SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

A SURVEY OF SOME RESULTS IN STOCHASTIC ADAPTIVE CONTROL*

P. R. KUMAR†

Abstract. Some results in discrete-time stochastic adaptive control are surveyed. The survey divides itself into two parts—Bayesian and non-Bayesian adaptive control. In the former area, the problems of converting an incompletely observed system into a completely observed one, multi-armed bandit processes, Bayesian adaptive control of Markov chains and Bayesian adaptive control of linear systems are exposed and surveyed. In the latter area, non-Bayesian adaptive control of Markov chains and the self-tuning regulator are dealt with. Proofs are given, where appropriate, to illustrate the methods involved.

Key words. adaptive control, stochastic adaptive control, Bayesian adaptive control, bandit problems, Bayesian control of Markov chains, non-Bayesian adaptive control, adaptive control of linear systems, self-tuning regulators, self-optimizing systems

CONTENTS

1. Introduction.

We shall be concerned with the control of discrete time stochastic systems. The distinguishing feature of the class of problems we address is that the system under control is unknown. In spite of this, we desire to design control laws which will result in adequate behaviour of the system. The adequacy of the system behaviour will, in turn, be mainly measured by a given cost criterion.

We shall *not* deal with the problem of adaptive control of deterministic systems. Rather, random or noisy behaviour will be an essential feature of the systems we study. We shall also not deal with *continuous* time stochastic systems.

* Received by the editors January 16, 1984, and in revised form June 18, 1984. This is a special expository paper written at the invitation of the editors. This research was supported in part by the National Science Foundation under grant ECS-8304435, and in part by the Army Research Office under contract DAAG29-84-K0005.

† Department of Electrical Engineering and Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801.

The survey divides itself naturally into two parts—Bayesian adaptive control problems (BACP's) and non-Bayesian adaptive control problems (NACP's). We shall suppose, throughout, that the system behaviour depends on a parameter θ, and it is the fact that the *value* of θ is unknown to us which makes the *system unknown.* Now we can distinguish the essential difference between the Bayesian and non-Bayesian formulations. In the former, we are given a probability distribution $q_0(d\theta)$ for the value of the unknown parameter θ. In the latter, we are *not* given any such initial prior distribution. Rather, we are only given a *set* Θ and a *guarantee* that the unknown parameter θ is some element of Θ.

The *Bayesian N*-armed bandit problem and the "dual" control problem are examples of BACP's. The self-tuning regulator and the non-Bayesian adaptive control problem for Markov chains are examples of NACP's. All these topics and others are surveyed in what follows. In §§2–5, we survey BACP's and in §§6–8, we survey NACP's. Each section is addressed to one type of problem.

Some comments on the goal and style of this paper are in order. It has been a primary goal in writing this paper to produce an exposé of the field which will cover, in an understandable way, some of the problems, ideas and mathematical techniques of this field. To achieve this, *proofs* of several results are provided throughout the paper. (However a theorem or proof attributed to an author is not necessarily an exact replica of the original; some modifications have sometimes been made in the interests of brevity, clarity, etc.). No attempt has been made to provide an exhaustive list of *all* papers in the field. Such an approach, it was felt, would only tend to make the narrative very disjointed. The emphasis, instead, is on the coverage of *ideas.* Apologies are thus owed to several authors for such omissions. Lastly, we have made no real attempt at attribution of results to authors.

2. Formulation and reduction of the Bayesian adaptive control problem. The Bayesian approach to adaptive control is this. There is a stochastic dynamic system which depends on a parameter θ. We are given an initial probability distribution (a *prior* distribution) $q_0(d\theta)$ for the unknown parameter. At each time instant $t = 1, 2, 3, \cdots$ we obtain a noisy (or, as a special case, perfect) observation y_t of the state x_t of the system. Our goal is to minimize some given cost criterion, such as $E\sum_0^\infty \beta^t c(x_t, u_t)$. Here u_t is the control input that we apply to the system on the basis of the observations $(u_0, y_1, u_1, y_2, \cdots, y_{t-1}, u_{t-1}, y_t)$ made on the system. β with $0 < \beta \leqq 1$ is the discount factor. The heart of the problem is that the expectation in the cost criterion $E\sum_0^\infty \beta^t c(x_t, u_t)$ is taken not only with respect to the random behaviour of the stochastic system, but *also* with respect to the random choice of θ according to $q_0(d\theta)$.

The standard approach to solving this problem is to transform the BACP into an equivalent *dynamic programming* problem and then to bring to bear the well-developed theory of dynamic programming. The "*state*" of this new dynamic programming problem, which we shall refer to as the *hyperstate*, will be the conditional probability distribution of the *old* state *and* the parameter value given the observations made.

However, in achieving this transformation, some delicate measurability questions must be resolved in order to ensure that one obtains a mathematically well formulated dynamic programming problem. It is one of the accomplishments of the past two decades that this problem of converting a partial or imperfect observations stochastic control problem (for this is what a BACP is) into an equivalent dynamic programming problem has been more or less satisfactorily resolved.

The problem to be examined below is, in some sense, purely technical, and it is somewhat unfortunate that in order to preserve logical continuity this purely technical problem is the first issue examined in detail in this survey. The reader without an appetite for technical issues may wish to gloss over this section and proceed to § 3 and the rest of this paper where, in contrast to this section, purely *structural* issues are addressed.

The development we follow is due to Bertsekas and Shreve [1] and consists of the following:

i) X, Y, U, Θ are Borel spaces (i.e. homeomorphic to a Borel subset of some complete separable metric space) which are, respectively, the state space, observation space, control set and parameter set.

ii) $q_0(dx_0, d\theta)$ is a given probability distribution for the initial state x_0 and the parameter θ.

iii) $p(dx_{t+1}|x_t, u_t, \theta)$ is a Borel measurable stochastic kernel (i.e. $p(B|\cdot)$ is Borel measurable for every Borel set $B \subseteq X$) which specifies the probability distribution of the new state x_{t+1} given the previous state x_t, applied control u_t and the parameter value θ.

iv) $r(dy_t, x_t, u_{t-1}, \theta)$ is a Borel measurable stochastic kernel which specifies the probability distribution of the observation y_t given the state x_t and control u_{t-1}.

v) $c(x_t, u_t)$ is a lower semi-analytic function which is the one-step cost function. (This means $\{(x, u): c(x, u) < a\}$ is an analytic set for every a. A subset of A is analytic if it is the projection on A of a Borel subset of $A \times B$ where B is some uncountable Borel space.) $\beta \in (0, 1]$ is a discount factor. If $\beta = 1$, we assume that either $c \geqq 0$ always or $c \leqq 0$ always. c will always be assumed to be bounded.

vi) A policy π is a sequence $\pi = (\pi_0, \pi_1, \cdots)$ where each $\pi_t(du_t|q_0, u_0, y_1, u_1, \cdots, y_t)$ is a universally measurable stochastic kernel (i.e. $\pi(B|\cdot)$ is universally measurable for every Borel $B \subseteq U$. A function f is universally measurable if the inverse image of every Borel set is measurable with respect to the completion of every Borel measure). Let Π be the set of all such policies.

vii) For every $\pi \in \Pi$ and q_0, one can define the associated cost function $J(\pi, q_0) = E \sum_0^\infty \beta^t c(x_t, u_t)$. Let $J(q_0) = \inf_{\pi \in \Pi} J(\pi, q_0)$ be the *optimal* cost function.

The interpretation of this model is as follows. At time 0, the initial state x_0 and the parameter value θ are distributed according to $q_0(dx_0, d\theta)$. Based on q_0, the controller chooses a $u_0 \in U$ according to the distribution $\pi_0(du_0|q_0)$. Based on (x_0, u_0, θ) the new state x_1 is distributed according to $p(dx_1|x_0, u_0, \theta)$. Based on (x_1, u_0, θ) the controller receives an observation y_1 distributed according to $r(dy_1|x_1, u_0, \theta)$. Based on (q_0, u_0, y_1) the controller chooses u_1 according to $\pi_1(du_1|q_0, u_0, y_1)$ etc.

THEOREM 2.1 (Bertsekas and Shreve).

i) *There exists a Borel measurable stochastic kernel* $q_n(A|q_0, u_0, \cdots, y_n) = E_\pi(1_A(x_n, \theta)|q_0, u_0, \cdots, y_n)$ *for all* π, q_0 *and a.e.* $(u_0, \cdots, y_n)$. *Here* $E_\pi(\cdot)$ *is the conditional expectation under the probability measure induced by the policy* π *and* $1_A(\cdot)$ *is the indicator function of the set* A.

ii) *There exists a Borel measurable stochastic kernel* $\hat{p}(dq_{k+1}|q_k, u_k)$ *such that* $\hat{p}(Q|q_k, u_k) = E_\pi(1_Q(q_{k+1}(\cdot|q_0, \cdots, y_{k+1}))|q_0, q_k(\cdot|q_0, \cdots, y_k), u_k)$ *for every* π, q_0 *and a.e.* $(u_0, \cdots, y_k)$.

iii) *There exists a lower semi-analytic function* $\hat{c}(q_k, u_k)$ *such that* $\hat{c}(q_k, u_k) = E_\pi(c(x_k, u_k)|q_0, q_k(\cdot|q_0, \cdots, y_k), u_k)$ *for every* π, q_0.

iv) *Let* $\hat{\Pi}$ *be the set of all policies* $\hat{\pi}_k = (\hat{\pi}_0, \hat{\pi}_1, \cdots)$ *such that* $\hat{\pi}_k(du_k|q_k)$ *is a universally measurable kernel. For every policy* $\pi \in \hat{\Pi}$, *there is a policy* $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$

such that $\pi_k(du_k|q_0, u_0, \cdots, y_k) = \hat{\pi}_k(du_k|q_k(\cdot|q_0, u_0, \cdots, y_k))$. *Thus* $\hat{\Pi}$ *can be identified with a subset of* Π.

v) *If* $\hat{\pi} \in \hat{\Pi}$ *is nonrandomized, then the corresponding element* $\pi \in \Pi$ *with which it is identified is also nonrandomized.* (*A policy* $\pi = (\pi_0, \pi_1, \cdots)$ *is nonrandomized if each* π_k *is a degenerate probability distribution concentrated on just one point.*)

vi) *Let* $\hat{J}(\hat{\pi}, q_0) = E_{\hat{\pi}}(\sum_0^\infty \hat{c}(q_k, u_k)\beta^k | q_0 = q)$ *be the cost function of a dynamic programming problem with transition kernel* $\hat{p}$, *policy set* $\hat{\Pi}$, *and cost function* $\hat{c}$. *Then* $J(\hat{\pi}, q) = \hat{J}(\hat{\pi}, q)$ *for every* $\hat{\pi} \in \hat{\Pi}$ *and* q. (*In the left-hand side, by* $\hat{\pi}$ *we mean the element of* Π *with which it is identified.*)

vii) *For every* q *and* $\pi \in \Pi$, *there exists a* $\hat{\pi} \in \hat{\Pi}$ *such that* $J(\pi, q) = \hat{J}(\hat{\pi}, q)$.

viii) *Let* $\hat{J}(q) = \inf_{\hat{\pi} \in \hat{\Pi}} \hat{J}(\hat{\pi}, q)$. *Then* $\hat{J}(q) = J(q)$ *for every* q.

ix) *If* $\hat{\pi}^*$ *satisfies* $\hat{J}(\hat{\pi}^*, q) = \hat{J}(q)$ (*or* $\leqq \hat{J}(q) + \varepsilon$) *for every* q, *then* $J(\hat{\pi}^*, q) = J(q)$ (*or* $\leqq J(q) + \varepsilon$) *for every* q.

x) *For every* $\varepsilon > 0$, *there exists a* $\hat{\pi} \in \hat{\Pi}$ *which is nonrandomized, such that* $\hat{J}(\hat{\pi}, q) \leqq \hat{J}(q) + \varepsilon$ *for every* q.

Proof. See [1, Lemma 10.2 and Propositions 10.2, 10.3, 10.5].

The interpretation is as follows. $q_k(dx_k, d\theta|q_0, u_0, \cdots, y_k)$ is the conditional probability distribution of (x_k, θ) given the initial probability distribution $q_0(dx_0, d\theta)$ and the observation history $(u_0, y_1, \cdots, y_k)$. This will be the *hyperstate* of the new dynamic programming problem. $\hat{p}(dq_{k+1}|q_k, u_k)$ specifies the transition probability function for this new problem while $\hat{c}(q_k, u_k)$ is the new one-step cost function. Together, (i)-(iii) show that q_k, $\hat{p}$, $\hat{c}$ satisfy the assumptions to provide us a well-formulated dynamic programming problem which is of the type studied in [1]. The main point is that the hyperstate is completely observed.

The relationship between this new dynamic programming problem and the original BACP is provided through (iv)-(x). (iv) shows that any policy for the new dynamic programming problem can also be implemented on the BACP. This is clear in the sense that policies in $\hat{\Pi}$ are *only* allowed to depend on $(q_0, q_1, \cdots, u_0, u_1, \cdots)$ while policies in Π are allowed to depend on $(q_0, y_1, y_2, y_3, \cdots, u_0, u_1, \cdots)$ and each q_n is itself calculated on the basis of $(q_0, y_1, \cdots, y_n, u_0, \cdots, u_{n-1})$. Thus $\hat{\Pi}$ can be identified as a subset of Π. (vi) shows that the cost of using $\hat{\pi}$ in the new dynamic programming problem is the same as using it in the BACP. (vii) is a deep result which shows that for the BACP, every policy $\pi \in \Pi$ can be replaced by a policy $\hat{\pi} \in \hat{\Pi}$ which has the same cost. The advantage of this is that one may then restrict attention to policies $\hat{\pi} \in \hat{\Pi}$ which depend only on the hyperstate, thus rendering the hyperstate a "sufficient statistic" in some sense. (viii) and (ix) complete the process of identifying the BACP with the new dynamic programming problem. (viii) shows that both have the same optimal cost functions. (ix) shows that a policy $\hat{\pi} \in \hat{\Pi}$ optimal (or ε-optimal) for the new dynamic programming problem is also optimal (or ε-optimal) for the BACP. (x) is a consequence of our allowing universally measurable policies and shows that there exists an ε-optimal nonrandomized policy for the new dynamic programming problem (and therefore also for the BACP). A development similar to the above can also be given for the finite horizon case, see [1].

For other treatments of the conversion of an imperfectly observed problem into a completely observed dynamic programming problem, the reader is referred to Bellman [2], Dynkin [3], Aoki [4], Åström [5], Shiryaev [6], Striebel [7], [8], Hinderer [9], Sawaragi and Yoshikawa [10], Rhenius [11], Martin [12], Rieder [13] and van Hee [14]. [2]-[5] are early references featuring examples. [7] examines the concept(s) of a "sufficient statistic". [10] deals with countable state spaces and shows versions of (vi) and (vii) above. [11] considers the same issues for general Borel spaces. [12]-[14] also

achieve the conversion to a completely observed dynamic programming problem, making specific reference to the BACP. We have chosen here to follow the approach of [1] because its allowance of universally measurable policies gives the useful and reassuring property (x) above.

3. Bandit processes. As we have seen in § 2, a BACP can, like other imperfectly, observed problems, be replaced by an equivalent dynamic programming problem. However, since a BACP is a *special* type of an imperfect observations problem, the question naturally arises as to whether and to what extent we can take advantage of the special structure.

There is one class of problems, the multi-armed bandit problems, which is very special even *within* the class of BACP's and for this class of problems we can exploit this highly special structure to provide a rather deep theory.

Suppose that there are N (slot) machines. For machine i, $1 \leqq i \leqq N$, θ_i is the probability that, if it is played, it yields a reward of one unit, while $(1-\theta_i)$ is the probability that it yields no reward. The parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)$ is unknown. At every time $t = 0, 1, 2, \cdots$ *one* of the machines *has* to be played, and suppose that the reward accrued at time t is c_t. The goal is to choose which machine to play at each time $t = 0, 1, 2, \cdots$ so that $E \sum_0^\infty \beta^t c_t$ is maximized. $\beta \in (0, 1)$ is a discount factor.

We are not told of the value of $\boldsymbol{\theta}$, but instead, we are given a prior probability distribution of $q_0(d\boldsymbol{\theta}) = q_0^1(d\theta_1) q_0^2(d\theta_2) \cdots q_0^N(d\theta_N)$ for the value of $\boldsymbol{\theta}$. This immediately renders this a BACP.

As in the previous section, we therefore define the *hyperstate* $\mathbf{q} = (q^1, q^2, \cdots, q^N)$ where each $q^i(d\theta_i)$ is the conditional distribution of the value of θ_i. This gives a dynamic programming problem characterized by the following dynamic programming equation:

$$J(\mathbf{q}) = \max_{i \leqq i \leqq N} \left\{ \int_0^1 \theta_i q^i(d\theta_i) + \beta \left[\int_0^1 \theta_i q^i(d\theta_i) J(q^1, q^2, \cdots, q^{i-1}, S(q^i), q^{i+1}, \cdots, q^N) \right.\right.$$
$$\left.\left. + \int_0^1 (1-\theta_i) q^i(d\theta_i) J(q^1, q^2, \cdots, q^{i-1}, F(q^i), q^{i+1}, \cdots, q^N) \right] \right\},$$

where

$$S(q^i)(d\theta_i) = \theta_i q^i(d\theta_i) \left[\int_0^1 \eta q^i(d\eta) \right]^{-1}$$

and

$$F(q^i)(d\theta_i) := (1-\theta_i) q^i(d\theta_i) \left[\int_0^1 (1-\eta) q^i(d\eta) \right]^{-1}$$

The interpretation of this equation is straightforward. If machine i is played, then the probability that it yields a reward of one unit is $\int_0^1 \theta_i q^i(d\theta_i)$, and such an occurrence will cause us to revise the *posterior* distribution of the probability of success on machine i to $S(q^i)$ by Bayes' rule. Thus the state $\mathbf{q} = (q^1, \cdots, q^i, \cdots, q^N)$ changes to $(q^1, \cdots, S(q^i), \cdots, q^N)$. A similar analysis holds when a failure results from our playing machine i. $J(\mathbf{q})$, the optimal reward for the problem starting in the hyperstate $\mathbf{q}$, thus satisfies the given dynamic programming equation.

However, this dynamic programming equation, as written, is a rather formidable functional equation to solve. Through the work of Gittins and Jones [15], there is now a deep theory for this problem.

