, U_n^1, for $n \ge 0$ as follows: we let $V_0(x, M) = V_0^1(x^1, M) = U_0^1(y^1, M) = M$, for all x, x^1, y^1. Also,

$$V_{n+1}(x^1, y^1, M) = \max \{M, R^1(x^1) + \alpha E[V_n(f^1(x^1, w^1), M)], \max_{j \ne 1} \{R^j(x^j) + \alpha E[V_n(x^1, F^j(y^1, w^j), M)]\}\} \quad (1)$$

where $F^j(y^1, w^j)$ is a vector with all components equal to those of y^1, except for the component with superscript j which becomes equal to $f^j(x^j, w^j)$. Similarly,

$$V_{n+1}^1(x^1, M) = \max \{M, R^1(x^1) + \alpha E[V_n^1(f^1(x^1, w^1), M)]\}, \quad (2)$$

$$U_{n+1}^1(y^1, M) = \max \{M, \max_{j \ne i} \{R^j(x^j) + \alpha E[U_n^1(F^j(y^1, w^j), M)]\}\}. \quad (3)$$

These are easily recognized to be the equations for the successive approximation algorithm for the original and the two auxiliary multiarmed bandit problems, respectively. Because of the boundedness assumption on the R^i's, it follows [4] that V_n^1, V^1, and U_n^1 converge to V, V^1, and U^1, respectively. It is therefore sufficient to prove that

$$V_n(x^1, y^1, M) \le V_n^1(x^1, M) + U_n^1(y^1, M) - M \quad (4)$$

holds for all n, x^1, y^1.

Equation (4) is trivially true for $n = 0$. Assume it is true for some n; we will demonstrate its validity for $n + 1$ as well. We first notice that $V_n^1(x^1, M) \ge M$, $U_n^1(y^1, M) \ge M$, $V_n^1(x^1, M) \le V_{n-1}^1(x^1, M)$ and $U_n^1(y^1, M) \le U_{n+1}^1(y^1, M)$ which follow from (2), (3) and a straightforward induction. As a consequence, we have

$$\alpha[U_n^1(y^1, M) - M] \le U_n^1(y^1, M) - M \le U_{n-1}^1(y^1, M) - M \quad (5)$$

and

$$\alpha[V_n^1(x^1, M) - M] \le V_n^1(x^1, M) - M \le V_{n+1}^1(x^1, M) - M. \quad (6)$$

Using (1)-(3) and the induction hypothesis we obtain

$$V_{n+1}(x^1, y^1, M) \le \max \{M, R^1(x^1) + \alpha E[U_n^1(y^1, M) + V_n^1(f^1(x^1, w^1), M) - M], \max_{j \ne i} \{R^j(x^j) + \alpha E[U_n^1(F^j(y^1, w^j), M) + V_n^1(x^1, M) - M]\}\}. \quad (7)$$

We will show that all terms in the right-hand side of (7) are less or equal than $V_{n+1}^1(x^1, M) + U_{n+1}^1(y^1, M) - M$. This is certainly true for the first term, which is equal to M because of the inequalities $V_{n+1}^1 \ge M$ and $U_{n+1}^1 \ge M$. Using (5), the second term is bounded above by $R^1(x^1) + \alpha E[V_n^1(f^1(x^1, w^1), M)] + U_{n+1}^1(y^1, M) - M$; this expression is in turn bounded above by $V_{n-1}^1(x^1, M) + U_{n-1}^1(y^1, M) - M$ because of (2). Concerning the last term, we use (6) to obtain, for any $j \ne 1$, the bound $R^j(x^j) + \alpha E[U_n^1(F^j(y^1, w^j), M)] + V_{n+1}^1(x^1, M) - M$; this expression is again bounded above by $V_{n-1}^1(x^1, M) + U_{n+1}^1(y^1, M) - M$, because of (3). ∎