THEOREM 3.1.

i) (Gittins and Jones). *There is a real valued index function* $\gamma(\cdot)$ *with the property that if at any time t, the hyperstate is* $\mathbf{q}(t) = (q^1(t), \cdots, q^N(t))$, *then it is optimal to play any machine i (there may be more than one) for which* $\gamma(q^i(t))$ *is largest.*

ii) (Gittins and Glazebrook). *Suppose the state of a machine (it does not matter which one) is q. Let* $\{d_0, d_1, d_2, \cdots\}$ *be the stochastic process representing the successive random rewards obtainable by continuously playing this machine. Let* $\mathcal{F}_t := \sigma(d_0, d_1, \cdots, d_t)$ *be the* σ*-algebra generated by the reward sequence up to time t. Then*

$$\gamma(q) = \max_{\tau \geqq 1} \frac{E \sum_0^{\tau-1} \beta^t d_t}{E \sum_0^{\tau-1} \beta^t}$$

will suffice in (i), *where the maximum is taken over all stopping times* τ *of* $\{\mathcal{F}_t\}$.

iii) (Nash). *In* (ii), *let* $q(t)$ *be the hyperstate of the machine at time t (after t successive plays of the machine). Then a maximizing stopping time* τ *in* (ii) *is given by*

$$\tau = \inf\{t \geqq 1 \colon \gamma(q(t)) \leqq \gamma(q)\}.$$

iv) *Consider a problem for which there is* only *one machine which initially, as in* (ii), *has a hyperstate q. At any given time, one may either continue to play the machine, or collect M units of reward and quit forever. Let* $V(q, M)$ *be the optimal expected reward for this optimal stopping problem. Then*

$$V(q, M) = \max\left\{M; \int_0^1 \theta q(d\theta) + \beta\left[\int_0^1 \theta q(d\theta) V(S(q), M) + \int_0^1 (1-\theta) q(d\theta) V(F(q), M)\right]\right\}.$$

v) (Gittins and Jones). $\gamma(\cdot)$, *the index function of* (i) *and* (ii) *is given by*

$$\gamma(q) = \inf\{M(1-\beta) \colon V(q, M) = M\} = \sup\{M(1-\beta) \colon V(q, M) > M\}.$$

vi) (Whittle). *The optimal reward function J is related to V by*

$$J(\mathbf{q}) = (1-\beta)^{-1} - \int_0^{(1-\beta)^{-1}} \prod_{j=1}^{N} \frac{\partial}{\partial M} V(q^j, M)\, dM.$$

Proof. It is clear that $\gamma(q)$ in (ii) is well defined. It can be shown that this optimal stopping problem of (ii) actually has a *maximizing* stopping time, which demonstration we omit. Now we will show that the τ defined in (iii) attains the maximum in (ii). Let σ be *any* optimal stopping time (we have already assumed that there exists at least one such). Then consider the new stopping time $\mu := \tau \wedge \sigma$. Elementary calculations show that

$$\begin{aligned}
&\left(E \sum_0^{\mu-1} \beta^t d_t\right)\left(E \sum_0^{\mu-1} \beta^t\right)^{-1} - E\left(\sum_0^{\sigma-1} \beta^t d_t\right)\left(E \sum_0^{\sigma-1} \beta^t\right)^{-1} \\
&= \left\{E\left[-1(\tau < \sigma) \sum_\tau^{\sigma-1} d_t \beta^t\right] E \sum_0^{\sigma-1} \beta^t - E\left[-1(\tau < \sigma) \sum_\tau^{\sigma-1} \beta^t\right] E \sum_0^{\sigma-1} \beta^t d_t\right\} \\
&\quad \cdot \left\{E \sum_0^{\mu-1} \beta^t E \sum_0^{\sigma-1} \beta^t\right\}^{-1} \\
&= \left\{E\left[-1(\tau < \sigma) \sum_\tau^{\sigma-1} d_t \beta^t\right] - \gamma(q) E\left[-1(\tau < \sigma) \sum_\tau^{\sigma-1} \beta^t\right]\right\}\left\{E \sum_0^{\mu-1} \beta^t\right\}^{-1}
\end{aligned}$$

$$
\begin{aligned}
&= \left\{ \gamma(q) E\left[1(\tau<\sigma) \sum_{\tau}^{\sigma-1} \beta^t \right] - EE\left[1(\tau<\sigma) \sum_{\tau}^{\sigma-1} d_t \beta^t | \mathscr{F}_\tau \right] \right\} \left\{ E \sum_{0}^{\mu-1} \beta^t \right\}^{-1} \\
&\geqq \left\{ \gamma(q) E\left[1(\tau<\sigma) \sum_{\tau}^{\sigma-1} \beta^t \right] - E\left[1(\tau<\sigma)\gamma(q(\tau)) \sum_{\tau}^{\sigma-1} \beta^t \right] \right\} \left\{ E \sum_{0}^{\mu-1} \beta^t \right\}^{-1} \\
&\geqq \left\{ \gamma(q) E\left[1(\tau<\sigma) \sum_{\tau}^{\sigma-1} \beta^t \right] - E\left[1(\tau<\sigma)\gamma(q) \sum_{\tau}^{\sigma-1} \beta^t \right] \right\} \left\{ E \sum_{0}^{\mu-1} \beta^t \right\}^{-1} \\
&= 0.
\end{aligned}
$$

This shows that if σ is optimal, then $\sigma \wedge \tau = \mu$ is also optimal. We now show that $\mu = \tau$ a.s. Suppose not, then $P(\mu < \tau) > 0$, i.e. $P(\gamma(q(u)) > \gamma(q)) > 0$. For some $\varepsilon > 0$, therefore $P(\gamma(q(\mu)) > \gamma(q) + 2\varepsilon) > 0$. Define $\tilde{\Omega} := \{\omega\colon \gamma(q(\mu)) > \gamma(q) + 2\varepsilon\}$ and a new stopping time ξ by $\xi := \mu$ on $\tilde{\Omega}^c$, and $\xi := \mu + \alpha(q(\mu))$ on $\tilde{\Omega}$, where $\alpha(q(\mu))$ is ε-optimal starting from the state $q(\mu)$, i.e.

$$
E\left[\sum_{\mu}^{\mu+\alpha(q(\mu))-1} \beta^t d_t | \mathscr{F}_\mu \right] \left\{ E\left[\sum_{\mu}^{\mu+\alpha(q(\mu))-1} \beta^t | \mathscr{F}_\mu \right] \right\}^{-1} \geqq \gamma(q(\mu)) - \varepsilon \geqq \gamma(q) + \varepsilon \quad \text{on } \tilde{\Omega}.
$$

Another calculation shows that

$$
\begin{aligned}
\gamma(q) &\geqq E \sum_{0}^{\xi-1} \beta^t d_t \left\{ E \sum_{0}^{\xi-1} \beta^t \right\}^{-1} \\
&= \left\{ E \sum_{0}^{\mu-1} \beta^t d_t + E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t d_t \right] \right\} \left\{ E \sum_{0}^{\mu-1} \beta^t + E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t \right] \right\}^{-1} \\
&\geqq \left\{ \gamma(q) E \sum_{0}^{\mu-1} \beta^t + E\left[1(\xi>\mu)(\gamma(q)+\varepsilon) \sum_{\mu}^{\xi-1} \beta^t \right] \right\} \\
&\quad \cdot \left\{ E \sum_{0}^{\mu-1} \beta^t + E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t \right] \right\}^{-1} \\
&= \gamma(q) + \varepsilon E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t \right] \left\{ E \sum_{0}^{\mu-1} \beta^t + E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t \right] \right\}^{-1} \\
&> \gamma(q),
\end{aligned}
$$

which is impossible. This shows that $\mu = \tau$ a.s., showing that τ is optimal. This proves (iii).

Now we follow essentially the interchange argument of Glazebrook [19], which in turn is essentially the argument in [15], in showing (i) and (ii). Let π be a particular policy satisfying (i) with γ defined as in (ii) and which breaks ties between competing maximizers by choosing the lexicographically smallest machine. Suppose that π chooses machine i at time 0. Then $\gamma(q^i(0)) = \max_j \gamma(q^j(0))$.

Consider $k \neq i$, and let $\hat{\pi}$ be the (nonstationary) policy which chooses machine k at time $t = 0$, and thereafter proceeds according to π. Note that under $\hat{\pi}$, machine k will be chosen at times $t = 0, \cdots, \tau_k - 1$ where $\tau_k := \min\{t \geqq 1\colon \gamma(q^k(\tau_k)) \leqq \gamma(q^i(0))\}$ if $k > i$ or $\tau_k := \min\{t \geqq 1\colon \gamma(q^k(\tau_k)) < \gamma(q^i(0))\}$ if $k < i$. Thereafter, under $\hat{\pi}$, machine i will be chosen at times $t = \tau_k, \cdots, \tau_k + \tau_i - 1$ where $\tau_k + \tau_i = \min\{t > \tau_k\colon \gamma(q^i(t)) \leqq \gamma(q^i(\tau_k)) = \gamma(q^i(0))\}$.

On the other hand, consider the policy $\hat{\hat{\pi}}$ which chooses machine i at times $t = 0, \cdots, \tau_i - 1$ and thereafter chooses machine k at times $t = \tau_i, \cdots, \tau_i + \tau_k - 1$. After

this, $\hat{\hat{\pi}}$ coincides with $\hat{\pi}$. Hence

$$
\begin{aligned}
&E_{\hat{\hat{\pi}}} \sum_0^\infty \beta^t d_t - E_{\hat{\pi}} \sum_0^\infty \beta^t d_t \\
&= E_{\hat{\hat{\pi}}}\left[\sum_0^{\tau_i - 1} \beta^t d_t + \sum_{\tau_i}^{\tau_i + \tau_k - 1} \beta^t d_t\right] - E_{\hat{\pi}}\left[\sum_0^{\tau_k - 1} \beta^t d_t + \sum_{\tau_k}^{\tau_i - 1} \beta^t d_t\right] \\
&= E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t d_t + E_{\hat{\hat{\pi}}} \beta^{\tau_i} E_{\hat{\pi}} \sum_0^{\tau_k - 1} \beta^t d_t - E_{\hat{\pi}} \sum_0^{\tau_k - 1} \beta^t d_t - E_{\hat{\pi}} \beta^{\tau_k} E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t d_t \\
&= E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t d_t [1 - E_{\hat{\pi}} \beta^{\tau_k}] - E_{\hat{\pi}} \sum_0^{\tau_k - 1} \beta^t d_t [1 - E_{\hat{\hat{\pi}}} \beta^{\tau_i}] \\
&\geqq \gamma(q^i(0)) E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t E_{\hat{\pi}} \sum_0^{\tau_k} \beta^t (1-\beta)^{-1} - \gamma(q^k(0)) E_{\hat{\pi}} \sum_0^{\tau_k - 1} \beta^t E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t (1-\beta)^{-1} \\
&\geqq 0.
\end{aligned}
$$

Thus $\hat{\hat{\pi}}$ is an improvement of $\hat{\pi}$.

Our main goal is to show that π is an improvement of $\hat{\pi}$. Now, at time τ_i, $\hat{\hat{\pi}}$ *might not* use the machine that π does. In this case, by shifting the time origin to τ_i and by repeating the above argument, one can obtain an improvement $\hat{\hat{\hat{\pi}}}$ of $\hat{\hat{\pi}}$. This policy $\hat{\hat{\hat{\pi}}}$ will coincide with π over a strictly longer initial time segment than $\hat{\hat{\pi}}$ does. Continuing in this way, we can obtain policies which coincide with π over arbitrarily large initial time segments and which are improvements of $\hat{\pi}$. Since this is a discounted cost problem with discount factor $\beta < 1$, it follows that π itself is an improvement of $\hat{\pi}$.

Since $\hat{\pi} = (\hat{\pi}_0, \tilde{\pi}, \tilde{\pi}, \tilde{\pi}, \cdots)$ where $\pi = (\tilde{\pi}, \tilde{\pi}, \tilde{\pi}, \cdots)$, it follows by standard results on the discounted cost problem that π is optimal.

π was special in that it used the natural ordering of $\{1, 2, \cdots, N\}$ to break ties. Clearly *any* ordering of $\{1, 2, \cdots, N\}$ could have been used. Standard results again show that at any given time t, one can use *any* machine i with the largest value $\gamma(q^i(t))$ and still achieve optimality. This proves (i) and (ii).

For (v), consider the problem equivalent to (iv) where the option of retiring and collecting M is replaced by a machine which has a *known* probability $M(1-\beta)$ of success (for $0 \leqq M \leqq 1/(1-\beta)$). By (ii) the index of the *known* machine is $M(1-\beta)$. By (i) the index of the unknown machine is therefore that value of $M(1-\beta)$ when one is exactly indifferent to playing the unknown machine once or the known machine once.

The result (vi) is from Whittle [18] and for a proof we refer the reader to Whittle [20] or Ross [21]. (In the attribution of (iii) we have followed [16]). □

The above results possess some very special features which we now discuss. First, (i) shows that to each machine one may assign a (desirability) index $\gamma(q)$ which *just* depends on the *state* of the machine and *nothing else*. The optimal rule then, according to (i), has the intuitively appealing interpretation that at any given time one should merely *compare* the indices of the available machines, and then play the machine with the largest index. Thus, we have the very useful property that the problem of dealing with N machines simultaneously, with attendant state $\mathbf{q} = (q^1, \cdots, q^N)$ *simplifies* (or *decouples*) into N *separate* machines each with state q^i, $1 \leqq i \leqq N$.

Second, the index of a machine which from now on we refer to as its *Gittins* index, can be interpreted according to (ii) as the maximum reward per unit time (where both rewards and time are progressively discounted by a factor β) given that one may

stop anytime. Thus the problem of computation of the Gittins index is an optimal stopping problem.

Third, (v) affords yet another interpretation of the Gittins index. Consider the problem of (iv) where there is a machine and a retirement option of M units. Then, the Gittins index is that value of $M(1-\beta)$ at which one is exactly indifferent to retiring or playing the machine once and preserving the retirement option.

Last, (vi) shows that one can actually obtain an expression for the optimal reward function $J(\mathbf{q})$ of the N-armed bandit problem in terms of the optimal reward functions of *single*-armed bandits with retirement options. This is quite remarkable in view of the formidable functional equation defining $J(\cdot)$.

From the characterization of τ in (iii) it follows that for $\sigma < \tau$, $\gamma(q(\sigma)) > \gamma(q(0))$. Thus at any time *prior* to τ, the Gittins index is larger than the Gittins indices of the other machines if it were so at time 0. Hence, one should play the machine with the largest index continuously for a period of at least τ units. Then only need one consider changing machines.

This result can be generalized in several ways. The first is as follows. Consider the situation where there are N Markov processes with states $q^1(t), q^2(t), \cdots, q^N(t)$. At each time t one can choose to let *one* of the processes evolve while freezing *all* the rest. If process i is chosen for evolution, then the next state $q^i(t+1)$ is chosen according to a transition probability function $P_i(dq^i(t+1)|q^i(t))$ and a reward (or a cost) $c^i(q^i(t))$ is obtained. All other processes have $q^j(t+1) = q^j(t)$ for $j \neq i$. The generalization here is that the process $\{q^i(t)\}$ need not represent the "hyperstates" or conditional probability distributions of some other process and the transition probability function $P_i(\cdot|\cdot)$ can be quite general. Further, the reward function $c^i(\cdot)$ can also be general. This general version is the one treated in [15] and it is quite clear that the proof above works without modifications. One can also consider an adaptive version of this problem where the transition probabilities are unknown, see [16].

The next generalization is to allow each of the above Markov processes to be a Markov decision process, i.e., after choosing one process for evolution, one also has to choose a *control* action to apply to this process, see Gittins [15] where each process is now referred to as a "superprocess". Now, however, it is not generally true that the index result generalizes; some restrictions have to be imposed. Whittle [18] shows that if *each* individual superprocess i with a retirement option of M units has an optimal stationary policy (in the sense of Markov decision processes) with the property that it is optimal for *all* values of M, then an index rule is optimal. (Note that in the previously considered processes the control set is a singleton and so clearly this condition is satisfied.) Glazebrook [23] shows that this condition is both weakly and strongly necessary in a certain sense.

Another generalization of the bandit problem is to the case where new (slot) machines arrive in the course of time; this is treated by Whittle [24].

Varaiya, Walrand and Buyukkoc [25] consider a different sort of extension and show that the process involved need not even be Markov. They also exhibit certain problems which are equivalent to bandit problems, but which are nevertheless more useful from the viewpoint of certain applications.

There are many applications of these results in the areas of stochastic scheduling (of jobs by a server), searching (for an object hidden in one of several boxes), planning of research (which option should be investigated next), design of experiments etc. Just as an illustration; if there are many customers and only one server, then the problem of deciding which customer to serve is analogous to deciding which of several machines to play in a bandit problem. We refer the reader to [19] for references.

Let us revert to the N-armed bandit formulation. For each $\beta \in (0, 1)$, the implementation of each β-optimal policy converges a.s. onto *some* machine, i.e., after some time only one machine is exclusively played and the others are ignored (under some technical conditions), see Rothschild [26] and Kelly [27]. Thus every β-optimal policy *experiments* with the machines for only a certain amount of time after which it settles down to playing some machine exclusively, i.e., *stops* experimenting. It is of course of interest to obtain the probability P_β that the β-optimal policy will settle down ultimately to playing the machine with the *largest* probability of success. As $\beta \to 1$ it is shown in [27] that $P_\beta \to 1$. However $P_\beta < 1$ for all β, see Rothschild [26].

Now we consider another aspect. First we define a policy, the "least failures policy", which operates as follows. At each instant of time we play the machine with the least number of previous failures, except that if there is more than one such machine, then we select the one with the greatest number of successes, and again if there is more than one such, then we choose in a uniformly randomized way from among all the contenders. Kelly [27] shows that as $\beta \to 1$, the β-optimal policies evaluated at each state converge (but *not* uniformly over all states) to the least failures policy (we assume all machines start alike).

For β close to 1 each β-optimal policy therefore plays the least failures rule for a certain amount of time (until it leaves a certain set of states over which the β-optimal policy coincides with the least failures rule). Thus we now have a very nice interpretation of the behaviour of the β-optimal policies. There are three phases of behaviour. During the first phase, there is *experimentation* with the various machines (because the least failures rule constantly switches between machines). Then there is an intermediate period and finally, in the last phase, *learning* completely stops and the policy only plays one machine to the exclusion of all others. This explanation of Kelly [27] is as good and rigorous as any we have seen regarding the so-called "dual" effect in Bayesian adaptive control, see § 5.

There is a discontinuity in the behaviour at $\beta = 1$. As $\beta \to 1$, the limiting policy, the least failures rule is *not* optimal with respect to maximizing the average reward per unit time. However for β close to 1, the β-optimal policies do perform well with respect to the average cost criterion in the sense that $P_\beta \approx 1$.

This discussion has shown that we can obtain policies which are optimal with respect to the discounted cost criterion and *close* to optimal with respect to the average cost criterion. The converse can also be done. By considering randomized policies, Glazebrook [28] shows that one can obtain policies which are optimal with respect to the average cost criterion and *close* to optimal with respect to the discounted cost criterion; see also Bather [29].

As we have seen in the above theorem, we can reduce the N-armed bandit problem to N separate one-armed bandit problems of the type featured in (iv) of the theorem, which, of course, leaves us with the task of dealing with these one-armed bandit problems. Gittins [15] shows how one can develop recursive approximation schemes. Glazebrook [20] shows how the index results can be used to obtain bounds on the quality of *any* stationary strategy.

Explicit analytic results on these one-armed bandit problems are available in the case of "improving" or "deteriorating" arms, see [15]. In [31] by using careful bounds for the dynamic programming equation it is shown that if the known arm has a probability $m/(n+m)$ of success where m and n are relatively prime integers, then if the unknown machine has a probability of success which is Beta distributed with integer parameters x and y, then the optimal policy is to play the unknown machine or the known machine, respectively, according as $x/(x+y) \geqq m/(m+n)$ or $x/(x+y) <$

$m/(m+n)$ for all $0<\beta\leqq 1/(n+1)$. Another result in the same vein is described in [30] where it is indicated that by computation it was determined that for all $0<\beta<0.801$ and $m=n=1$, the optimal policy is to play the unknown machine until it has more failures (x) than successes (y).

4. Bayesian control of Markov chains. As we have seen in the previous § 3, the bandit family of problems possesses a highly special structure which can be usefully exploited to provide a deep theory. What results are available in the general BACP where such special structure does not exist? If one considers either the total discounted or total undiscounted cost criteria, deep results do not appear to be available, other than in highly degenerate situations, some examples of which are given in van Hee [14].

Attention has therefore been turned to the problem of obtaining accurate approximations to the solution of the dynamic programming equation. We consider below the case where one obtains perfect (as opposed to noisy or incomplete) observations of the state of the system. In such situations one can take advantage of the special structure with which BACP's are endowed.

Let X, U, Θ, all finite, be the state, control and parameter spaces, and suppose that the transition probabilities are given by the function $p(i, j, u; \theta) =$ Prob $(x_{t+1}=j|x_t=i, u_t=u, \theta)$. The cost function is $E\sum_0^\infty \beta^t c(x_t, u_t, x_{t+1})$ where $0<\beta<1$.

First consider the problem of (nonadaptive) control when the parameter value is *known* to be θ. Let $\Pi := \{\pi|\pi: X\to U\}$ be the set of stationary policies, $J^\pi(i, \theta)$ the expected cost when π is employed and the initial state is i, and $J(i, \theta)$ the optimal cost function.

Turning now to the BACP, let $P(\Theta)$ be the set of probability measures on Θ and let $J(i, q)$ be the optimal cost function where q is the prior distribution. (We have used the same notation as for the nonadaptive problem since one can identify a known parameter θ with a degenerate distribution concentrated on θ.) Define T, the standard dynamic programming operator by

$$(TV)(i, q)=\min_{u\in U}\sum_{\theta\in\Theta}\sum_{j\in X} q(\theta)p(i, j, u; \theta)[c(i, u, j)+\beta V(j, S(i, u, j; q))].$$

Here $S(i, u, j; q)\in P(\Theta)$ is the posterior distribution of θ, if q is the prior distribution and a transition of the state from i to j under u is observed. For a standard reference on discounted dynamic programming, the reader is referred to Blackwell [32].

THEOREM 4.1 (van Hee). i) *Let*

$$L(i, q) := \sum_{\theta\in\Theta} q(\theta)J(i, \theta),$$

$$U(i, q) := \min_{\pi\in\Pi}\sum_{\theta\in\Theta} q(\theta)J^\pi(i, \theta),$$

$$T^n := \textit{nth iterate of the map } T.$$

Then

$$L(i, q)\leqq TL(i, q)\leqq\cdots\leqq T^nL(i, q)\leqq\lim_N T^NL(i, q)=J(i, q)$$

$$=\lim_N T^NU(i, q)\leqq T^nU(i, q)\leqq\cdots\leqq TU(i, q)\leqq U(i, q).$$

ii)

$$\sup_{i,q}|T^nU(i, q)-T^nL(i, q)|\leqq\beta\sup_{i,q}|T^{n-1}U(i, q)-T^{n-1}L(i, q)|.$$

Proof. The lower and upper bounds $L(i, q) \leqq J(i, q) \leqq U(i, q)$ are obvious. It is well known that $TJ = J$. Since T is monotone, it follows that $T^n L \leqq T^n J = J \leqq T^n U$. Now

$$\begin{aligned} TL(i, q) &= \min_{u \in U} \sum_{\theta \in \Theta} \sum_{j \in X} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta L(j, S(i, u, j; q))] \\ &= \min_{u} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta \sum_{\theta' \in \Theta} S(i, u, j; q)(\theta') J(j, \theta'). \end{aligned}$$

By Bayes' rule we can obtain

$$\sum_{\theta} p(i, j, u; \theta) q(\theta) S(i, j, u; q)(\theta') = q(\theta') p(i, j, u; \theta').$$

So

$$\begin{aligned} TL(i, q) &= \min_{u} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta J(j, \theta)] \\ &\geqq \sum_{\theta} q(\theta) \left\{ \min_{u} \sum_{j} p(i, j, u; \theta)[c(i, u, j) + \beta J(j, \theta)] \right\} \\ &= \sum_{\theta} q(\theta) J(i, \theta) = L(i, q). \end{aligned}$$

By the monotonicity of T, it follows that $L \leqq TL \leqq T^2 L \leqq \cdots \leqq T^n L$. Also

$$\begin{aligned} TU(i, q) &= \min_{u \in U} \sum_{\theta \in \Theta} \sum_{j \in X} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta U(j, S(i, u, j; q))] \\ &= \min_{u} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta) \left[c(i, u, j) + \beta \min_{\pi \in \Pi} \sum_{\theta' \in \Theta} S(i, u, j; q)(\theta') J^{\pi}(j, \theta') \right] \\ &\leqq \min_{u} \min_{\pi \in \Pi} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta) \left[c(i, u, j) + \beta \sum_{\theta'} S(i, u, j; q)(\theta') J^{\pi}(j, \theta') \right] \\ &= \min_{u} \min_{\pi \in \Pi} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta J^{\pi}(j, \theta)] \\ &\leqq \min_{\pi \in \Pi} \sum_{\theta} \sum_{j} q(\theta) p(i, j, \pi(i); \theta)[c(i, \pi(i), j) + \beta J^{\pi}(j, \theta)] \\ &= \min_{\pi \in \Pi} \sum_{\theta} q(\theta) J^{\pi}(i, \theta) = U(i, q). \end{aligned}$$

Again by monotonicity and contractivity of T, the result follows. □

The usefulness of the above theorem lies in the fact that $T^n L$ and $T^n U$ are lower and upper bounds for J for every n, which moreover converge monotonically to J as $n \to \infty$. How does one use these results? First, $J^{\pi}(i, \theta)$ and $J(i, \theta)$ are obtainable by standard algorithms such as the policy iteration algorithm, see Howard [33]. Thus $L(i, q)$ and $U(i, q)$ are obtainable by standard methods. The determination of $T^n L(i, q)$ (or similarly of $T^n U(i, q)$) for a fixed (i, q) clearly requires the values $T^{n-1} L(j, S(i, u, j; q))$ for every (i, u, j), of which there are finitely many. To see exactly what is involved in the computation of $T^n L(i, q)$, define $R(q) := \cup_{i,u,j} \{S(i, u, j; q\}$ and $R^j(q) = R(R^{j-1}(q))$ (= image of $R^{j-1}(q)$ under R). Thus $R^j(q)$ is the set of posterior distributions which could possibly result after j stages. It is clear that the data needed for the computation of $T^n L(i, q)$ for fixed (i, q) are $\cup_{j=1}^{n-1} \{T^j L(\cdot, \bar{q}): \bar{q} \in R^{n-j}(q)\}$. In theory, therefore, one can calculate $T^n L(i, q)$ after a finite set of computations. Since, by (ii) of the theorem, or by any other way, one can choose n so large that $T^n U(i, q) - T^n L(i, q) \leqq \varepsilon$, it follows that one can approximate $J(i, q)$ to within an error of ε. However, one should note that the computation of $T^{n+1} L(i, q)$ requires the data

$\cup_{j=1}^{n} \{T^j L(\cdot, \bar{q}): \bar{q} \in R^{n+1-j}(q)\}$. Since $R^{n-j}(q)$ and $R^{n+1-j}(q)$ are different sets, and quite possibly disjoint sets, it follows that one cannot proceed from $T^n L(i, q)$ to $T^{n+1} L(i, q)$ in a recursive way.

Thus, the above process of approximation can be quite cumbersome. Some special situations are identified in [14] where one may use simpler procedures. [14] also considers a discretization approach when Θ is not finite.

Martin [12] and Satia and Lave [34] have provided other upper and lower bounds. The bounds in [34] are obtained through an intermediate process (involving a worst case choice of θ and a best case choice of θ, i.e., by solving a min-max and a max-max situation) and are generally poorer than the bounds L and U given above.

A branch and bound algorithm. An interesting "branch and bound" algorithm to obtain an ε-optimal control for the state (i, q) is given by Satia and Lave [34]. This uses, in an essential way, upper and lower prior bounds for the optimal cost function, and we may suppose here that these bounds are U and L. Basically, one examines the possible branches (trajectories) leading out of the state (i, q). At any node, say state $(j, \bar{q})$, one has, at each stage, bounds for the costs of the states resulting from the application of, say, u. One can clearly eliminate from consideration those u's for which the lower bound on the cost-to-go is *larger* than the upper bound on the cost-to-go available through some other control. Thus, at each stage of the algorithm some decisions at nodes are eliminated from consideration and, in addition, the upper and lower bounds at all nodes are refined. Then the decision tree is extended by considering one more time unit of horizon. The prior bounds for these newly introduced nodes are U and L, and the whole process of eliminating decisions and refining the bounds is carried out all over again.

The algorithm is guaranteed to yield an ε-optimal decision in a finite number of iterations. Systems with four states and two decisions are described as being of moderate size, and for such systems the algorithm is regarded as being efficient [34]. However, convergence is reported to be slow when β is close to 1.

In contrast with the discounted cost problem considered above, optimal policies for the BACP are obtainable when the cost criterion is of the average cost type: $E \lim 1/N \sum_1^N c(x_t, u_t)$. This sort of a cost criterion is however more properly examined within a non-Bayesian context, as we see in §§ 6 and 7.

5. Bayesian adaptive control of linear systems with quadratic costs. An example will clarify the situation vis-à-vis BACP's in this category. Consider the Auto Regressive Moving Average System with EXogeneous Inputs (ARMAX) system,

$$y(t+1) = a_0 y(t) + \cdots + a_n y(t-n) + b_0 u(t) + \cdots + b_n u(t-n) + w(t+1)$$

where $(a_0, a_1, \cdots, a_n, b_0, b_1, \cdots, b_n)^T := \theta$. θ, the unknown parameter, will be regarded as having a prior distribution which is normal $N(\bar{\theta}, \bar{\Sigma})$. $\{w(t)\}$ is a sequence of independent, identically distributed $N(0, \sigma^2)$ random variables, i.e., white Gaussian noise. The goal is to minimize $E \sum_1^N y^2(t)$ (say). (We assume that $(y(0), y(-1), \cdots, y(-n-1), u(0), \cdots, u(-n-1)$ are known initial conditions.)

We have seen in § 2 that a central part of the BACP is to obtain the posterior distribution of the unknown parameter θ, given the observations $y(0)$, $u(0)$, $y(1), u(1), \cdots, y(t), u(t)$ made up to time y. To solve this problem of obtaining the posterior distribution, we rewrite the above system as

$$\theta(t+1) = \theta(t), \qquad \theta(0) = \theta \sim N(\bar{\theta}, \bar{\Sigma}),$$

$$y(t+1) = \phi^T(t)\theta(t) + w(t+1)$$

where $\phi(t) := (y(t),\ y(t-1), \cdots, y(t-n),\ u(t),\ u(t-1), \cdots, u(t-n))^T$. Now it is clear that the posterior distribution is a normal distribution, the mean $\hat{\theta}(t)$ and covariance $\Sigma(t)$ of which are obtained through the Kalman filtering equations.

(If θ initially is *not* normally distributed, then some recent results for the continuous time problem, Makowski [35], may prove useful.)

The hyperstate for the BACP is thus $(\hat{\theta}(t), \Sigma(t), y(t), \cdots, y(t-n), u(t), \cdots, u(t-n))$. Thus it is perfectly clear that in view of the highly nonlinear manner in which the variables $\hat{\theta}(t)$ and $\Sigma(t)$ depend on the past $(y(t), u(t), y(t-1), u(t-1), \cdots)$, we really have a highly nonlinear system with a quadratic cost criterion. Indeed one can solve the problem of minimizing the finite horizon cost criterion $\sum_1^N y^2(t)$ when $N=1$, but no analytical solutions are available when $N \geqq 2$, see Åström and Wittenmark [36].

This lack of solutions has spurred many attempts at qualitatively understanding the nature of the optimal input sequence. The two qualitative features which are most popular are "caution" and "probing", which we shall briefly explain. Consider the simplest system of the type considered:

$$y(t+1) = y(t) + bu(t) + w(t+1)$$

where the prior distribution for b is $N(\bar{b}, \Sigma)$ with $\bar{b} > 0$. $\{w(t)\}$ is, as before, an i.i.d. $N(0, \sigma^2)$ sequence. $y(0)$ is known.

First consider the cost criterion $E(y^2(1))$. The optimal value of $u(0)$ is easily calculated to be

$$u(0) = \left(\frac{\bar{b}}{\bar{b}+\Sigma}\right) y(0).$$

Thus we note that as Σ increases, $|u(0)|$ decreases, i.e., as the uncertainty (here, variance) of b increases, the control (in absolute value) decreases. The controller is said to be *cautious.*

On the other hand, if the cost criterion is $E(y^2(2))$, then we know that

$$\min_{u(1)} E[y^2(2) | y(0), u(0), y(1)] = \frac{\Sigma(1)}{\hat{b}^2(1) + \Sigma(1)} y^2(1) + \sigma^2$$

where $\hat{b}(1) = E(b | u(0), y(1))$ and $\Sigma(1) = E[(b - \hat{b}(1))^2 | u(0), y(1)]$. Thus at time $t=1$, it is preferable to have a smaller value of $\Sigma(1)$—all other considerations remaining the same. Since $\Sigma(1)$, the conditional variance of the estimation error, is given by

$$\Sigma(1) = \frac{\sigma^2 \Sigma}{\sigma^2 + u^2(0)\Sigma}$$

it would appear at first sight that choosing a large value of $|u(0)|$ is helpful. The choice of large values of the control input to enhance identification is called "*probing*". However, increasing $|u(0)|$ could also increase $|y(1)| (= |y(0) + bu(0) + w(1)|)$ and hence also $y^2(1)$. This of course makes the cost larger. Thus there is a tradeoff between increasing the control to probe the system and decreasing the control to reduce the uncurred cost. This sort of a problem is therefore referred to as a "dual control" problem, in view of the possibly dual purposes in applying a control action.

For another discussion of the "dual" control effect, the reader is referred to that portion of § 3 where the work of Kelly [27] is discussed.

Problems of the type described here have been considered by Feldbaum [37], Jacobs and Patchell [38], Tse and Bar-Shalom [39]-[41], Bar-Shalom and Tse [42],

Wittenmark [43], Wenk and Bar-Shalom [44], Bar-Shalom [45], Deshpande, Upadhyay and Lainiotis [46], Lainiotis [47], Dersin, Athans and Kendrick [48]. [37] is an early reference. [38] treats an incompletely observed problem which allows computation of the optimal solution and examines the nature of resulting optimal control law. [39] proposes a control law based on replacing the nonlinear system by a version linearized (in some sense) around a trajectory resulting from a nominal control law. [41] and [42] show when one may expect a certainty-equivalence control law to be optimal. In [45] a decomposition of an approximation of the cost-to-go function into three components is made, which supposedly reflect caution, probing and the residual cost. In [44] the performance of an approximation of [46], [47], which consists of using a control which is the weighted average (given by the posterior distribution) of the optimal controls for the various parameters, is examined as well as another algorithm which first averages over the parameter values and then chooses the control based on it. [48] examines a particular problem and evaluates the results of [45].

In the continuous time case, Rishel [140] has recently shown that the optimal control can be written in terms of the solution of a certain stochastic two-point boundary value problem. Also, Hijab has shown that for a particular measurement equation, and a particular cost criterion including an entropy term, the optimal control is the conditional mean of the optimal controls in the various parameter values, see [141].

6. Non-Bayesian adaptive control. In the *Bayesian* adaptive control problem (BACP), the specification of the problem is fairly rigid. Given an a priori probability distribution for the unknown parameter, one has to obtain a control law which minimizes the *expected* value of a certain cost criterion.

In the *non-Bayesian* adaptive control problem (NACP), more flexibility is allowed in the *design* of control laws. However, the designed control law must meet certain other (asymptotic) criteria in order to be deemed acceptable. To illustrate some of the central concepts, we start with a very simple example due to Robbins [49]—the non-Bayesian two-armed bandit adaptive control problem.

6.1. The non-Bayesian two-armed bandit problem. There are two slot machines: A and B. When machine A (or B) is used, one obtains one unit of reward with probability p_A (or p_B) and zero units of reward with probability $1-p_A$ (or $1-p_B$). Without loss of generality, we assume $p_A > p_B$.

Consider first the case when p_A and p_B are known. If at *each* time $t = 1, 2, 3, \cdots$ we play machine A exclusively, then $\lim (1/N)\sum_1^N r_t = p_A$ a.s., where $r_t :=$ random reward earned at time t. This is, almost surely, the maximum long-term average reward that one could possibly gain even by switching between machines, and obviously this is achievable by playing machine A exclusively.

Suppose now that we do *not* know the values of p_A and p_B. All we know is that $p_A \in (0, 1)$ and $p_B \in (0, 1)$. (Note that we are *not* provided with an initial probability distribution for $\theta := (p_A, p_B)$. This is what distinguishes this from a BACP.) However, we are as ambitious as before. We would still like to have a policy of playing the machines which ensures that $\lim (1/N)\sum_1^N r_t = \max (p_A, p_B)$ a.s.

We now exhibit a policy which achieves this goal. Let $u_t \in \{A, B\}$ be the control input chosen at time t, where $u_t = A$ or B according to whether A or B is played. Let $y_t \in \{0, 1\}$ denote the observation made at time t, where $y_t = 0$ or 1 according to whether a reward was not earned or was earned at time t. Here $r_t = y_t$.

To define the policy, we first choose any increasing sequence $\{a_n\}$ of positive integers, such that $\lim (1/N)\sum_{t=1}^N 1(t = a_i \text{ for some } i) = 0$. At each time t, we make

"estimates" $\hat{p}_A(t)$ and $\hat{p}_B(t)$ of p_A and p_B respectively, by

$$\hat{p}_A(t) := \frac{\sum_0^{t-1} 1(u_n = A, y_n = 1)}{\sum_0^{t-1} 1(u_n = A)} \quad \text{and} \quad \hat{p}_B(t) := \frac{\sum_0^{t-1} 1(u_n = B, y_n = 1)}{\sum_0^{t-1} 1(u_n = B)}.$$

We note, in passing, that these are the maximum likelihood estimates of p_A and p_B. We now choose u_t according to:

$$u_t = \begin{cases} A & \text{if } (t = a_{2n} \text{ for some } n) \text{ or } (t \neq a_n \text{ for all } n \text{ and } \hat{p}_A(t) \geqq \hat{p}_B(t)), \\ B & \text{if } (t = a_{2n+1} \text{ for some } n) \text{ or } (t \neq a_n \text{ for all } n \text{ and } \hat{p}_A(t) < \hat{p}_B(t)). \end{cases}$$

Basically, except for the times $t = a_1, a_2, a_3, \cdots$ we play whichever of A or B has the larger "estimated probability" of a win. However, the times $t = a_1, a_2, a_3, \cdots$ are *reserved* for experimentation.

Now we show that the above policy attains our goal $\lim (1/N) \sum_1^N r_t = \max (p_A, p_B)$ a.s., and does so *without* knowing the values of (p_A, p_B). To see this, first note that by the reservation of the experimentation times, each of A and B is played infinitely often (i.o.) a.s. By the law of large numbers therefore, $\hat{p}_A(t) \to p_A$ and $\hat{p}_B(t) \to \hat{p}_B$ a.s., and so for all $t \geqq$ some $T(\omega)$, $\hat{p}_A(t) > \hat{p}_B(t)$. So, A is exclusively played after time T, except at some of the *reserved* times. But these reserved times are so sparse, that they make no contribution to the *average* cost. More precisely:

$$\begin{aligned} \lim \frac{1}{N} \sum_1^N r_t &\geqq \lim \frac{1}{N} \sum_1^N 1(u_t = A, y_t = 1) \\ &= \lim \frac{1}{N} \sum_1^N 1(u_t = A) \lim \frac{\sum_1^N 1(u_t = A, y_t = 1)}{\sum_1^N 1(u_t = A)} \\ &= p_A \lim \frac{1}{N} \sum_1^N 1(u_t = A) \\ &\geqq \lim \frac{p_A}{N} \left(N - T - \sum_{t=1}^N 1(t = a_i \text{ for some } i) \right) = p_A \quad \text{a.s.} \end{aligned}$$

Three properties of this non-Bayesian adaptive control scheme should be noted.

i) $\lim (1/N) \sum_1^N r_t = \max (p_A, p_B)$ a.s. Thus, the cost of this scheme is *optimal*, i.e., it could not be bettered even if we knew the values of p_A and p_B.

ii) $\lim (\hat{p}_A(t), \hat{p}_B(t)) = (p_A, p_B)$. Thus the parameter estimates are consistent, i.e., the true parameters are identified.

iii) $\lim u_t$ does not exist. Hence the control scheme does not converge, and therefore, of course, it does not converge to an optimal control scheme. (However, it does converge in a Cesaro sense, $\lim (1/N) \sum_1^N 1(u_t = A) = 1$ a.s.)

6.2. Non-Bayesian adaptive control versus Bayesian adaptive control. We shall discuss some of the differences between the Bayesian and non-Bayesian approaches to adaptive control.

In the previous section, we have obtained a policy π for which

$$\lim \frac{1}{N} \sum_1^N r_t = \max (p_A, p_B) \quad \text{a.s.}$$

A more precise way of stating the above fact is as follows. There exists a policy π such that

$$lim \frac{1}{N} \sum_1^N r_t = \max J(\theta^0), \qquad P_{\theta^0}^{\pi}\text{-a.s. for every } \theta^0 \in \Theta.$$

(Here $\theta^0 := (p_A, p_B)$, $J(\theta^0) := \max(p_A, p_B)$, $\Theta = (0,1)\times(0,1)$ and $P_{\theta^0}^{\pi}$ is the probability measure induced on the trajectories of the system by the policy π when the parameter value is θ^0.) This clearly shows that no matter *what* the value of θ^0 is, the policy π attains the maximum reward attainable for *that* value of θ^0. In non-Bayesian adaptive control with respect to an *average* cost criterion, we shall frequently impose such a requirement on a policy, viz. it should be optimal *uniformly* for all $\theta^0 \in \Theta$ a.s.

If the requirement above can be met by some policy π, then it is clear that *all* Bayesian problems with the above cost criterion are also immediately solved. The reason is that if $q(\theta)$ is the prior distribution of θ, then the policy π, when implemented, attains the expected cost $\sum_\theta q(\theta)J(\theta)$, and clearly no policy can do better. (Thus π attains the obvious lower bound in Theorem 4.1 of § 4.) Hence π is optimal even in a Bayesian framework *irrespective* of the prior distribution q.

We will show in the sequel that one can often obtain a policy π meeting the requirements above, and frequently there will be several π's which do so. Thus, insofar as just the long-term average cost criterion is concerned, the non-Bayesian formulation is unquestionably superior to the Bayesian formulation.

Optimal policies for the long term average cost criterion are nonunique in an essential way, since what happens in the initial period does not alter the cost. Since one is often (practically) interested in attaining a fast rate of convergence to optimal control laws (or a fast rate of convergence of the parameter estimates etc.), one may choose between several π's meeting the above requirement by imposing some other criterion, such as rate of convergence, to judge them. Alternatively, one could pose, as in § 3 in the discussion of [27]–[29], the problem of obtaining a policy which is optimal for the average cost criterion and nearly optimal for the discounted cost problem or vice-versa. Or, one could study the rate at which the difference between the finite horizon cost of adaptive control and its optimal value grows with the horizon.

However, if the cost criterion is of the discounted type $\sum_0^\infty \beta^t c(x_t, u_t)$, then the BACP and NACP are rather different. For the NACP, it is clear that one cannot expect that $\sum_0^\infty \beta^t c(x_t, u_t)$ will converge a.s. to a constant, as it did in the average cost case. If we replace this by the requirement that $E_{\theta^0}^{\pi} \sum_0^\infty \beta^t c(x_t, u_t) = J(x_0, \theta^0)$ for all $\theta^0 \in \Theta$, then this is clearly too strong to be met by a single π. So finally, we arrive at a requirement of the type,

$$\lim_N \left\{ E_{\theta^0}^{\pi} \sum_N^\infty \beta^{t-N} c(x_t, u_t) - E_{\theta^0}^{\pi} J(x_N, \theta^0) \right\} = 0 \quad \text{for all } \theta^0 \in \Theta$$

which is of a reasonable nature, see Schäl [63].

So far, however, we have only paid attention to the convergence of the costs, and *not* the controls. So, we now address the problem of convergence of control laws in a NACP. For each $\theta \in \Theta$, let π^θ be an optimal stationary policy for the discounted cost problem (and for simplicity of discussion we assume that it is unique). Then one can ask the following question: Is there a policy $\pi = (\pi_0, \pi_1, \cdots)$ for which

$$\lim \{\pi_N(x_0, u_0, x_1, u_1, \cdots, x_N) - \pi^{\theta^0}(x_N)\} = 0, \qquad P_{\theta^0}^{\pi}\text{-a.s. for all } \theta^0 \in \Theta \quad ?$$

Thus we are requiring that the controls generated by the adaptive scheme π should converge asymptotically to the optimal controls, uniformly for all $\theta^0 \in \Theta$. This, again, is a reasonable requirement in many instances.