It is obvious from the method of the proof that Lemma 1 admits the following generalization. Let $\{S_1, S_2\}$ be a partition of the set of projects. Let z^1, z^2 be vectors having as components the states of the projects in the sets S^1, S^2, respectively. Let $U^1(z^1, M)$, $U(z^2, M)$ be the optimal reward functions when we are allowed to either retire or work on a project in the set S^1, S^2, respectively. Then,

$$V(x, M) \le U^1(z^1, M) + U^2(z^2, M) - M.$$

II. Optimality of Writeoff Policies

A write off policy has been defined by Whittle [1] as any policy "in which project i is written off (i.e., abandoned) when first its state x^i enters a writeoff set S^i. One continues as long as there are projects which have not been written off, working only on those projects; one retires as soon as all projects are written off. While it is known that writeoff policies are optimal (this is a consequence of the index rule [1]-[3]) no direct proof of this fact was known. However, we show below that this is a simple consequence of Lemma 1.

Let us define $S^i = \{x^i : V^i(x^i, M) = M\}$. Suppose that the state x of the projects is such that $x^i \in S^i$. It follows from Lemma 1 that $V(x, M) \le U^i(y^i, M)$. The reverse inequality is also trivially true. It follows that $V(x, M) = U^i(y^i, M)$. Equivalently, there exists an optimal policy such that one never works on project i, whenever $x^i \in S^i$. On the other hand, if for some project i, $x^i \notin S^i$, then $V(x, M) \ge V^i(x^i, M) > M$, which shows that it is not optimal to retire. We thus conclude that there exists a

Manuscript received December 6, 1985. This work was supported by an IBM Faculty Development Award.

The author is with the Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA 02139.

IEEE Log Number 8608354.

writeoff policy (with the writeoff sets S^i as defined above) which is optimal.

III. On the Proof of the Index Rule

For any project i we define its index m^i (as a function of its state x^i) by $m^i(x^i) = \min \{m: V^i(x^i, m) = m\}$. The index rule states that a policy is optimal if and only if it never works on a project whose index is less than M and whenever it works on a project, then this project has the largest index among all projects. Several proofs of this result are known [1]–[3]. The proof given by Whittle is essentially based on the equality $V(x, M) = B - \int_M^B \Pi_i \partial V^i(x^i, m)/\partial m \, dm$. However, a fairly indirect argument is used to prove this equality. On the other hand, Whittle demonstrates that this equality could be proved directly if there was a direct proof of optimality of writeoff policies. Since such a proof has been given in the previous section, it can be combined with the arguments in [1] for a new and fairly short proof of the index rule.

References

[1] P. Whittle, *Optimization Over Time*. New York: Wiley, 1982.

[2] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Royal Statist. Soc.*, B, vol. 41, pp. 148–164, 1979.

[3] P. P. Varaiya, J. C. Walrand, and C. Buyukkoc, "Extensions of the multiarmed bandit problem: The discounted case," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 426–439, 1985.

[4] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*. New York: Academic, 1976.

Adaptive Stabilization of Single-Input Single-Output Delay Systems

MOHAMMED DAHLEH AND WILLIAM E. HOPKINS, JR.

***Abstract*—For the problem of adaptive stabilization of linear systems with unknown, noncommensurate, real delays, the existence of a smooth controller that regulates the output to zero is proven. The assumptions imposed on the system are natural generalizations of the familiar minimum phase and relative degree one conditions in nondelay systems.**

I. Introduction

The problem of adaptive stabilization of linear systems with unknown high-frequency gains has received a great deal of attention in the past few years. Nussbaum [6] showed that it is possible to construct a smooth nonlinear controller which stabilizes any scalar system in the absence of any knowledge of the sign of the high-frequency gain. This idea was used by Willems and Byrnes [7] to obtain a smooth nonlinear controller capable of stabilizing any nth order single-input, single-output, minimum phase system with relative degree one. Morse [4] showed how to use the idea of Nussbaum to construct controllers for nth-order systems with unknown relative degree not exceeding two. Mudgett and Morse [5] extended the results to systems with arbitrary, but known relative degrees.