Since the asymptotic requirements on the costs and the controls are closely related, we shall refer to either of these requirements (perhaps with slight modifications), in an NACP, as a *self-optimizing* requirement. One of the main goals of non-Bayesian adaptive control is to obtain a self-optimizing policy.

This situation is in contrast to the BACP's for which, frequently, the optimal policy is not self-optimizing. An example is given in [26], and the reader is referred to the discussion in § 3 on [26], [27].

From a practical point of view, in a BACP, one is typically faced with problems where the computational burden is very high. For an NACP, one is typically interested in the rate of convergence of the self-optimizing policy, if indeed one can obtain one.

7. Non-Bayesian adaptive control of Markov chains. To begin, we consider the case where the state, control and parameter spaces, X, U and Θ are all finite. The transition probabilities are given, for each $\theta \in \Theta$, by $\{p(i, j, u; \theta): i, j \in X, u \in U\}$. The value of the true parameter is θ^0. All we know is that θ^0 is *some* element of Θ. Our goal is to design a policy π which a.s. attains the minimum of $\lim (1/N) \sum_1^N c(x_t, u_t)$, where $c(i, u)$ is a one-stage cost function.

We start by considering a scheme that, *at first sight*, looks very reasonable. At each time t we will have accumulated a history $(x_0, u_0, x_1, u_1, \cdots, x_t)$ and we can use this history to form (say) a maximum-likelihood estimate (MLE) $\hat{\theta}_t$ of the unknown parameter. Thus, we choose $\hat{\theta}_t \in \Theta$ so that

$$\prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \hat{\theta}_t) \geqq \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \theta) \quad \text{for all } \theta \in \Theta.$$

(In the event that there is more than one maximizer of the likelihood function, one can choose a particular maximizer according to some prespecified priority order on elements of Θ.) Then we choose a control input u_t which is optimal if the parameter value was $\hat{\theta}_t$, i.e., we choose

$$u_t = \phi(x_t, \hat{\theta}_t)$$

where, for each $\theta \in \Theta$, $\phi(\cdot, \theta): X \to U$ is an optimal stationary control law (policy).

Many questions arise.

i) Does $\hat{\theta}_t$ converge a.s.?

ii) Does $\hat{\theta}_t$ converge to θ^0 a.s.?

iii) Does $\phi(\cdot, \hat{\theta}_t)$ converge a.s.?

iv) Does $\phi(\cdot, \hat{\theta}_t)$ converge to $\phi(\cdot, \theta^0)$ a.s.?

v) Does $\lim (1/N) \sum_1^N c(x_t, u_t)$ converge a.s.?

vi) Does $\lim (1/N) \sum_1^N c(x_t, u_t)$ converge to $J(\theta^0)$ a.s.? Here $J(\theta^0)$ is the optimal cost achievable for the parameter θ^0, and we assume that it does not depend on the initial state x_0.

vii) At what rate do these quantities converge, if they do so?

In a nice counterexample, Borkar and Varaiya [50] demonstrate that (ii) need not hold. We provide below a counterexample in a similar vein, from [51], to illustrate that (ii), (iv) and (vi) do not hold.

Counterexample. Let $X = \{1, 2\}$, $U = \{1, 2\}$, $\Theta = \{1, 2, 3\}$, $\theta^0 = 1$. The transition probabilities are $p(1, 1, 2; 1) = p(1, 1, 2; 2) = 0.8$, $p(1, 1, 2; 3) = 0.2$, $p(2, 1, u; \theta) = 1$ for all u, θ. The cost function $c(x_t, u_t, x_{t+1})$ is $c(i, u, j) = 3 + (2 - i)\ (7.8 - 0.3u - b_j)$. It is easily calculated, see [33], that the optimal policies are $\phi(i, 1) = 1$, $\phi(i, 2) = \phi(i, 3) = 2$ for all $i \in X$.

To see that $\hat{\theta}_t$ need not converge to 1, consider the starting values $x_0 = 1$, $u_0 = 1$. With probability 0.5, the next state is $x_1 = 1$. But then $\hat{\theta}_1 = 2$. Hence $u_1 = \phi(1, 2) = 2$. It can then be checked easily that $\hat{\theta}_t = 2$ for all $t \geqq 1$. Thus, there is a probability of at least 0.5 that $\lim \hat{\theta}_t \neq \theta^0$. This shows that the answer to each of the questions (ii), (iv) and (vi) is a no. □

The basic problem here is that one cannot fully identify a process in closed loop. Borkar and Varaiya [50] show the best that one may expect.

THEOREM 7.1 (Borkar and Varaiya). *If* $p(i, j, u, \theta) \geqq \varepsilon$ *for all* i, j, u, θ *then* $\lim \hat{\theta}_t = \hat{\theta}_\infty$ *a.s. and* $p(i, j, \phi(i, \hat{\theta}_\infty), \theta^0) = p(i, j, \phi(i, \hat{\theta}_\infty), \hat{\theta}_\infty)$ *a.s.*

Proof. Define $L_t(\theta) := \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \theta) p^{-1}(x_s, x_{s+1}, u_s; \theta^0)$, the likelihood ratio. For each $\theta \in \Theta$, $\{L_t(\theta), \mathcal{F}_t\}$ is a positive martingale (where $\mathcal{F}_t := \sigma\{x_0, u_0, x_1, u_1, \cdots, x_t\}$ is the σ-algebra generated by the observed past). Hence $\{L_t(\theta)\}$ converges a.s. for every $\theta \in \Theta$. Fix $\omega \in \Omega$, the sample space. If $\{\hat{\theta}_t(\omega)\}$ has $\bar{\theta} \in \Theta$ as a limit point, then $\hat{\theta}_t(\omega) = \bar{\theta}$ i.o. Since $L_t(\hat{\theta}_t) \geqq L_t(\theta^0) = 1$, it follows that $L_t(\bar{\theta}, \omega) \geqq 1$ i.o., and so $\lim L_t(\bar{\theta}, \omega) \geqq 1$ (as long as ω is not in a certain null set). This shows that $p(x_t(\omega), x_{t+1}(\omega), u_t(\omega); \bar{\theta}) p^{-1}(x_t(\omega), x_{t+1}(\omega), u_t(\omega); \theta^0) = L_{t+1}(\bar{\theta}, \omega) L_t^{-1}(\bar{\theta}, \omega) \to 1$. Hence, for all $t \geqq T(\omega)$, $p(x_t(\omega), x_{t+1}(\omega), u_t(\omega), \bar{\theta}) = p(x_t(\omega), x_{t+1}(\omega), u_t(\omega), \theta^0)$. If $\bar{\bar{\theta}}$ is any other limit point of $\{\hat{\theta}_t(\omega)\}$, then a similar result holds for all $t \geqq \bar{T}(\omega)$. Hence $L_t(\bar{\theta}, \omega) = L_t(\bar{\bar{\theta}}, \omega)$ for all $t \geqq \max(T(\omega), \bar{T}(\omega)) =: \bar{\bar{T}}(\omega)$. Since one always breaks ties by picking the particular maximizer which is highest in the priority ordering, it follows that $\bar{\theta} = \bar{\bar{\theta}}$, showing that $\hat{\theta}_t \to \hat{\theta}_\infty$ a.s. Hence we obtain $p(x_t, x_{t+1}, u_t; \theta_\infty) = p(x_t, x_{t+1}, u_t; \theta^0)$ a.s. By the Martingale Stability Theorem, $\lim (1/N) \sum_1^N 1(x_t = i, x_{t+1} = j) - E(1(x_t = i, x_{t+1} = j) | \mathcal{F}_{t-1}) = 0$ a.s. and since $E(1(x_t = i, x_{t+1} = j) | \mathcal{F}_{t-1}) \geqq \varepsilon^2$ it follows that $(x_t = i, x_{t+1} = j)$ i.o. a.s. Hence, the result follows. □

The above result is fundamental and needs elaboration. Why does $\hat{\theta}_t$ not converge to θ^0? The answer is this. If one could guarantee that $(x_t = i, u_t = u)$ i.o. for every (i, u), then one could hope to identify $p(i, j, u; \theta^0)$. But in closed loop one uses only those u_t's for which $u_t = \phi(x_t, \hat{\theta}_t)$. Thus if $\hat{\theta}_t \to \bar{\theta}$, then $(x_t = i, u_t = u)$ i.o. *only* if $u = \phi(i, \bar{\theta})$, and so one can only hope to identify $p(i, j, \phi(i, \bar{\theta}); \theta^0)$, and this is the content of the above theorem.

Sagalovsky [52] treats the situation where the unknown probabilities depend in an affine way on a real valued unknown parameter, and shows how this structure can be exploited. [53] shows what further difficulties are encountered in generalizing Theorem 7.1 to the case where Θ is compact, instead of finite as above. Borkar and Varaiya [54] consider the situation where the state space is countable.

Earlier, Mandl [55]–[57] had considered the problem where U, Θ are compact, $p(i, j, u; \theta)$ and $c(i, u)$ continuous and, for simplicity, $p(i, j, u; \theta) > 0$. [55] considers the class of estimators based on general "contrast" functions. This class includes the maximum likelihood estimator considered above, if the following assumption is made.

Identifiability condition. For every $\theta, \theta' \in \Theta$ and $\theta \neq \theta'$, there exists an $i = i(\theta, \theta') \in X$ such that for every $u \in U$, there is a $j = j(i, u, \theta, \theta') \in X$ with $p(i, j, u; \theta) \neq p(i, j, u; \theta')$.

THEOREM 7.2 (Mandl). *If* U *and* Θ *are compact,* $p(\cdot)$, $c(\cdot)$, $\phi(\cdot)$ *and* $g(\cdot)$ *(see below) are continuous,* $p(i, j, u; \theta) > 0$ *and the Identifiability Condition is satisfied, then*

i) *For any policy* $\pi = (\pi_0, \pi_1, \cdots)$, $\lim \hat{\theta}_t = \theta^0$ *a.s.*

ii) *If* $\pi_t(x_0, u_0, x_1, \cdots, x_t) := \phi(x_t, \hat{\theta}_t)$, *and* ϕ *is continuous, then* $\lim (1/N) \sum_1^N c(x_t, u_t) = J(\theta^0)$ *a.s.*

Proof. The proof of part (i) is omitted, because the reader can deduce that at least for the policy of (ii), and when Θ is finite, Theorem 7.1 and the Identifiability Condition give the required result. For (ii), we use the theory of the long-term average cost criterion [33] to note that there exist $\{v_i(\theta): i \in X\}$ satisfying

$$J(\theta) + v_i(\theta) = c(i, \phi(i, \theta)) + \sum_j p(i, j, \phi(i, \theta); \theta) v_j(\theta)$$

$$\leqq c(i, u) + \sum_j p(i, j, u; \theta) v_j(\theta) \quad \text{for all } u, i, \theta.$$

Defining $g(i, u, \theta) := c(i, u) + \sum_j p(i, j, u; \theta) v_j(\theta) - J(\theta) - v_i(\theta)$, we see that $g(i, u, \theta) \geqq 0$ and $g(i, u, \phi(i, \theta)) = 0$. Now note that if $y(t+1) := c(x_t, u_t) - J(\theta^0) + v(x_{t+1}, \theta^0) - v(x_t, \theta^0) - g(x_t, u_t, \theta^0)$, then $E(y_{t+1}|\mathcal{F}_t) = 0$, and since $\{y_{t+1}\}$ is bounded, it follows by the Martingale Stability Theorem that $\lim (1/N) \sum_1^N y_{t+1} = \lim (1/N) \sum_1^N E(y_{t+1}|\mathcal{F}_t) = 0$ a.s. Hence, substituting for y_{t+1} and noting that $\{v_i(\theta^0)\}$ is bounded in i, gives $\lim (1/N) \sum_1^N c(x_t, u_t) = J(\theta^0) + \lim (1/N) \sum_1^N g(x_t, u_t, \theta^0)$. By part (i) however, $\lim (u_t - \phi(x_t, \theta^0)) = \lim (\phi(x_t, \hat{\theta}_t) - \phi(x_t, \theta^0)) = 0$, and so, by continuity of ϕ and g, $\lim g(x_t, u_t, \theta^0) = 0$. □

Note that a result such as (i) is very strong, since it guarantees that $\hat{\theta}_t \to \theta^0$ a.s. for *all* policies π. This of course points to the restrictive nature of the Identifiability Condition. However, in some problems, for example, in the control of queueing systems, such a condition is satisfied.

Kurano [58] considers a similar problem. Kolonko [59], [60] determines the appropriate generalization of the Identifiability Condition and other regularity assumptions which are sufficient to ensure that results such as (i) and (ii) of the above theorem hold in the adaptive control of Markov renewal processes. Applications to the adaptive control of queueing systems are also studied. Georgin [61] also considers the generalization of the above theorem to more general state spaces. Baranov [62] considers a different scheme where even the control laws are obtained by a recursive process.

(i) of the above theorem also has implications for cost criteria other than of the long-term average type. For example, consider the case of a discounted cost criterion, and suppose that the policy $\pi = (\pi_0, \pi_1, \pi_2, \cdots)$ is such that $\pi_t(x_0, \cdots, x_t) = \phi(x_t, \hat{\theta}_t)$ where $\phi(\cdot, \theta)$ is a stationary control law which is optimal for the discounted cost problem. Then, under reasonable continuity conditions, $\phi(\cdot, \hat{\theta}_t) \to \phi(\cdot, \theta^0)$. Under appropriate conditions, it then also follows that $E \sum_t^\infty \beta^{n-t} c(x_n, u_n) - EJ(x_t, \theta^0) \to 0$ a.s. Such a generalization is carried out in Schal [63].

How does one deal with the general situation where an Identifiability Condition may not hold? This is a difficult problem and essentially requires some procedure for overcoming the fundamental closed loop identification problem. There are several ways of doing this, which we now take up for consideration.

7.1. Forced choice schemes. Here, just as in the non-Bayesian bandit problem of § 6.1, some time instants $t = a_1, a_2, a_3, \cdots$ are set aside for experimentation. At these times, one must use forced choices of all the elements of U in (say) cyclic order. Thus, since $(x_t = i, u_t = u)$ occurs i.o. a.s., it follows that $\hat{\theta}_t \to \theta^0$ a.s. However if $\lim (1/N) \sum_{t=1}^N 1(t = a_i \text{ for some } i) = 0$, then the control actions taken at times $t = a_1, a_2, \cdots$ do not make any contribution to the average cost. At other times one just uses the control inputs $u_t = \phi(x_t, \hat{\theta}_t)$. Since one has $\hat{\theta}_t \to \theta^0$ a.s., it follows that optimal cost can be obtained. An approach of this type is followed in Fox and Rolph [64] for Markov renewal processes. Van Hee [14] considers a comparable scheme albeit in a Bayesian formulation.

7.2. Randomization schemes. These are schemes for which each $\pi_t(\cdot|x_0, u_0, x_1, \cdots, x_t)$ is allowed to be a probability measure on U according to which u_t is picked. Since u_t is random, every $u \in U$ has (in some sense) a positive probability of being applied at each time.

Doshi and Shreve [65] impose a less restrictive condition than the identifiability condition given earlier. At each time instant t, a parameter $\hat{\theta}_t$ is randomly chosen from among all those which very nearly maximize the likelihood function. This is then shown to overcome the identifiability problem. Borkar and Varaiya [54] also consider similar randomization schemes for countable Markov chains.

Sato, Ake and Takeda [66] have proposed quite a different scheme for the *discounted* cost problem, which we now describe. Let us assume that:

i) $p(i, j, u; \theta^0) > 0$ for all i, j, u.

ii) There is an *unique* optimal stationary control law $\phi(\cdot, \theta^0)$ for the parameter value θ^0.

The algorithm to generate the controls proceeds as follows.

SATO, ABE, TAKEDA ALGORITHM

Step 1. Set $t = 0$ and choose $\{\hat{p}_0(i, j, u)\}$, "estimates" of the transition probabilities, so that there is an unique stationary optimal control law ϕ_0 for these estimates.

Step 2. Choose $\{n_0(i, j, u)\}$ and $\{n_0(i, u)\}$, all positive, so that $\hat{p}_0(i, j, u) = n_0(i, j, u) n_0^{-1}(i, u)$ for all i, j, u.

Step 3. Choose $\gamma_0 > 0$ so small that if one defines

$$S_0(i, j, u) := \frac{n_0(i, j, u)}{n_0(i, u) + g(\gamma_0)} \quad \text{and} \quad B_0(i, j, u) := \frac{n_0(i, j, u) + g(\gamma_0)}{n_0(i, u) + g(\gamma_0)}$$

where $g(\gamma) := \gamma/(1-\gamma)$, then "for all models M satisfying $S_0 \leqq M \leqq B_0$", the policy ϕ_0 is still optimal. By a model M, we shall mean a set of transition probabilities $\{p(i, j, u)\}$ satisfying $p(i, j, u) \geqq 0$ and $\sum_j p(i, j, u) = 1$ for all i, u. By $S_0 \leqq M \leqq B_0$ we shall mean $S_0(i, j, u) \leqq p(i, j, u) \leqq B_0(i, j, u)$ for all i, j, u.

Step 4. Apply $u_t = \phi_t(x_t)$ with probability γ_t and all other elements of U with equal probability.

Step 5. Set $t = t + 1$.

Step 6. Set $n_t(i, j, u) = n_{t-1}(i, j, u) + 1(x_{t-1} = i, x_t = j, u_{t-1} = u)$ for all i, j, u and $n_t(i, u) = n_{t-1}(i, u) + 1(x_{t-1} = i, u_{t-1} = u)$ for all i, u.

Step 7. Set

$$S_t(i, j, u) := \frac{n_t(i, j, u)}{n_t(i, u) + g(\gamma_{t-1})} \quad \text{for all } i, j, u,$$

$$B_t(i, j, u) = \frac{n_t(i, j, u) + g(\gamma_{t-1})}{n_t(i, u) + g(\gamma_{t-1})} \quad \text{for all } i, j, u.$$

Step 8. If there is a single control law ϕ_t which is optimal for all models M with $S_t \leqq M \leqq B_t$, then set $\gamma_t = h(\gamma_{t-1})$ where $h(\gamma) := (1+\gamma)/2$ and go to Step 4. If a ϕ_t as above cannot be chosen, then set $\gamma_t := \gamma_{t-1}$, $\phi_t := \phi_{t-1}$ and go to Step 4.

THEOREM 7.3 (Sato, Abe, Takeda). *Assume*

i) $p(i, j, u; \theta^0) > 0$ *for all* i, j, u;

ii) *for* θ^0 *there is an* <u>unique</u> *stationary optimal control law* $\phi(\cdot, \theta^0)$ *for the discounted cost problem.*

Then for the above algorithm, $\lim \phi_t = \phi(\cdot, \theta^0)$ *a.s.*

Proof. First note that $\{\gamma_t\}$ is an increasing sequence. Assume that on a subset $\tilde{\Omega} \subseteq \Omega$ of the sample space of positive measure, $\gamma_t \to \gamma_\infty < 1$. By Step 4, $u_t = u$ i.o. for every u on $\tilde{\Omega}$, and by our assumptions $n_t(i, u) \to \infty$. By the law of large numbers, therefore, $n_t(i, j, u)/n_t(i, u) =: \hat{p}_t(i, j, u) \to p(i, j, u; \theta^0)$ on $\tilde{\Omega}$. This means for every $\omega \in \tilde{\Omega}$, for every $\varepsilon > 0$, there is a $T(\omega)$ large enough so that $[S_t(i, j, u), B_t(i, j, u)] \subseteq [p(i, j, u; \theta^0) - \varepsilon, p(i, j, u; \theta^0) + \varepsilon]$ for all $t \geqq T$. Now note that by (ii), *all* control laws other than $\phi(\cdot, \theta^0)$ produce a cost strictly larger than $\phi(\cdot, \theta^0)$ does. By continuity, therefore, this must also hold for all models M with $p(\cdot, \theta^0) - \varepsilon \leqq M \leqq p(\cdot, \theta^0) + \varepsilon$ for $\varepsilon > 0$ sufficiently small. Hence, for all $t \geqq$ some T', $\phi(\cdot, \theta^0)$ is optimal for all models M with $S_t \leqq M \leqq B_t$ on $\tilde{\Omega}$. Hence by Step 8, it must be the case that $\gamma_t = h(\gamma_{t-1})$ for all $t \geqq T'$ on $\tilde{\Omega}$. But this shows that $\gamma_t \to 1$ on $\tilde{\Omega}$, contradicting our original assumption. Hence we can deduce that $\gamma_t \to 1$ a.s.

For every (i, j, u), either $\lim n_t(i, j, u) = +\infty$ or not. If the former, then we have already seen a demonstration that $[S_t(i, j, u), B_t(i, j, u)] \subseteq [p(i, j, u; \theta^0) - \varepsilon, p(i, j, u; \theta^0) + \varepsilon]$ for all $t \geqq$ some T. If the latter, then since $\gamma_t \to 1$, it follows that $g(\gamma_t) \to +\infty$ and so by Step 7, $S_t(i, j, u) \to 0$ and $B_t(i, j, u) \to 1$. In either case therefore we see that $[p(i, j, u; \theta^0) - \varepsilon, p(i, j, u; \theta^0) + \varepsilon] \cap [S_t(i, j, u), B_t(i, j, u)]$ is nonempty for all $t \geqq$ some T. Choose $\varepsilon > 0$ so small that $\phi(\cdot, \theta^0)$ is the *only* stationary control law which is optimal for all models M with $p(\cdot, \theta^0) - \varepsilon \leqq M \leqq p(\cdot, \theta^0) + \varepsilon$. Then, for all $t \geqq$ some $\bar{T}$, the only stationary control law which is possibly optimal for all models M with $S_t \leqq M \leqq B_t$, is $\phi(\cdot, \theta^0)$, provided there is one such control law.

Now let ϕ^* be any limit point of $\{\phi_t\}$. Since the set of stationary control laws is finite, it follows that $\phi^* = \phi_t$ i.o., and since $\gamma(t) \to 1$, it follows that ϕ^* is optimal for all models M with $S_t \leqq M \leqq B_t$ for some $t \geqq \bar{T}$. Hence $\phi^* = \phi(\cdot, \theta^0)$, showing that $\phi_t \to \phi(\cdot, \theta^0)$ a.s. □

Another approach to the problem of generating an adaptive control policy is by using methods common in the theory of learning automata. Lyubchik and Poznyak [67] have proposed several such schemes. No proofs are provided and the identifiability issue is not alluded to. El-Fattah [68], [69] has analyzed one recursive identification and control scheme, which we now examine.

Assume that $\Theta \subseteq \mathbb{R}^n$ is a Cartesian product of closed intervals. Consider the following stochastic approximation type scheme, see Tsypkin [70], [71], for the generation of the parameter estimates:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \frac{\gamma}{t+1} \frac{\nabla(\sum_u p(x_t, x_{t+1}, u; \hat{\theta}_t)\pi_t(u|x_t))}{\sum_u p(x_t, x_{t+1}, u; \hat{\theta}_t)\pi_t(u|x_t)}.$$

Here $\pi_t(u|x_t)$ is the probability of using the control $u_t = u$ at time t, given x_t. This is a reasonable updating scheme since, at each time t, we take a step $(\hat{\theta}_{t+1} - \hat{\theta}_t)$ in the direction of the gradient, i.e., in the direction in the parameter space in which the probability of the transition from (x_t, u_t) to x_{t+1} increases most rapidly. The probabilities $\{\pi_t(u|x); u \in U, x \in X\}$ which specify the randomization scheme, are updated as follows:

$$\pi_{t+1}(u|x) = \begin{cases} \pi_t(u|x) & \text{if } x \neq x_t, \\ \pi_t(u|x) + \dfrac{a\Delta}{t+1} & \text{if } u = u_t \text{ and } x = x_t, \\ \pi_t(u|x) - \dfrac{a\Delta}{t+1} \cdot \dfrac{1}{|U|-1} & \text{if } u \neq u_t \text{ and } x = x_t \quad (|U| = \text{cardinality of } U). \end{cases}$$

Here $\Delta := f_{x_t, u_t}(\pi_t, \hat{\theta}_{t+1})$ where we have assumed that there are functions $F_x(\pi, \theta)$ and $f_{x,u}(\pi, \theta)$ satisfying $F_x(\pi, \theta) > 0$ and such that $\partial J(\pi, \theta)/\partial \pi(u|x) = F_x(\pi, \theta) f_{x,u}(\pi, \theta)$. Here $J(\pi, \theta)$ is the cost of using the stationary policy π when the parameter value is θ. Ignoring this assumption on the representation of $\partial J(\pi, \theta)/\partial \pi(u|x)$, the updating scheme for $\pi_{t+1}(u|x)$ is reasonable because it merely increases the probability of using u in state x, if this tends (infinitesimally) to reduce the cost. The term $1/(t+1)$ tends to reduce the "step-sizes" at a rate appropriate for convergence, and is standard in stochastic approximation theory.

THEOREM 7.4 (El-Fattah). *Consider the following conditions*:

i) $\Theta \subseteq \mathbb{R}^n$ *is a Cartesian product of closed intervals.*

ii) $p(i, j, u; \theta) > 0$ *for all* i, j, u, θ.

iii) *For each state* $x \in X$, *there is a possibly empty set* $S_x \subseteq \{1, 2, \cdots, n\}$ *such that*

$\bigcup_x S_x = \{1, 2, \cdots, n\}$ *and there is a* $c > 0$ *such that*

$$\sum_{k \in A} R_k^{\pi}(x, \theta)(\theta^k - \theta^{0k}) \leqq -c \sum_{k \in A} (\theta^k - \theta^{0k})^2 \quad \textit{for all } \pi, \textit{ if } A = S_x$$

$$= 0 \quad \textit{for all } \pi, \textit{ if } A = S_x^c.$$

Here

$$R^{\pi}(x, \theta) := E\left[\frac{\nabla \sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t)}{\sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t)}\middle| x_t = x, \theta^0\right].$$

iv) *There exist* $\lambda_1 > 0$, $\lambda_2 > 0$ *so that, for every* $k = 1, 2, \cdots, n$; $\operatorname{tr} K^{\pi}(x, \theta) \leqq \lambda_1 + \lambda_2 \|\theta - \theta^0\|^2$ *for all* π, *where*

$$K^{\pi}(x, \theta) := E\left[\frac{(\nabla \sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t))(\nabla^T \sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t))}{(\sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t))^2}\middle| x_t = x, \theta^0\right].$$

v) *There exist functions* $F_x(\pi, \theta) > 0$ *and* $f_{x,u}(\pi, \theta)$ *such that* $\partial J(\pi, \theta)/\partial \pi(u|x) = F_x(\pi, \theta) f_{x,u}(\pi, \theta)$ *for all* π, θ, u, x.

vi) *There exists* $c_1 > 0$ *such that*

$$\sum_u (f_{x,u}(\pi, \theta) - f_{x,u}(\pi, \theta^0))^2 \leqq c_1 \sum_{k \in S_x} (\theta^k - \theta^{0k})^2 \quad \textit{for all } x \in X.$$

vii)

$$\sum_u (f_{x,u}(\pi, \theta))^2 \pi(u|x) \geqq \lambda > 0 \quad \textit{for all } x \in X.$$

Then:

viii) *If* (i)-(iv) *hold, then for* <u>*any*</u> *nonanticipative randomized policy,*

$$\lim \hat{\theta}_t = \theta^0 \quad a.s. \quad \textit{and} \quad \sum_t \frac{1}{t+1} \sum_{k \in S_{x_t}} (\hat{\theta}_t^k - \theta^{0k})^2 < \infty \quad a.s.$$

ix) *If* (i)-(vii) *hold, then the adaptive scheme above is such that* $\lim (1/N) \sum_1^N c(x_t, u_t) = J(\theta^0)$ *a.s.*

Proof. Define the "stochastic Lyapunov function", $V_t := \|\hat{\theta}_t - \theta^0\|^2$. By calculation, we obtain (with $\mathcal{F}_t := \sigma(x_0, u_0, x_1, \cdots, x_t)$)

$$E(V_{t+1}|\mathcal{F}_t) = V_t + \frac{2\gamma}{t+1} \sum_{k \in S_{x_t}} (\hat{\theta}_t^k - \theta^{0k}) R_k^{\pi_t}(x_t, \hat{\theta}_t) + \frac{\gamma^2}{(t+1)^2} \operatorname{tr} K^{\pi_t}(x_t, \hat{\theta}_t)$$

$$\leqq V_t\left(1 + \frac{\gamma^2 \lambda_2}{(t+1)^2}\right) + \frac{\lambda_1 \gamma^2}{(t+1)^2} - \frac{2\gamma c}{(t+1)} \sum_{k \in S_{x_t}} (\hat{\theta}_t^k - \theta^{0k})^2.$$

Hence, $\{V_t, \mathcal{F}_t\}$ is "nearly a positive supermartingale" in the sense of Neveu [72] or Robbins and Siegmund [73]. Applying the convergence theorem for such, we deduce that $\{V_t\}$ converges a.s. and $\sum_t 1/(t+1) \sum_{k \in S_{x_t}} (\hat{\theta}_t^k - \theta^{0k})^2 < \infty$ a.s. By positive recurrence of $\{x_t = x\}$ for every x, [68] deduces that $\liminf \sum_{k \in S_x} (\hat{\theta}_t^k - \theta^{0k})^2 = 0$ a.s. for every $x \in X$. Then $\liminf \|\hat{\theta}_t - \theta^0\|^2 = 0$ a.s. also follows, and since $\{V_t\}$ converges, shows that $V_t \to 0$ a.s. This completes the proof of part (viii). The proof of part (ix), being algebraically tedious is omitted; the reader being referred to [68]. □

It is clear that (viii) is a strong result which shows that under the conditions (i)-(iv) one can identify θ^0 under *any* policy. Thus, it is clear that even though the assumptions (iii) and (iv) are not identical to Mandl's condition [55], they are nevertheless *restrictive.* (iii) is the "pseudo-gradient" condition of Polyak and Tsypkin [74] and

guarantees that the *expected value* of the step $\hat{\theta}_{t+1} - \hat{\theta}_t$ forms an "acute angle" with the *desired* direction $\theta^0 - \hat{\theta}_t$. The strict negativity for $A = S_x$ in (iii) appears therefore to play a role analogous to that of an Identifiability Condition.

The proof used here is remarkably similar to the proof of Goodwin, Ramadge and Caines [75] in their treatment of their version of the self-tuning regulator. This only serves to illustrate how closely connected all these problems are.

7.3. The cost biased maximum likelihood method. This approach requires no restrictive identifiability condition of any sort and is motivated by the following argument. Without the imposition of identifiability conditions of any sort, many reasonable parameter identification schemes will provide a limit θ^* of $\{\hat{\theta}_t\}$ which satisfies $p(i, j, \phi(i, \theta^*); \theta^*) = p(i, j, \phi(i, \theta^*); \theta^0)$ (Theorem 7.1 for example). This has an immediate consequence, to see which we define

$\pi(i, \phi, \theta) :=$ steady state probability of i when control law ϕ is used in θ (for simplicity we assume $p(i, j, u; \theta) > 0$);

$J(\phi, \theta) :=$ long-term average cost of using ϕ in θ;

$\phi_\theta :=$ optimal control law for θ;

$J(\theta) :=$ optimal long-term average cost for θ.

Now,

$$\begin{aligned} J(\theta^*) &= J(\phi_{\theta^*}, \theta^*) \\ &= \sum_{i,j} \pi(i, \phi_{\theta^*}, \theta^*) p(i, j, \phi(i, \theta^*), \theta^*) c(i, \phi(i, \theta^*)) \\ &= \sum_{i,j} \pi(i, \phi_{\theta^*}, \theta^0) p(i, j, \phi(i, \theta^*), \theta^0) c(i, \phi(i, \theta^*)) \\ &= J(\phi_{\theta^*}, \theta^0) \\ &\geqq J(\theta^0). \end{aligned}$$

Here we have used the well-known, see [33], representation of the long-term average cost in terms of the steady state probabilities *and* the implication $\{p(i, j, \phi(i, \theta^*); \theta^*) = p(i, j, \phi(i, \theta^*); \theta^0)$ for all $i, j\} \Rightarrow \{\pi(i, \phi_{\theta^*}, \theta^*) = \pi(i, \phi_{\theta^*}, \theta^0)$ for all $i\}$. So $\hat{\theta}_t \to \{\theta: J(\theta) \geqq J(\theta^0)\}$. To capitalize on this, one can "bias" the parameter identification scheme in "favour" of parameters θ for which $J(\theta)$ is small, knowing of course that θ^0 has the smallest value $J(\theta^0)$ in $\{\theta: J(\theta) \geqq J(\theta^0)\}$. However, this biasing has to be done delicately, so that we do *not* destroy the parameter identification scheme itself. This is done below.

THEOREM 7.5 (Kumar and Becker). *Let*

i) $p(i, j, u; \theta) > 0$ *for all* i, j, u, θ;

ii) $c(i, u) > 0$ *for all* i, u;

iii) Θ *be finite*;

iv) $o(t)$ *be such that* $\lim o(t) = +\infty$ *and* $\lim t^{-1} o(t) = 0$. *Choose* $\hat{\theta}_t$ *and* u_t *so that*

$$\hat{\theta}_t = \begin{cases} \arg\max_\theta J(\theta)^{-o(t)} \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \theta) & \text{for } t = 0, 2, 4, 6, \cdots, \\ \hat{\theta}_{t-1} & \text{for } t = 1, 3, 5, \cdots \end{cases}$$

and $u_t = \phi(x_t, \hat{\theta}_t)$. (*If there is more than one maximizer above, choose one according to some prespecified priority order on* Θ.) *Then*:

v) $\lim (1/N) \sum_1^N c(x_t, u_t) = J(\theta^0)$ *a.s.*

vi) $\lim (1/N)\sum_1^N 1(\hat{\theta}_t = \theta^*) = 1$ *for some* θ^* *a.s.*

vii) $p(i, j, \phi(i, \theta^*), \theta^*) = p(i, j, \phi(i, \theta^*), \theta^0)$ *a.s.*

viii) ϕ_{θ^*} *is optimal for* θ^0 *a.s.*

Proof. Since $\hat{\theta}_t$ is a maximizer of the criterion according to which it is chosen,

$$J(\hat{\theta}_t)^{-o(t)} \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \hat{\theta}_t) \geqq J(\theta^0)^{-o(t)} \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \theta^0).$$

Taking logarithms,

$$\frac{o(t)}{t} \log\left(\frac{J(\theta^0)}{J(\hat{\theta}_t)}\right) + \frac{1}{t}\sum_{s=0}^{t-1} \log \frac{p(x_s, x_{s+1}, u_s; \hat{\theta}_t)}{p(x_s, x_{s+1}, u_s; \theta^0)} \geqq 0.$$

Hence, if $\hat{\theta}_t = \theta^*$ i.o., for some $\omega \in \Omega$ the sample space, it follows that

$$\liminf \frac{1}{N} \sum_0^{N-1} \log \frac{p(x_s, x_{s+1}, u_s; \theta^*)}{p(x_s, x_{s+1}, u_s; \theta^0)} \geqq 0.$$

By the Martingale Stability Theorem, it follows that for every $\theta \in \Theta$,

$$\lim \frac{1}{t} \sum_0^{t-1} \frac{p(x_s, x_{s+1}, u_s; \theta)}{p(x_s, x_{s+1}, u_s; \theta^0)} = 1 \quad \text{a.s.}$$

Putting these two facts together and making use of the positive recurrence of each event $(x_t = i)$, it follows that a.s. $p(i, j, \phi(i, \theta^*); \theta^*) = p(i, j, \phi(i, \theta^*); \theta^0)$ whenever $\limsup (1/N)\sum_0^{N-1} 1(\hat{\theta}_t = \theta^*) > 0$. On the other hand,

$$\left(\frac{J(\theta^0)}{J(\hat{\theta}_t)}\right)^{o(t)} \prod_{s=0}^{t-1} \frac{p(x_s, x_{s+1}, u_s; \hat{\theta}_t)}{p(x_s, x_{s+1}, u_s; \theta^0)} \geqq 1.$$

Since we know that

$$\left\{\prod_0^{t-1} \frac{p(x_s, x_{s+1}, u_s; \theta)}{p(x_s, x_{s+1}, u_s; \theta^0)}\right\}$$

is a positive martingale and therefore converges a.s., we see that if $\hat{\theta}_t = \theta^*$ i.o., then $J(\theta^0) \geqq J(\theta^*)$. Together with the argument preceding this theorem, this shows that $\limsup (1/N)\sum_0^{N-1} 1(\hat{\theta}_t = \theta^*) > 0$ implies $\phi(\cdot, \theta^*)$ is optimal for θ^0. The argument in Theorem 7.2 can now be used to show (v). (vi), (vii) and (viii) will follow if we can show that a.s. θ^* is unique. This is a consequence of the priority ordering of elements of Θ. The full details are in [51]. □

The unique feature of the adaptive control scheme given here is that without resorting to either forced choices or randomization, it can attain optimal performance. Condition (ii) of the theorem is unnecessary and condition (i) can be replaced by a positive recurrence condition, see [51].

In [76] the above result has been generalized to the situation where Θ is not finite, but is the class of *all* models. In [77], the generalization to the case where X and U are Borel spaces is developed. The case of linear systems with quadratic cost criteria is analyzed in [78] when the parameter set Θ is finite. In this situation, an additional complexity is to prove that the system is stable as well, and this is also done in [78].

If Θ is not finite, then implementation of schemes of this type will depend on the availability of methods to perform the requisite on-line computations. In the general case such methods are not yet available. In some specific situations however, see [79], explicit solutions can be obtained.

The above approach can also be used for discounted (or other, for example, finite horizon) cost problems to obtain adaptive policies which converge in a Cesaro sense to the optimal policy. This is done in [77].

Before ending the topic of this section, we note the early work of Riordan [80] who has also simulated the control of a heat treatment process by modeling it as a problem in the adaptive control of Markov chains.

8. Self-tuning regulators. This class of non-Bayesian adaptive control problems (NACP's) has recently enjoyed much practical success. See [81]–[85] for a sampling of some of the applications culled from the last two years' issues of one journal.

The basic problem can be posed as follows. There is a system

$$y(t+1)=\sum_{i=0}^{n} a_i y(t-i)+\sum_{i=0}^{p} b_i u(t-i-d+1)+\sum_{i=0}^{m} c_i w(t-i)+w(t+1).$$

Here u is the input, y the output and $\{w(t)\}$ is a "white" noise sequence (defined more precisely later on). $d \geqq 1$ is called the *delay*, and is assumed to be known. The problem is that the coefficients $\{a_i, b_i, c_i\}$ of the linear system are unknown. In spite of this, the goal is to choose $u(t)$ based on $(y(0), u(0), y(1), u(1), \cdots, y(t))$, for each t, so that $\lim (1/N)\sum_1^N y^2(t)$—the sample path variance of the output—is almost surely a minimum. As is clear, this is a standard NACP of the sort defined in § 6.

To obtain a full appreciation of this problem, however, it is necessary first to understand the complexities involved in a) *identifying* the coefficients of an Auto Regressive Moving Average System with EXogeneous Inputs (ARMAX system) and b) the problem of *controlling* such processes. Accordingly, we first take up these two issues.

8.1. Least squares estimation of coefficients. We shall *very* briefly explain the various issues involved in identifying the coefficients of an ARMAX model.

8.1.1. Suppose that we have available a finite sequence $\{y(-n), y(-n+1), \cdots, y(-1)\} \cup \{y(0), y(1), \cdots, y(N)\}$ and we wish to model this sequence by a relationship of the form $y(t+1) \approx \alpha_0 y(t)+\alpha_1 y(t-1)+\cdots+\alpha_n y(t-n)$ for $t=0, 1, 2, \cdots, N-1$. The question is: What are the "best" coefficients $(\alpha_0, \alpha_1, \cdots, \alpha_n)$ to choose? One way, of course, is to select them so that they minimize $\sum_1^N (y(t)-\alpha_0 y(t-1)-\alpha_1 y(t-2)-\cdots-\alpha_n y(t-n-1))^2$. This is a standard "curve fitting" sort of procedure which seeks to minimize the sum of the squares of the errors of "fit".

Note that this method of fitting a model to data is *totally nonprobabilistic.*

One way of solving the minimization problem is to define $\phi_{t-1} := (y_{t-1}, y_{t-2}, \cdots, y_{t-n-1})^T$ and $\theta := (\alpha_0, \alpha_1, \cdots, \alpha_n)^T$. (We use both $y(t)$ and y_t interchangeably.) Then the goal is to choose $\theta \in \mathbb{R}^{n+1}$ to minimize $\|(y_1, y_2, \cdots, y_N)-\theta^T(\phi_0, \phi_1, \cdots, \phi_{N-1})\|^2$. A standard application of the projection theorem shows that the minimizing θ, denoted by $\hat{\theta}_N$, is given by

$$\begin{aligned}\hat{\theta}_N &= ((\phi_0, \phi_1, \cdots, \phi_{N-1})(\phi_0, \phi_1, \cdots, \phi_{N-1})^T)^{-1} \\ &\quad \cdot (\phi_0, \phi_1, \cdots, \phi_{N-1})(y_1, y_2, \cdots, y_N)^T \\ &= \left(\sum_0^{N-1} \phi_i \phi_i^T\right)^{-1} \sum_0^{N-1} \phi_i y_{i+1} \\ &= \left(\frac{1}{N}\sum_0^{N-1} \phi_i \phi_i^T\right)^{-1} \left(\frac{1}{N}\sum_0^{N-1} \phi_i y_{i+1}\right).\end{aligned}$$

This will be called the "least squares estimate".