This note addresses the problem of extending the previous results to a special class of delay systems. It will be shown that the simple controller of [7] is capable of stabilizing a large class of delay systems with unknown coefficients. Conditions are imposed on the transfer functions of these systems which are similar to relative degree one and minimum phase conditions. This result suggests that it may be possible to treat many infinite-dimensional linear systems with techniques presently used only for finite-dimensional systems.

Manuscript received November 14, 1985; revised February 5, 1986. This work was supported by the National Science Foundation under Grants ECS-8307433 and ECS-8351621.

M. Dahleh is with the Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544.

W. E. Hopkins, Jr. is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544.

IEEE Log Number 8608351.

II. The Adaptive Control Problem

The systems considered in this note are assumed to be modeled by the transfer function

$$g(s, e^{-sh_1}, e^{-sh_2}, \cdots, e^{-sh_m}) = \frac{b_0\left(s^{n-1} + \sum_{j=1}^{n-1}\sum_{i=0}^{m} s^{n-1-j} b_j^i e^{-sh_i}\right)}{s^n + \sum_{j=1}^{n}\sum_{i=0}^{m} s^{n-j} a_j^i e^{-sh_i}} \quad (1)$$

where b_0, b_j^i, h_i, a_j^i, $m \geq 0$, $n \geq 1$ are unknown constants satisfying $b_0 \neq 0$, $0 = h_0 < h_1 < h_2 < \cdots < h_m$. Extending the usual definition of relative degree to be the difference between the highest power of s in the denominator and the highest power of s in the numerator, it is clear that g is of relative degree one. Generalizing the standard minimum phase condition, assume all the zeros of $s^{n-1} + \sum_{j=1}^{n-1}\sum_{i=0}^{m} s^{n-1-j} b_j^i e^{-sh_i}$ have negative real parts. The problem then is to construct a continuous controller which will drive the output of the system to zero. In the next section it will be shown that one such controller is

$$u = N(|k(t)|)k(t)y(t)$$

$$\frac{d}{dt}k(t) = y(t)^2$$

where $N(x): R^+ \to R$ (called a Nussbaum gain) is any locally Lipschitz function satisfying

$$\sup_{x>0}\left\{\frac{1}{x}\int_0^x zN(z)\,dz\right\} = \infty$$

$$\inf_{x>0}\left\{\frac{1}{x}\int_0^x zN(z)\,dz\right\} = -\infty. \quad (2)$$

The proof that this controller stabilizes the system will be given in three steps. The first is the decomposition of the system into a feedback configuration with a first-order system in the forward loop and a stable system in the feedback loop. The second step is to generalize results on L_p stability of linear time-invariant systems to include systems governed by functional differential equations. (The theorem that will be given here is more general than needed for the stability analysis, since it includes the case of an infinite number of delays.) The final step is the stability analysis where the properties of the Nussbaum gain will be used to show that the output converges to zero as t tends to infinity.

III. Stability Analysis

The transfer function (1) can be written as $g = b_0 q/p$ where p, q are polynomials in s, $e^{-h_i s}$. Treating the exponentials in p and q as parameters and using the division algorithm, p can be written as $p = q(s + \sum_{i=0}^{m}(a_1^i - b_1^i)e^{-sh_i}) + w'$ where w' is a polynomial of degree at most $n - 2$.[1] Therefore, the transfer function may be rewritten

$$g = \frac{b_0}{r}\left[\frac{1}{1 + \frac{b_0}{r}\frac{w}{q}}\right]$$

where $w = w'/b_0$, degree $(w) <$ degree (q), and $r = s + \sum_{i=0}^{m}(a_1^i - b_1^i)e^{-sh_i}$. Thus, the system can be represented in the feedback form of Fig. 1. This configuration permits the system to be described by the

[1] The degree is defined to be the highest power of s, treating the exponentials as parameters.

0018-9286/86/0600-0577$01.00