8.1.2. Suppose one more piece of data y_{N+1} becomes available; then one can obtain another estimate $\hat{\theta}_{N+1}$ which minimizes $\sum_1^{N+1} (y_t-\alpha_0 y_{t-1}-\cdots-\alpha_n y_{t-n-1})^2$. The

relationship between $\hat{\theta}_{N+1}$ and $\hat{\theta}_N$ can be written in a recursive manner as

$$\hat{\theta}_{N+1} = \hat{\theta}_N + R_N^{-1}\phi_N(y_{N+1} - \hat{\theta}_N^T\phi_N)$$

where $R_N := \sum_{i=0}^{N} \phi_i\phi_i^T$.

One can also obtain a recursive expression for R_N^{-1} as

$$R_{N+1}^{-1} = R_N^{-1} - \frac{R_N^{-1}\phi_{N+1}\phi_{N+1}^T R_N^{-1}}{1 + \phi_{N+1}^T R_N^{-1}\phi_{N+1}}.$$

This is done by using the well-known Matrix Inversion Lemma.

8.1.3. So far we have only examined the question of fitting a model to data, without mentioning in *any way whatsoever*, how the data were generated in the first place. Let us suppose that the sequence $\{y_t\}$ is actually generated by the *system*

$$y_{t+1} = a_0 y_t + a_1 y_{t-1} + \cdots + a_n y_{t-n} + v_t$$

where $\{v_t\}$ is "white" noise.

By employing the ergodic theorem (assuming all roots of the polynomial $1 - a_0 z - a_1 z^2 - \cdots - a_n z^{n+1}$ are strictly outside the unit circle) it can be seen that

$$\left(\lim \frac{1}{N} \sum_{i=0}^{N-1} \phi_i\phi_i^T\right)_{jk} = r_{|j-k|} \quad \text{where } r_j = E(y_t y_{t-j}) \text{ in steady state.}$$

Also

$$\lim \frac{1}{N} \sum_{0}^{N-1} \phi_i y_{i+1} = (r_1, r_2, \cdots, r_{n+1})^T.$$

Now we can examine the *probabilistic* behaviour of the sequence $\{\hat{\theta}_N\}$. From the above it follows that

$$\lim \hat{\theta}_N = [r_{|j-k|}]^{-1}(r_1, r_2, \cdots, r_{n+1})^T.$$

Since $y_{t-j}y_{t+1} = \sum_{i=0}^{n} a_i y_{t-j}y_{t-i} + y_{t-j}v_{t+1}$, taking expectations, gives $r_{j+1} = \sum_{i=0}^{n} a_i r_{j-i}$ for $j \geqq 0$. Hence $(r_1, \cdots, r_{n+1})^T = [r_{|j-k|}](a_0, \cdots, a_n)^T$ and so $(a_0, \cdots, a_n)^T = [r_{|j-k|}]^{-1}(r_1, \cdots, r_{n+1})^T$. This shows that $\lim \hat{\theta}_N = (a_0, a_1, \cdots, a_n)^T$ a.s.

To summarize, if the noise entering into the system (here v_t) is "white", then least squares estimates of the parameters $\{\hat{\theta}_N\}$ are consistent.

For a general treatment of this, the reader is referred to Lai and Wei [86].

8.1.4. Crucial use was made in the above of the fact that $\{v_t\}$ was a "white" noise. Suppose now that it is *not* white; then, in general, $Ey_{t-j}v_{t+1} \neq 0$ for $j \geqq 0$. It then follows, that $\hat{\theta}_N \to (a_0, a_1, \cdots, a_n)^T + \Delta$ where $\Delta \neq 0$. See Goodwin and Payne [87] for an example.

Now if $v_{t+1} = w_{t+1} + \sum_{i=0}^{m} c_i w_{t-i}$ where $\{w_t\}$ is "white", then $\{v_t\}$ is a moving average of white noise and is *not* itself white. Thus if one uses the least squares procedure to obtain estimates $\{\hat{\theta}_N\}$ of the coefficients $(a_0, a_1, \cdots, a_n)$ in the system

$$y(t+1) = \sum_{i=0}^{n} a_i y(t-i) + \sum_{i=0}^{m} c_i w(t-i) + w(t+1),$$

then the least squares estimates are asymptotically *biased.*

8.1.5. For a system of the type just mentioned one might also be interested in estimating the coefficients $(c_0, c_1, \cdots, c_m)$ in *addition* to the coefficients $(a_0, a_1, \cdots, a_n)$.

If one had access to $(y_0, y_1, \cdots, y_N)$ *and* $(w_0, w_1, \cdots, w_N)$, then one could form asymptotically (as $N \to \infty$) consistent estimates $(\alpha_0, \cdots, \alpha_n, \gamma_0, \cdots, \gamma_m)$ by minimizing

$$\sum_{t=1}^{N} (y_t - \alpha_0 y_{t-1} - \cdots - \alpha_n y_{t-n-1} - \gamma_0 w_{t-1} - \gamma_1 w_{t-2} - \cdots - \gamma_m w_{t-m-1})^2.$$

However $\{w_t\}$ is *not* generally available to observe. It is merely an innovations *representation* of the coloured noise $v_t = w_t + c_0 w_{t-1} + \cdots + c_m w_{t-m-1}$ in the system. But one could *estimate* w_t by $y_t - \phi_{t-1}\hat{\theta}_{t-1}$ or $y_t - \phi_{t-1}^T \hat{\theta}_t$, and use these in place of the true values of w_t.

A scheme of this *general* sort is examined by Solo [88]. It is shown that a sufficient condition to obtain strongly consistent estimates is that the polynomial $C(z) = 1 + c_0 z + \cdots + c_m z^{m+1}$ satisfies the condition

$$\operatorname{Re}\left[\frac{1}{C(e^{i\omega})} - \frac{1}{2}\right] > 0 \quad \text{for all } \omega.$$

(Here $i = \sqrt{-1}$.) Note that since $v_{t+1} = w_{t+1} + c_0 w_t + \cdots + c_m w_{t-m}$ is just a *representation* of $\{v_t\}$, we can assume without loss of generality that all roots of $C(z)$ are outside the unit circle, see Åström [89].

Conditions of the type $\operatorname{Re}(C e^{i\omega}) > \varepsilon$, called Positive Real Conditions, occur in other analyses of identification algorithms also, see Solo [148]. Proofs, both in identification and adaptive control, as we shall see in §§ 8.4 and 8.5, have been built around such conditions. To this author, however, the full extent of their role is not completely clear: some insights, however, are offered by Ljung [94]. It should be noted that the Positive Real Condition also plays a role in deterministic adaptive control, where it is used via the concept of "hyperstability".

8.1.6. Suppose that we also have control inputs $\{u_t\}$ in the system, i.e.,

$$y(t+1) = \sum_{i=0}^{n} a_i y(t-i) + \sum_{i=0}^{p} b_i u(t-d-i+1) + \sum_{i=0}^{m} c_i w(t-i) + w(t+1).$$

In such a case, besides estimating $\{a_i, c_i\}$ we may also wish to estimate $\{b_i\}$.

Clearly, if $u_t = 0$ for all t, then one could not possibly identify $\{b_0, \cdots, b_p\}$. Thus, some "excitation" conditions have to be imposed on $\{u_t\}$ in order to guarantee asymptotic consistency of the estimates of $\{a_0, \cdots, a_n, b_0, \cdots, b_p, c_0, \cdots, c_m\}$. These *sufficiency* conditions are usually of the form

$$\lim \frac{1}{N} \sum_{1}^{N} \phi(t)\phi^T(t) \quad \text{is positive define}$$

where $\phi(t) := (y(t), \cdots, y(t-n), u(t-d+1), \cdots, u(t-d-p+1), w(t), \cdots, w(t-m))^T$ and are called "persistency of excitation" conditions, see Åström [90] and Solo [88].

8.2. Minimum variance control of an ARMAX model. We now examine the problem of *control* of an ARMAX model, when the coefficients $\{a_i, b_i, c_i\}$ are *known*. We shall proceed by examining a sequence of special cases until we obtain full generality. Only the fully general case is proved; for the special cases we provide informal arguments which are quick and easy to understand.

8.2.1. For the system

$$y_{t+1} = a_0 y_t + \cdots + a_n y_{t-n} + b_0 u_t + \cdots + b_p u_{t-p} + w_{t+1}$$

where $\{w_t\}$ is "white" noise, it is clear that the control law

$$u_t = \frac{-1}{b_0}(a_0 y_t + \cdots + a_n y_{t-n} + b_1 u_{t-1} + \cdots + b_p u_{t-p})$$

minimizes the sample path variance $\lim (1/N) \sum_1^N y^2(t)$. For then, $y_{t+1} = w_{t+1}$ and so

$$\lim \frac{1}{N} \sum_1^N y_t^2 = \lim \frac{1}{N} \sum_1^N w_t^2 = \sigma^2 \quad \text{a.s.}$$

the best achievable, where $\sigma^2 := E(w_t^2)$.

8.2.2. One special feature of the above is that the noise is "white". Suppose now that the system is

$$y(t+1) = a_0 y_t + \cdots + a_n y_{t-n} + b_0 u_t + \cdots + b_p u_{t-p} + w_{t+1} + c_0 w_t + \cdots + c_n w_{t-n}.$$

If we *could* use $u_t = (-1/b_0)(a_0 y_t + \cdots + a_n y_{t-n} + b_1 u_{t-1} + \cdots + b_p u_{t-p} + c_0 w_t + \cdots + c_n w_{t-n})$, then this is clearly the best achievable, since then $y_{t+1} = w_{t+1}$ and so $\lim (1/N) \sum_1^N y_t^2 = \sigma^2$ a.s. However $\{w_t\}$ is *not* accessible, and so we replace it by $\{y_t\}$, which is what it *would* be if the above control law could be implemented. This gives

$$u_t = \frac{-1}{b_0}[(a_0 + c_0) y_t + \cdots + (a_n + c_n) y_{t-n} + b_1 u_{t-1} + \cdots + b_p u_{t-p}]$$

the optimal control law.

It is important to note that in order to implement this control law, one needs knowledge only of $\{a_i + c_i, b_i\}$ and *not* $\{a_i, c_i, b_i\}$ separately.

8.2.3. The special feature of the above model is that the delay d is *exactly* one unit. Consider now the general case $d \geqq 1$. The system is

$$y_{t+1} = a_0 y_t + \cdots + a_n y_{t-n} + b_0 u_{t-d+1} + \cdots + b_p u_{t-d-p+1} + w_{t+1} + c_0 w_t + \cdots + c_n w_{t-n}.$$

This is more conveniently represented in the "polynomial" format

$$A(z) y_{t+1} = z^d B(z) u_{t+1} + C(z) w_{t+1}$$

where $A(z) := 1 - a_0 z - \cdots - a_n z^{n+1}$; $B(z) = b_0 + b_1 z + \cdots + b_p z^p$; $C(z) = 1 + c_0 z + \cdots + c_n z^{n+1}$, and z is the *backward shift operator* $z y_t := y_{t-1}$. (Note that often in the literature z^{-1} is the backward shift operator; but we find this more convenient.) Let us *divide* the polynomial C by the polynomial A, carrying out d steps of the long division process to get

$$C = AF + z^d G$$

where $F(z) = 1 + f_1 z + \cdots + f_{d-1} z^{d-1}$ is the "quotient" and $G(z) = g_0 + g_1 z + \cdots + g_n z^n$. Now we can formally represent the system as

$$y_{t+1} = z^d \frac{B}{A} u_{t+1} + \frac{C}{A} w_{t+1} = z^d \left(\frac{B}{A} u_{t+1} + \frac{G}{A} w_{t+1} \right) + F w_{t+1}.$$

Clearly u_{t+1-d} cannot be based on the "future" $(w_{t+1}, \cdots, w_{t+2-d})$. Thus one chooses $u_{t+1} = (-G/B) w_{t+1}$. This results in $y_{t+1} = F(z) w_{t+1}$, thus giving us the optimal control

law

$$u_t = \frac{-1}{BF} y_t.$$

Further, the minimum output variance is $\lim (1/N) \sum_1^N y_t^2 = (1 + f_1^2 + \cdots + f_{d-1}^2)\sigma^2$. For more details, the reader is referred to Åström [89].

8.2.4. There is one potential "practical" flaw with this procedure. To see the nature of this, consider the very simple example

$$y_{t+1} = -y_t + u_t - 2u_{t-1} + w_{t+1}.$$

Applying the procedure of § 8.2.1, which is but a special case of § 8.2.3, shows that the optimal control law is

$$u_t = y_t + 2u_{t-1}.$$

Under this optimal control law, $y_{t+1} = w_{t+1}$ and so the sequence of applied controls is $u_t = 2u_{t-1} + w_t$. This is clearly an *unstable* difference equation, in view of the coefficient 2. Thus, although theoretically this is an optimal control law, from a practical point of view this is *unacceptable.* One reason is that if the true system is actually $y_{t+1} = -y_t + u_t - (2+\varepsilon)u_{t-1} + w_{t+1}$ where ε is very small, then after the control law $u_t = 2u_{t-1} + w_t$ is applied, the closed loop system is $y_{t+1} = \varepsilon u_{t-1} + w_{t+1}$ which with $u_t = 2u_{t-1} + y_t$ is an *unstable* system. In any case, u_t is an "exploding" sequence.

The source of this instability is that the polynomial $B(z) = 1 - 2z$ does *not* have all its roots strictly outside the unit circle. This added requirement that the polynomial $B(z)$ have all its roots strictly outside the unit circle is called a *minimum phase condition,* and should necessarily be imposed if the *optimal* control law of § 8.2.3 is to be useful *practically.*

8.2.5. Thus, we see that in general the true practical problem at hand is *not* to minimize $\lim (1/N) \sum_1^N y_t^2$ *unconditionally,* but to minimize it *conditionally* subject to the *constraint* that the closed loop system is stable.

THEOREM 8.1 (Peterka). *Let* i) $C(z)$ *have no roots on or inside the unit circle*;
ii) $b_p \neq 0$.

Then the control law which minimizes the output variance subject to the stability requirement above is

$$u_t = -\frac{S(z)}{R(z)} y_t$$

where $S(z)$ *and* $R(z)$ *are polynomials determined by the polynomial equation*

$$B^* C = RA + z^d BS$$

with $\deg R = p + d - 1$. *Here*

$B^+ :=$ *factor of* B *containing all roots of* B *lying outside or on the unit circle normalized so that* $B^+(0) = 1$;

$B^- :=$ *factor of* B *containing all roots of* B *lying inside the unit circle and satisfying* $B^+ B^- = B$;

$$\tilde{B}^- := \frac{1}{b_p} z^p B^-(z^{-1});$$

$$B^* := B^+ \tilde{B}^-.$$

Proof. For any polynomial P, define P^+, P^-, P^* as above and also $\tilde{P} := (1/p_k)z^k P(z^{-1})$, $\bar{P} := P(z^{-1})$. Note that a general linear system can be represented as $y_t = z^d(b(z)/a(z))u_t + (\beta(z)/\alpha(z))w_t$ where $\beta = \beta^+$, $\alpha = \alpha^+$. Now this can be rewritten as $a\alpha y_t = z^d b\alpha u_t + a\beta w_t$, which in turn is the same as $a\alpha y_t = z^d b\alpha u_t + a^*\beta w_t$. So we shall assume for our purposes that $A = a\alpha$, $B = b\alpha$ and $C = a^*\beta$ where $\alpha = \alpha^+$, $\beta = \beta^+$, $\alpha(0) = \beta(0) = a(0) = 1$. Now if a control law $u_t = -(s/r)y_t$ is applied, then $\text{var}\,(y_t) = \sigma^2(1/2\pi i)\oint W(z)W(z^{-1})\,dz/z$ where $W = (\beta/\alpha)ar/(ar + z^d bs)$. One can rewrite W as

$$W = \frac{z^d a^- b^-}{\tilde{a}^-\tilde{b}^-}\left(\frac{\tilde{a}^-\tilde{b}^-\beta}{z^d a^- b^-\alpha} - \frac{\tilde{a}^-\beta b^* s}{a^-\alpha(ar + z^d bs)}\right).$$

A partial fraction expansion gives,

$$\frac{\tilde{a}^-\tilde{b}^-\beta}{z^d a^- b^-\alpha} = \frac{p}{z^d b^-} + \frac{q}{a^-\alpha}$$

where $\deg p = (\deg b^-) + d - 1$, $p(0) = 1$. This gives the polynomial equation $\tilde{a}^-\tilde{b}^-\beta = a^-\alpha p + z^d b^- q$. Now W can be rewritten as

$$W = \frac{z^d a^- b^-}{\tilde{a}^-\tilde{b}^-}\left(\frac{p}{z^d b^-} + \psi\right)$$

which defines ψ. ψ can be rewritten as

$$\psi = \frac{a^+ qr - \alpha p b^+ s}{\alpha(ar + z^d bs)}.$$

Now we see that

$$\text{var}\,(y_t) = \text{const}\left[\frac{1}{2\pi i}\oint \frac{p\bar{p}}{b^-\bar{b}^-}\frac{dz}{z} + \frac{1}{2\pi i}\oint \psi\bar{\psi}\frac{dz}{z} + \frac{2}{2\pi i}\oint \frac{\bar{p}}{z^{-d}\bar{b}^-}\psi\frac{dz}{z}\right]$$

where $\text{const} := (a^- b^-/\tilde{a}^-\tilde{b}^-)(\bar{a}^-\bar{b}^-/\bar{\tilde{a}}^-\bar{\tilde{b}}^-)$. The point of all this algebraic manipulation is that, for stability, the denominator of ψ must have its roots strictly outside the unit circle, i.e., ψ must be holomorphic inside and on the unit circle. This makes the last integral in the above expression zero. The first integral is unaffected by the choice of the control polynomials r and s. Thus, to minimize $\text{var}\,(y_t)$ subject to the stability constraint, the best one can do is to make $\psi = 0$, which happens when $s/r = a^+ q/\alpha p b^+$. This can be further simplified if one notes that by multiplying the polynomial equation determining p and q by $a^+ b^+$, we get $a^* b^*\beta = a\alpha p b^+ + z^d b a^+ q$. After defining $v = pb^+$, this simplifies to $a^* p^*\beta = a\alpha v + z^d bs$ and $r = \alpha v$. Multiplying by α, and using the original notation gives $CB^* = Ar + z^d Bs$, and the optimal control law is as claimed. □

Recently, [138] has also obtained the stable control laws which minimize the variance for general multivariable ARMAX systems.

8.3. The self-tuning regulator. We now take up for consideration the self-tuning regulator. Given a finite sequence of data $\{y_0, u_0, y_1, u_1, \cdots, y_N\}$ we shall first fit a model of the form

$$y_{t+1} \approx \alpha_0 y_{t-d+1} + \cdots + \alpha_l y_{t-l} + \beta_0 u_{t-d+1} + \cdots + \beta_m u_{t-d-m+1} \quad \text{for some } d \geqq 1$$

by minimizing $\sum_{t=0}^{N-1}(y_{t+1} - \alpha_0 y_{t-d+1} - \cdots - \alpha_l y_{t-l-d+1} - \beta_0 u_{t-d+1} - \cdots - \beta_m u_{t-d-m+1})^2$ over $(\alpha_0, \cdots, \alpha_l, \beta_0, \cdots, \beta_m)^T \in \mathbb{R}^{l+m+2}$. Denoting the minimizer by $\hat{\theta}_N$, we know that it can be written in the recursive form

$$\hat{\theta}_N = \hat{\theta}_{N-1} + R_{N-1}^{-1}\phi_{N-1}(y_N - \hat{\theta}_{N-1}^T\phi_{N-d})$$

where $R_N = R_{N-1} + \phi_N \phi_N^T$ and $\phi_N := (-y_N, -y_{N-1}, \cdots, -y_{N-l}, u_N, \cdots, u_{N-m})^T$. (The scheme of Åström and Wittenmark [92] fixes β_0 arbitrarily and minimizes the least-squares criterion subject to this constraint.) Then the control input u_N is chosen so that

$$u_N = \frac{1}{\hat{\beta}_0}[\hat{\alpha}_0 y_N + \cdots + \hat{\alpha}_l y_{N-l} + \hat{\beta}_1 u_{N-1} + \cdots + \hat{\beta}_m u_{N-m}]$$

where $\hat{\theta}_N =: (\hat{\alpha}_0, \cdots, \hat{\alpha}_l, \hat{\beta}_0, \cdots, \hat{\beta}_m)$. Equivalently u_N is chosen so as to make $\phi_N^T \hat{\theta}_N = 0$.

This is a recursive scheme which operates strictly off incoming data, and in "real time". From data $\{y_0, u_0, \cdots, y_N\}$ a control u_N is calculated. This control input u_N is then applied to some system; it does not matter for the present what it is. The main feature is that some output y_{N+1} is obtained from the system. This gives the enlarged data set $\{y_0, \cdots, y_N, u_N, y_{N+1}\}$ from which u_{N+1} is calculated and then applied to the system etc.

8.3.1. The above self-tuning regulator (STR) scheme can be used to control any system. The question is: For what classes of systems is the scheme asymptotically optimal? Clearly, the control law we are applying is of the type $u_t = \sum_{i=1}^{m} h_i u_{t-i} + \sum_{i=0}^{l} g_i y_{t-i}$ and so it is clear that one must only consider systems for which control laws of the above type (and with the given orders) are optimal. A natural candidate class of systems is those of the type surveyed in § 8.2.3 with the appropriate orders. Thus we consider systems which are of the form

$$y_{t+1} = a_0 y_t + \cdots + a_n y_{t-n} + b_0 u_{t-d+1} + \cdots + b_p u_{t-d-p+1} + w_{t+1} + c_0 w_t + \cdots + c_n w_{t-n}$$

where $p \leqq m - d + 1$, $n \leqq l$.

8.3.2. Åström and Wittenmark [92] were, apparently, the first to attempt an analysis of the STR scheme when it is applied to the above system.

The basic contention of [92] is that the vectors $\{\hat{\theta}_N\}$ *cannot* converge to *arbitrary* values. If $\{\hat{\theta}_N\}$ converges, it can *only* converge to values which result in a feedback control law which is *optimal* for the true system.

We now show the arguments used in [92] on an example—where the arguments used are most transparent. The general case proceeds along *exactly* the same lines.

Generic example 8.2 (Åström and Wittenmark). The true system is $y(t) - a_0 y(t-1) = b_0 u(t-2) + b_1 u(t-3) + w(t) + c_0 w(t-1)$. We choose estimates $(\hat{\alpha}(N), \hat{\beta}_1(N), \hat{\beta}_2(N))$ which minimize

$$\sum_{t=1}^{N} (y(t) - \alpha y(t-2) - \beta_0 u(t-2) - \beta_0 \beta_1 u(t-3) - \beta_0 \beta_2 u(t-4))^2$$

over all $(\alpha, \beta_1, \beta_2)$. Here β_0 is prechosen and fixed (this makes it slightly different from the scheme above). Then u_N is chosen as

$$u(N) = -\frac{\hat{\alpha}(N)}{\beta_0} y(N) - \hat{\beta}_1(N) u(N-1) - \hat{\beta}_2(N) u(N-2).$$

Now we examine the possible limits of $\{(\hat{\alpha}(N), \hat{\beta}_1(N), \hat{\beta}_2(N))\}$. If $\{(\hat{\alpha}(N), \hat{\beta}_1(N), \hat{\beta}_2(N))\}$ does converge to $(\alpha, \beta_1, \beta_2)$ (say), then asymptotically we

will have, approximately,

$$\lim \frac{1}{N}\sum_1^N \begin{pmatrix} y_i^2 & \beta_0 y_i u_{i-1} & \beta_0 y_i u_{i-2} \\ \beta_0 y_i u_{i-1} & \beta_0^2 u_{i-i}^2 & \beta_0^2 u_{i-1}u_{i-2} \\ \beta_0 y_i u_{i-2} & \beta_0^2 u_{i-1}u_{i-2} & \beta_0^2 u_{i-2}^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\approx \lim \frac{1}{N}\sum_1^N \begin{pmatrix} y_{i+2}y_i - \beta_0 u_i y_i \\ y_{i+2}u_{i-1} - \beta_0 u_i u_{i-1} \\ y_{i+2}u_{i-2} - \beta_0 u_i u_{i-2} \end{pmatrix}$$

and $u_N \approx -(\alpha/\beta_0)y_N - \beta_1 u_{N-1} - \beta_2 u_{N-2}$. Substituting in the above gives

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \approx \lim \frac{1}{N}\sum_1^N \begin{pmatrix} y_{i+2}y_i \\ y_{i+2}u_{i-1} \\ y_{i+2}u_{i-2} \end{pmatrix}.$$

From the relationship for u_N, we additionally obtain $\lim (1/N)\sum_1^N y_{i+2}u_i \approx 0$. Note that, so far, we have *not* made use of the fact that the true system is modeled as an ARMAX process.

Proceeding, we see that asymptotically the closed loop system is $Ay_t \approx z^2 B(-\mathscr{A}/\mathscr{B})y_t + Cw_t$ where $A(z) = 1 - a_0 z$, $B(z) = b_0 + b_1 z$, $C(z) = 1 + c_0 z$, $\mathscr{A}(z) = \alpha/\beta_0$, $\mathscr{B}(z) = 1 + \beta_1 z + \beta_2 z^2$. This is equivalent to $(A\mathscr{B} + z^2 B\mathscr{A})y_t \approx C\mathscr{B}w_t$. Define the process $v_t := C/(A\mathscr{B} + z^2 B\mathscr{A})w_t$. Then $y_t \approx \mathscr{B}v_t$ and $u_t \approx -(\mathscr{A}/\mathscr{B})y_t \approx -\mathscr{A}v_t$. Since $\lim (1/N)\sum_1^N u_i y_{i+j} \approx 0$ for $j = 2$, 3 and 4 it follows, through $u_t \approx (-\alpha/\beta_0)v_t$, that $\lim (1/N)\sum_1^N v_i y_{i+j} \approx 0$ for $j = 2$, 3 and 4. Now y_{i+5} is a linear combination of $(y_{i+4}, y_{i+3}, y_{i+2}, w_{i+5}, w_{i+4}, w_{i+3}, w_{i+2})$. Hence $\lim (1/N)\sum_1^N v_i y_{i+j} \approx 0$ for $j \geqq 5$ also, and so for all $j \geqq 2$. Since $y_i \approx v_i + \beta_1 v_{i-1} + \beta_2 v_{i-2}$, it follows that $\lim (1/N)\sum_1^N y_i y_{i+j} \approx 0$ for all $j \geqq 2$. Hence $\{y_i\}$ is a pure *moving average process* of order 2, i.e., $y_t \approx f_1 w_t + f_2 w_{t-1}$ Hence $C\mathscr{B}/(A\mathscr{B} + z^2 B\mathscr{A}) \approx F$ where $F(z) = f_1 + f_2 z$, i.e., $C\mathscr{B} = FA\mathscr{B} + z^2 BF\mathscr{A}$ or $C = FA + z^2 BF\mathscr{A}/\mathscr{B}$. Since C and FA are polynomials, so is $z^2 BF\mathscr{A}/\mathscr{B} =: \hat{G}$. Now $z^2 BF\mathscr{A} = \hat{G}\mathscr{B}$ implies that $G := (1/z^2)\hat{G}$ is also a polynomial. Hence $C = FA + z^2 G$, but this shows, see § 2.3, that $u_t = -(G/BF)y_t$ is an optimal control law. But $G/BF = \mathscr{A}/\mathscr{B}$ and so the limiting control law $u_t = -(\mathscr{A}/\mathscr{B})y_t$ is also optimal. □

To the author's knowledge, there is as yet no self-contained proof of convergence of the above original self-tuning regulator. Much progress, though, has been made, as we shall see in the sequel.

In § 8.5.3, we give an explanation of the above result for the case $d = 1$.

8.4. The ordinary differential equation method of analysis. Åström and Wittenmark [92] do not answer the question of *when* the parameter estimates will converge, but indicate, by simulations, that in many cases they do.

Ljung [93], [94] addresses this convergence question. In [93], [94] an ordinary differential equation (ODE) is associated with the pair of recursions

$$\hat{\theta}_{N+1} = \hat{\theta}_N + R_N^{-1}\phi_N(y_{N+1} - \hat{\theta}_N^T\phi_N),$$

$$R_{N+1} = R_N + \phi_{N+1}\phi_{N+1}^T.$$

The justification for replacing the above stochastic difference equations by deterministic ODE's is based on the following idea. Since $R_N = \sum_1^N \phi_i\phi_i^T$, we suspect that elements of R_N are $O(N)$. Thus let us consider the pair, $\Delta\hat{\theta}(N) := \hat{\theta}(N+1) - \hat{\theta}(N)$ and $\Delta(R(N)/N) := R(N+1)/(N+1) - R(N)/N$.

Simple calculations show that

$$\Delta\hat{\theta}(N) := \frac{1}{N}\bar{R}(N)^{-1}\phi_N(y_{N+1} - \hat{\theta}_N^T\phi_N),$$

$$\Delta\bar{R}(N) = \frac{1}{N+1}(\phi_{N+1}\phi_{N+1}^T - \bar{R}(N))$$

where $\bar{R}(N) := R(N)/N$ and $\Delta\bar{R}(N) := \bar{R}(N+1) - \bar{R}(N)$. Note that because of the presence of the multiplier $1/N$, $\hat{\theta}(N)$ changes very slowly for large N. Thus changes in the *control law* take place very slowly. Over long periods of time, it may be reasonable to expect that the control law applied is approximately constant. In that case, perhaps $\phi_N y_{N+1}$ and $\phi_{N+1}\phi_{N+1}^T$ can be replaced by their ergodic limits. Thus, the following ODE's appear to be a plausible representation of the asymptotic behaviour of the sample paths of the above stochastic difference equations:

$$\frac{d\theta(\tau)}{d\tau} = \bar{R}^{-1}(\tau)E_{\theta(\tau)}(\phi y), \qquad \frac{d\bar{R}(\tau)}{d\tau} = E_{\theta(\tau)}(\phi\phi^T) - \bar{R}(\tau)$$

where we have also used the time rescaling $\tau = \log t$ to eliminate the factor $1/t$. Here $E_\theta(\phi y)$ is the expected value, assuming stationarity, of $\phi_t y_{t+1}$ and $E_\theta(\phi\phi^T) = E(\phi_t\phi_t^T)$, when the fixed control law corresponding to an estimate θ is used.

The exact technical justification for the above procedure is, at the least, very complex, and in any case, for the particular problem at hand, some boundedness conditions etc., have to be either assumed or proved by other methods.

However, this should not obscure the *great* value of the above ODE's. They were initially obtained to help in the study of a problem of great complexity and have provided the essential breakthroughs which have permitted further development of the field. At the moment these ODE's are an irreplaceable tool for problems in which there is no theory. An analysis of the ODE's suggests what one may expect without the need for extensive simulations. Moreover, because of the time scaling $\tau = \log t$, phenomena which in simulation would be observed for very large values of t would occur in the ODE solutions for modest values of τ. Ljung [93], [94] presents many convincing examples of the usefulness of the ODE's.

Kushner [95] also addresses the problem of obtaining ODE's to model the behaviour of stochastic difference equations. This approach is based on the weak convergence of measures; also see Kushner and Clark [96]. For a recent, very elegant martingale approach, the reader is referred to Metivier and Priouret [139].

A more direct approach, also recent, is due to Kushner and Shwartz [149], [150]. Besides the usefulness of the invariant measure approach, a distinct advantage of [149] is that it addresses the problem of projecting the estimates into D_s, see below.

We now study the behaviour of the above ODE's for the STR.

THEOREM 8.3 (Ljung). *Let*

i) $Ay_t = zBu_t^\theta + Cw_t$ *where* $A(z) := 1 - a_0 z - a_1 z - \cdots - a_n z^{n+1}$, $B(z) := b_0 + b_1 z + \cdots + b_n$ *and* $C(z) := 1 + c_0 z + \cdots + c_n z^{n+1}$;

ii) $\phi_t := (y_t, y_{t-1}, \cdots, y_{t-m}, u_{t-1}, u_{t-2}, \cdots, u_{t-p})^T$;

iii) $u_t^\theta := -\frac{1}{b_0}\theta^T\phi_t$;

iv) $E_\theta(\phi y)$ *is the expectation, assuming a steady state, of* $\phi_{t-1}y_t$. *Thus* θ *is restricted*

to the subset $D_s \subseteq \mathbb{R}^{2n+1}$ *for which the system is strictly stable. Similarly* $E_\theta(\phi\phi^T)$ *is the expectation, in steady state, of* $\phi_t\phi_t^T$.

Suppose that

v) *Real* $(1/C(e^{i\omega})-\frac{1}{2})>0$ *for all* ω *and all roots of* $C(z)$ *are strictly outside the unit circle*;

vi) $B(z)$ *is minimum phase, i.e., all roots of* $B(z)$ *are strictly outside the unit circle*;

vii) $(a_0+c_0)+(a_1+c_1)z+\cdots+(a_n+c_n)z^n$ *and* $B(z)$ *are exactly of degree* n *and contain no common factors.*

Suppose that the ODE*'s*

$$\frac{d\theta(\tau)}{d\tau} = R(\tau)^{-1}E_{\theta(\tau)}(\phi y), \qquad \theta(0)\in D_s,$$

$$\frac{dR(\tau)}{d\tau} = E_{\theta(\tau)}(\phi\phi^T)-R(\tau), \qquad R(0)=I,$$

are such that $\theta(\tau)\in D_s$ *for all* $\tau\geqq 0$. *Define* $V(\theta, R) := (\theta-\theta_{MV})^T R(\theta-\theta_{MV})$ *where* $\theta_{MV} := (a_0+c_0, a_1+c_1, \cdots, a_n+c_n, b_1, b_2, \cdots, b_n)$. *Then*

i) θ_{MV} *is the unique equilibrium point of the first* ODE.

ii) $V(\theta(\tau), R(\tau))>0$ *whenever* $\theta(\tau)\neq\theta_{MV}$.

iii) $(d/d\tau)V(\theta(\tau), R(\tau))<0$ *whenever* $\theta(\tau)\neq\theta_{MV}$.

Proof. First we need to calculate $E_\theta(\phi y)$ and $E_\theta(\phi\phi^T)$, i.e., we need to calculate $E(\phi_{t-1}y_t)$ and $E(\phi_t\phi_t^T)$ assuming (i), (ii) and (iii) are in steady state. Let $\theta^0 := (a_0, a_1, \cdots, a_n, b_1, \cdots, b_p)^T$. Then $y(t)=\phi^T(t-1)\theta^0+b_1u(t-1)+C(z)w(t)$ and substituting from (iii) gives $y(t)=\phi^T(t-1)(\theta^0-\theta)+C(z)w(t)=\phi^T(t-1)(\theta^0-\theta_{MV})+\phi^T(t-1)(\theta_{MV}-\theta)+C(z)w(t)$. But since $\phi^T(t-1)(\theta^0-\theta_{MV})=-\sum_{i=0}^n c_iy(t-i)=(1-C(z))y(t)$, we get $C(z)y(t)=\phi^T(t-1)(\theta_{MV}-\theta)+C(z)w(t)$ or equivalently, $y(t)=\tilde{\phi}^T(t-1)(\theta_{MV}-\theta)+w(t)$ where $\tilde{\phi}(t) := C(z)^{-1}\phi(t)$. Hence $E(\phi_{t-1}y_t)=E(\phi_{t-1}\tilde{\phi}_{t-1}^T)(\theta_{MV}-\theta)$.

Now we want to show that $E(\phi_t\tilde{\phi}_t^T+\tilde{\phi}_t\phi_t^T)$ is positive definite, but this is true since $1/C(z)$ is positive real, and $\phi_t=(1/C(z))\tilde{\phi}(t)$. Similarly $E(\phi_t\tilde{\phi}_t^T+\tilde{\phi}_t\phi_t^T-\phi_t\phi_t^T)$ is positive definite because $1/C(z)-\frac{1}{2}$ is positive real.

Now we see that the ODE's can be written as

$$\frac{d\theta(\tau)}{d\tau} = R(\tau)^{-1}E_{\theta(\tau)}(\phi\tilde{\phi}^T)(\theta_{MV}-\theta(\tau)),$$

$$\frac{dR(\tau)}{d\tau} = E_{\theta(\tau)}(\phi\phi^T)-R(\tau).$$

Clearly $\theta=\theta_{MV}$ is an equilibrium point of the first ODE and now we show that it is unique. Suppose θ^* is any other equilibrium point, then $E_{\theta^*}(\phi\tilde{\phi}^T)(\theta_{MV}-\theta^*)=0$ and $(\theta_{MV}-\theta^*)^TE_{\theta^*}(\phi\tilde{\phi}^T+\phi\tilde{\phi}^T)(\theta_{MV}-\theta^*)=0$. By positive realness, again, it follows that $(\theta_{MV}-\theta^*)^T\phi_{t-1}=0$, but then both θ^* and θ_{MV} are minimum variance control laws. But by (vii), there is a unique such law, and so $\theta^*=\theta_{MV}$, showing (i). (ii) is clearly true, and by simple computation

$$\frac{d}{d\tau}V(\theta(\tau), R(\tau)) = -(\theta(\tau)-\theta_{MV})^T[E_{\theta(\tau)}(\phi\tilde{\phi}^T+\tilde{\phi}\phi^T-\phi\phi^T)+R(\tau)](\theta(\tau)-\theta_{MV})$$
$$<0. \qquad \square$$

(Note that the above theorem is associated with the situation where the parameter b_0

in the model is exactly known, and so only the other parameters are estimated by a least squares procedure.)

This theorem points to the central role played by the positive real condition on the polynomial $C(z)$ in the convergence of the STR.

To show that $\theta(\tau) \to \theta_{MV}$ as $\tau \to \infty$, one could, for example, show that

$$\beta(\|\theta(\tau)-\theta_{MV}\|) \leqq V(\theta(\tau), R(\tau)) \leqq \alpha(\|\theta(\tau)-\theta_{MV}\|) > 0$$

and

$$\frac{dV}{d\tau}(\theta(\tau), R(\tau)) \leqq -\delta(\|\theta(\tau)-\theta_{MV}\|) < 0$$

where α, β and δ are nondecreasing continuous functions such that $\alpha(0)=\beta(0)=\delta(0)=0$. This would (for example) prove the uniform asymptotic stability of θ_{MV}.

8.5. Martingale methods to exhibit asymptotic cost optimality. We now show asymptotic optimality of the incurred cost in self-tuning schemes by using martingale methods.

8.5.1. To start, we consider a slightly restricted model.

$$y(t+1) = \sum_{i=0}^{n} a_i y(t-i) + \sum_{i=0}^{n} b_i u(t-i) + w(t+1) + \sum_{i=0}^{n} c_i w(t-i)$$

where the restriction lies in the fact that we have taken the delay to be exactly 1, in contrast to the general case of § 8.3.

For this model, we use a slight modification of the STR algorithm of § 8.3. Specifically, the recursions

$$\hat{\theta}(N+1) = \hat{\theta}(N) + \frac{\gamma}{r(N)}\phi(N)[y(N+1) - \hat{\theta}^T(N)\phi(N)], \qquad \hat{\theta}(0), \quad \gamma > 0,$$

$$r(N+1) = 1 + \sum_{i=0}^{N} \phi^T(i)\phi(i)$$

where $\phi(i) := (y(i), y(i-1), \cdots, y(i-n), u(i), u(i-1), \cdots, u(i-n))^T$, are used. The difference with the scheme of § 8.3 lies in the fact that $R(N)$ has been replaced by its *trace* $r(N)$. This recursion is called the *stochastic approximation* (or stochastic gradient) algorithm; also see [71].

The controls are, as before, generated by

$$u(t) := -\frac{1}{\hat{\beta}_0}(\hat{\alpha}_0 y(t) + \cdots + \hat{\alpha}_n y(t-n) + \hat{\beta}_1 u(t-1) + \cdots + \hat{\beta}_n u(t-n))$$

where $\hat{\theta}(t) =: (\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_n, \hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_n)^T$. Or, implicitly, $u(t)$ is defined through the relation $\hat{\theta}^T(t)\phi(t) = 0$. (It will follow, under the assumptions listed below, that $\hat{\beta}_0 = 0$ is a zero probability event, and so the scheme is well defined.)

THEOREM 8.4 (Goodwin, Ramadge, Caines). *Suppose*

i) $\{w(t)\}$ *satisfies* $E(w(t+1)|\mathcal{F}_t) = 0$, $E(w^2(t+1)|\mathcal{F}_t) = \sigma^2$ *and* $E(|w(t+1)|^{2+\delta}|\mathcal{F}_t) < +\infty$ *for some* $\delta > 0$, *for all t. Here* $\mathcal{F}_t := \sigma(w(0), \cdots, w(t))$ *is the* σ*-algebra generated by the "past". The probability distribution of* $w(t)$ *is mutually absolutely continuous with respect to Lebesgue measure.*

ii) *The polynomials* $b_0 + b_1 z + \cdots + b_n z^n$ *and* $1 + c_0 z + \cdots + c_n z^{n+1}$ *have all their roots strictly outside the unit circle. Further,* $\operatorname{Re}(1 - (\gamma/2) + c_0 e^{i\omega} + \cdots + c_n e^{ni\omega + i\omega}) > 0$ *for all* ω. ($i = \sqrt{-1}$.)

Then $\lim (1/N)\sum_1^N y^2(t) = \sigma^2$ *and* $\limsup (1/N)\sum_1^N u^2(t) < +\infty$ *a.s.*

Proof. Consider the stochastic Lyapunov function, $V(t) = \|\tilde{\theta}(t)\|^2$ where $\tilde{\theta}(t) := \hat{\theta}(t) - \theta^0$ and $\theta^0 := (a_0 + c_0, \cdots, a_n + c_n, b_0, b_1, \cdots, b_n)^T$. A simple calculation shows that

$$\begin{aligned} E(V(t+1)|\mathcal{F}_t) &= V(t) + \frac{2\gamma}{r(t)}\phi^T(t)\tilde{\theta}(t)E(y(t+1)|\mathcal{F}_t) \\ &\quad + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)(E(y(t+1)|\mathcal{F}_t))^2 + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 \\ &\leqq V(t) + \frac{2\gamma}{r(t)}\phi^T(t)\tilde{\theta}(t)E(y(t+1)|\mathcal{F}_t) \\ &\quad + \frac{\gamma^2}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2 + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 \\ &= V(t) - \frac{2\gamma}{r(t)}\left\{\phi^T(t)\tilde{\theta}(t) - \frac{(\gamma+\varepsilon)}{2}E(y(t+1)|\mathcal{F}_t)\right\}E(y(t+1)|\mathcal{F}_t) \\ &\quad - \frac{\varepsilon\gamma}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2 + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2, \end{aligned}$$

where $\varepsilon > 0$ is chosen so small that the inequality in (ii) is still true when γ is replaced by $(\gamma+\varepsilon)$. Then

$$\begin{aligned} \left[C(z) - \frac{\gamma+\varepsilon}{2}\right]E(y(t+1)|\mathcal{F}_t) &= C(z)(y(t+1) - w(t+1)) - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(t+1)|\mathcal{F}_t) \\ &= y(t+1) - w(t+1) + [C(z)-1](y(t+1) - w(t+1)) - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(t+1)|\mathcal{F}_t) \\ &= \phi^T(t)\theta^0 - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(t+1)|\mathcal{F}_t) \\ &= -\phi^T(t)\tilde{\theta}(t) - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(t+1)|\mathcal{F}_t). \end{aligned}$$

The right-hand side above can thus be viewed as $E(y(t+1)|\mathcal{F}_t)$ "filtered" through the system with transfer function $[C(z) - (\gamma+\varepsilon)/2]$. Because of the property of positive real transfer functions, it follows that

$$S(t) := 2\gamma \sum_{n=1}^{t}\left\{-\phi^T(n)\tilde{\theta}(n) - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(n+1)|\mathcal{F}_n)\right\}E(y(n+1)|\mathcal{F}_n) + K \geqq 0$$

for all t, for some K, a.s. Now it follows that if $M(t) := V(t) + S(t-1)r^{-1}(t-1)$, then

$$\begin{aligned} E(M(t+1)|\mathcal{F}_t) &\leqq V(t) - \frac{1}{r(t)}(S(t) - S(t-1)) + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 \\ &\quad - \frac{\varepsilon\gamma}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2 + \frac{S(t)}{r(t)} \\ &= V(t) + \frac{S(t-1)}{r(t)} + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 - \frac{\varepsilon\gamma}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2 \\ &\leqq M(t) + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 - \frac{\varepsilon\gamma}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2. \end{aligned}$$

Thus, $\{M_t, \mathcal{F}_t\}$ is "nearly a positive supermartingale" if $\sum(\phi^T(t)\phi(t)/r^2(t)) < \infty$, see [72]. But this is true because

$$\sum_1^N \frac{\phi^T(t)\phi(t)}{r^2(t)} = \sum_1^N \frac{r(t)-r(t-1)}{r^2(t)} \leqq \sum_1^N \frac{1}{r(t-1)} - \frac{1}{r(t)} \leqq 1.$$

Hence, by [73], $\{M_t\}$ converges a.s., and furthermore,

$$\sum \frac{(E(y(t+1)|\mathcal{F}_t))^2}{r(t)} < +\infty \quad \text{a.s.}$$

If $r(t) \to +\infty$, then $(1/r(N))\sum_1^N E(y(t+1)|\mathcal{F}_t)^2 \to 0$ by Kronecker's lemma. If not, then $y^2(t) \to 0$ and $u^2(t) \to 0$ and so $C(z)w(t) \to 0$, which happens only on a null set. Due to the minimum phase condition (that all roots of $b_0 + b_1 z + \cdots + b_n z^n$ are strictly outside the unit circle) it follows that for some k_1, k_2

$$\frac{1}{N}\sum_1^N u^2(t) \leqq \frac{k_1}{N}\sum_1^N y^2(t+1) + \frac{k_2}{N}.$$

Hence

$$\frac{r(N)}{N} \leqq \frac{k_3}{N}\sum_1^N y^2(t+1) + \frac{k_4}{N}.$$

Since $E(y(t+1)|\mathcal{F}_t) = y(t+1) - w(t+1)$, and since $(1/N)\sum_1^N w^2(t+1) \to \sigma^2$, it follows that

$$\frac{r(N)}{N} \leqq \frac{k_5}{N}\sum_1^N (E(y(t+1)|\mathcal{F}_t))^2 + k_6,$$

$$\frac{1}{r(N)}\sum_1^N (E(y(t+1)|\mathcal{F}_t))^2 \geqq \frac{r(N) - k_6 N}{k_5 r(N)}.$$

Suppose $\{r(N)/N\}$ is unbounded, then along some subsequence, $(1/r(N_k))\sum_1^{N_k}(E(y(t+1)|\mathcal{F}_t))^2 \geqq 1/2k_5 > 0$ which is a contradiction, and so $\{r(N)/N\}$ is bounded, which in turn shows that $(1/N)\sum_1^N (E(y(t+1)|\mathcal{F}_t))^2 \to 0$ a.s. Now

$$\begin{aligned}\frac{1}{N}\sum_1^N y^2(t+1) &= \frac{1}{N}\sum_1^N [E(y(t+1)|\mathcal{F}_t)^2 + w^2(t+1) + 2w(t+1)E(y(t+1)|\mathcal{F}_t)] \\ &= \frac{1}{N}\sum_1^N w^2(t+1) + \frac{1}{N}\sum_1^N E(y(t+1)|\mathcal{F}_t)^2 \\ &\quad + 2\alpha(N)\left(\frac{1}{N}\sum_1^N E(y(t+1)|\mathcal{F}_t)^2\right)^{1/2}\left(\frac{1}{N}\sum_1^N w^2(t+1)\right)^{1/2}\end{aligned}$$

for some $|\alpha(N)| \leqq 1$ by the Cauchy Schwarz inequality. Taking limits gives $(1/N)\sum_1^N y^2(t) \to \sigma^2$ a.s. □

The above result shows that optimality is achieved for this NACP in the sense of § 6.2. (Actually we have used the slightly stronger condition $E|w_t|^{2+\delta} < +\infty$ to show a result slightly stronger and slightly more relevant than the conclusion $(1/N)\sum_1^N E(y^2(t+1)|\mathcal{F}_t) \to \sigma^2$ a.s. of [75].)

8.5.2. Let us call the problem of minimizing $\lim(1/N)\sum_1^N y^2(t)$, the *regulation* problem. Consider now the slightly different problem of minimizing $\lim(1/N)\sum_1^N (y(t) - y^*(t))^2$ where $\{y^*(t)\}$ is some prespecified reference trajectory

which we want the output of the system to follow. We shall call this the *tracking* problem. Note that the regulation problem is a special case of the tracking problem where one wishes to track the identically zero trajectory.

For the general tracking problem, a new twist arises. Algorithms for which proofs of asymptotic optimality exist do require estimation, in one form or another, of the coefficients $\{c_0, \cdots, c_n\}$ describing the spectrum of the noise. Hence, while in § 8.5.1, the vector $\hat{\theta}(t)$ was of dimension $(2n+2)$, now it is of dimension $(3n+3)$.

THEOREM 8.5 (Goodwin, Ramadge and Caines). *Let the algorithm be as in Theorem* 8.4, *subject only to the following changes*:

i) $\phi(t) := (y(t), y(t-1), \cdots, y(t-n), u(t), u(t-1), \cdots, u(t-n), -y^*(t), \cdots, -y^*(t-n))^T$;

ii) $\hat{\theta}(t) \in \mathbb{R}^{3n+3}$;

iii) $\{y^*(t)\}$ *is bounded*;

iv) $$u(t) = -\frac{1}{\hat{\beta}_0}(\hat{\alpha}_0 y(t) + \cdots + \hat{\alpha}_n y(t-n) + \hat{\beta}_1 u(t-1) + \cdots + \hat{\beta}_n u(t-n) - y^*(t+1) - \hat{\gamma}_0 y^*(t) - \cdots - \hat{\gamma}_n y^*(t-n))$$

or when $\hat{\theta}(t) =: (\hat{\alpha}_0, \cdots, \hat{\alpha}_n, \hat{\beta}_0, \cdots, \hat{\beta}_n, \hat{\gamma}_0, \cdots, \hat{\gamma}_n)$, $u(t)$ *is implicitly specified by* $\phi^T(t)\hat{\theta}(t) = y^*(t+1)$.

Then

$$\lim \frac{1}{N} \sum_1^N (y(t) - y^*(t))^2 = \sigma^2 \quad a.s.$$

and

$$\limsup \frac{1}{N} \sum_1^N u^2(t) < +\infty \quad a.s.$$

Proof. Similar to Theorem 8.4, or the reader may refer to [75] for explicit details. □

It should be noted that if $y^*(t) = 0$ for all t, then the last $(n+1)$ components of $\phi(\cdot)$ and $\theta(\cdot)$ make no contribution at all to the algorithm, and the vectors can therefore be collapsed, giving the same algorithm as was used in § 8.5.1 for the regulation problem.

8.5.3. How is one to understand *why* the algorithm of this section or that of the previous § 8.5.1 leads to asymptotic optimality? We proceed as follows.

Define $\hat{y}(t+d) = E(y(t+d)|\mathcal{F}_t)$ the d-step ahead prediction of $y(t+d)$. We first obtain a recursive formula for $\hat{y}(t)$ by quick formal manipulation of polynomials; these can be done rigorously by the techniques of Åström [89]. Consider the system of § 8.2.3 with $p = n$. Then

$$y(t+1) = z^d \frac{B}{A} u(t+1) + \frac{C}{A} w(t+1)$$

$$= z^d \frac{B}{A} u(t+1) + Fw(t+1) + z^d \frac{G}{A} w(t+1) \quad \text{(remembering } C = AF + z^d G\text{)}.$$

Thus the d-step ahead predictor is $\hat{y}(t+1) = y(t+1) - Fw(t+1)$. Hence

$$\hat{y}(t+1) = z^d \frac{B}{A} u(t+1) + z^d \frac{G}{A} w(t+1).$$

Since the error of prediction is $y(t+1) - \hat{y}(t+1) = Fw(t+1)$, we can substitute for

$w(t+1)$ to obtain $y(t+1)=z^d(B/A)u(t+1)+z^d(G/AF)(y(t+1)-\hat{y}(t+1))$. Multiplying through by AF, and using $C=AF+z^dG$ gives

$$C\hat{y}(t+1)=z^dBFu(t+1)+z^dGy(t+1).$$

Now consider the tracking problem if the system is known. Clearly one will choose $u(s)$ at each instant of time so that $\hat{y}(s+d)=y^*(s+d)$. Hence we will want the relation $BFu(t)+Gy(t)=Cy^*(t+d)$, i.e., one should choose $u(t)$ so that

$$\begin{aligned}u(t)=-\frac{1}{(bf)_0}[&g_0y(t)+\cdots+g_ny(t-n)+(bf)_1u(t-1)+\cdots\\&+(bf)_{n+d-1}u(t-n-d+1)-c_0y^*(t+d-1)-\cdots\\&-c_ny^*(t+d-n-1)-y^*(t+d)]\end{aligned}$$

where $B(z)F(z)=:(bf)_0+\cdots+(bf)_{n+d-1}z^n$. Moreover since $y(t+1)-y^*(t+1)=Fw(t+1)$, we can write

$$\begin{aligned}y(t+1)=&g_0y(t-d+1)+\cdots+g_ny(t-d-n+1)+(bf)_0u(t-d+1)+\cdots\\&+(bf)_{n+d-1}u(t-d-n+1)-c_0y^*(t)-\cdots-c_ny^*(t-n)+Fw(t+1).\end{aligned}$$

Now consider three cases.

Case 1. *Regulation, unit delay, coloured noise.* Here $y^*(t)=0$ for all t. So under *optimal* control, the true system looks like

$$y(t+1)=g_0(t)y(t)+\cdots+g_ny(t-n)+(bf)_0u(t)+\cdots+(bf)_nu(t-n)+w(t+1),$$

$$u(t)=-\frac{1}{(bf)_0}[g_0y(t)+\cdots+g_ny(t-n)+(bf)_1u(t-1)+\cdots+(bf)_nu(t-n)].$$

Now it is transparent that in Theorem 8.4 we are actually identifying the coefficients of the first equation, and then adopting these estimated coefficients to calculate the control law as in the second equation. Since the noise in the system is now *white*, one can expect that this scheme does work (see §§ 8.1.3 and 8.1.4). (Actually since $d=1$, we obtain $F=1$, $zG=C-A$ and $BF=B$.)

Schemes of this type are called *implicit* or direct because they identify some *closed loop* parameters (here $g_0,\cdots,g_n,(bf)_0,\cdots,(bf)_n$) and *not* the open loop parameters ($a_0,\cdots,a_n,b_0,\cdots,b_n,c_0,\cdots,c_n$). For a discussion of this point, see [75] and Fuchs [97].

Case 2. *Tracking, unit delay, coloured noise.* Here $y^*(t)$ is *not* identically zero. Hence one *cannot* neglect $y^*(\cdot)$ as we did in Case 1. So under *optimal* control, the true system can be modeled as:

$$\begin{aligned}y(t+1)=&g_0y(t)+\cdots+g_ny(t-n)+(bf)_0u(t)+\cdots\\&+(bf)_nu(t-n)-c_0y^*(t)-\cdots-c_ny^*(t-n)+w(t+1),\end{aligned}$$

$$\begin{aligned}u(t)=-\frac{1}{(bf)_0}[&g_0y(t)+\cdots+g_ny(t-n)+(bf)_1u(t-1)+\cdots\\&+(bf)_nu(t-n)-c_0y^*(t)-\cdots-c_ny^*(t-n)-y^*(t+1)].\end{aligned}$$

Again, the adaptive scheme of Theorem 8.5 *assumes* the system is in this form, and estimates the coefficients of this assumed model, and chooses a control input which is optimal for the estimated parameters. Again the noise in the system is *white* and one can expect convergence at least under some conditions, which is the content of Theorem 8.5. Again, this is an implicit or direct scheme.

Case 3. *Interlacing.* Consider the system only at d time unit intervals. Then

$$y(t+1)=\sum_{i=0}^{n} g_i y(t-d-i+1)+\sum_{i=0}^{n+d-1}(bf)_i u(t-d-i+1)$$
$$+\sum_{i=0}^{n}(-c_i)y^*(t-i)+v_{t+1},$$
$$y(t+d+1)=\sum_{i=0}^{n} g_i y(t-i+1)+\sum_{i=0}^{n+d-1}(bf)_i u(t-i+1)$$
$$+\sum_{i=0}^{n}(-c_i)y^*(t+d-i)+v_{t+d+1},$$
$$y(t+2d+1)=\sum_{i=0}^{n} g_i y(t-i+d+1)+\sum_{i=0}^{n+d-1}(bf)_i u(t+d-i+1)$$
$$+\sum_{i=0}^{n}(-c_i)y^*(t+2d-i)+v_{t+2d+1}$$
$$\vdots$$

where

$$v_t := f_0 w_t + f_1 w_{t-1} + \cdots + f_{d-1} w(t-d+1).$$

Now we see that $\{v_{t+1}, v_{t+d+1}, v_{t+2d+1}, \cdots\}$ are *independent* of each other. Hence one can estimate the parameters and even hope for asymptotic consistency. Thus one updates $\hat{\theta}(t+1)$ to $\hat{\theta}(t+d+1)$ to $\hat{\theta}(t+2d+1)\cdots$ etc. This procedure is called *interlacing* because one needs to store $(\hat{\theta}(t+1), \hat{\theta}(t+2), \cdots, \hat{\theta}(t+d))$ and then update it to $(\hat{\theta}(t+d+1), \hat{\theta}(t+d+2), \cdots, \hat{\theta}(t+2d))$ etc. Goodwin, Ramadge and Caines [75] present a scheme based on this for the case where $C=1$, $d \geqq 1$ while Goodwin, Sin and Saluja [98] present one for the general case. Both use a stochastic approximation scheme to update the parameters.

8.5.4. Fuchs [97], [99], [100] has considered *indirect* or *explicit* schemes where one first estimates the coefficients of the (open loop) model and *not* the coefficients in the *prediction* form, as is done in [98]. In [99], [100] the case when the delay is $d \geqq 1$ is considered and a scheme is used which does *not* involve the cumbersome interlacing procedure.

THEOREM 8.6 (Fuchs). *Consider the system with the assumptions of Theorem* 8.5, *the only change being that* $d \geqq 1$. *Let* $\{y^*(t)\}$ *be a desired* bounded *reference trajectory. Consider the algorithm*:

$$\hat{\theta}(t+1)=\hat{\theta}(t)+\frac{\gamma}{r(t)}\phi(t)e(t+1), \qquad \gamma>0,$$
$$e(t+1) := y(t+1)-\hat{\theta}(t)^T\phi(t),$$
$$\phi^T(t) := (y(t),\cdots,y(t-n),u(t-d+1),\cdots,u(t-n-d+1),e(t),\cdots,e(t-n))$$

and

$$r(t+1)=r(t)+\phi^T(t+1)\phi(t+1), \qquad r(0)=1.$$

Let $\hat{y}(t+d|\hat{\theta}(t))=E(y(t+d)|\mathcal{F}_t, \hat{\theta}(t))$ *be the prediction of* $y(t+d)$ *if* $\hat{\theta}(t)$ *is the true parameter, as in* § 8.5.3. *Choose* $u(t)$ *so that* $\hat{y}(t+d|\hat{\theta}(t))=y^*(t+d)$. *Then* $\lim (1/N)\sum_1^N (y(t+d)-y^*(t+d))^2=0$ *and* $\limsup (1/N)\sum_1^N u^2(t)<+\infty$ *a.s.*

Proof. See [100]. □

The main difference, as explained above, is that one attempts to estimate *all* the coefficients of the true system. For this purpose an identification scheme (called RML_1) has been used (or rather, a stochastic approximation variant of it). The identification method estimates $w(t)$ by $e(t)$, the residuals $y(t)-\phi^T(t-1)\hat{\theta}(t-1)$, see § 8.1.5.

Another treatment of a stochastic approximation based adaptive control scheme is given by Kushner and Kumar [101]. This is based on the use of (truncated) bounded estimates and control inputs. Because of the latter, it is assumed that the system is (open loop) *stable*, i.e., all roots of A are outside the unit circle.

8.5.5. The most serious problem with the stochastic approximation based schemes of §§ 8.5.2 and 8.5.4 is that their *rate* of convergence has been observed in practice to be *very* slow compared to least squares based algorithms. (The rates of convergence of these algorithms are, to the author's knowledge, yet to be established rigorously.)

8.5.6. The asymptotic optimality of a strict least squares type scheme such as, say, in § 8.3 (the original self-tuning regulator) has yet to be established. However, algorithms *closely* related to a least squares algorithm have been analyzed by Kumar and Moore [102], Kumar [146] and Sin and Goodwin [103]. [102] considers an algorithm which gives more weight to past measurements than recent measurements. We now briefly discuss the scheme of [103]. They consider an estimation algorithm of the form

$$\hat{\theta}(t+1)=\hat{\theta}(t)+\gamma^{-1}(t)R^{-1}(t)\phi(t)[y(t+1)-\phi^T(t)\hat{\theta}(t)],$$
$$R(t+1)=\gamma(t+1)^{-1}[R(t)+\phi(t+1)\phi(t+1)^T]$$

where

$$\phi(t):=(y(t),\cdots,y(t-n),u(t-1),\cdots,u(t-n),-\bar{y}(t),\cdots,-\bar{y}(t-n))^T,$$
$$\bar{y}(t):=\phi^T(t-1)\hat{\theta}(t)$$

and a rule for calculating $\gamma(t)$, $0<\gamma(t)\leqq 1$, is specified.

Comparing this with the least squares recursion of §§ 8.1.2 and 8.1.3 we see the following salient differences. First $\gamma(t+1)$ is *not* always 1. Second, in the vector $\phi(t)$ occur the predictions $\bar{y}(t)=\phi^T(t-1)\hat{\theta}(t)$. The distinctive feature is that in calculating this prediction we use $\hat{\theta}(t)$ and *not* $\hat{\theta}(t-1)$. Thus this prediction is made on the basis of a *more recent* estimate, *and* is called an a posteriori prediction (contrast this with the recursions in §§ 8.5.1–8.5.3). Proofs and even convergence of identification schemes can rest on such fine differences, see Solo [88].

The use of a posteriori predictions (as opposed to a priori predictions) does indeed add one new wrinkle with respect to the *regulation* problem. Note that we choose $u(t)$ so that $\phi^T(t)\hat{\theta}(t)=0$. However $\bar{y}(t)\neq 0$ and so $(\bar{y}(t),\cdots,\bar{y}(t-n))$ *cannot* be dropped from the vector $\phi(t)$—as they could before, see Case 1 of § 8.5.3.

The computation of $\gamma(t)$ involves a cumbersome process and it would be of interest to obtain schemes which do not involve such computations.

You-Hong [104] extends the results of [103] to the general delay case. Gawthrop [105] contains an analysis of least squares based schemes. Kumar [106] exhibits the consequences of assuming that a certain "regularity condition" is satisfied by the closed loop system.

8.6. Convergence of parameter estimates and control laws. So far, we have not justified the use of the *phrase* (adjective) "self-tuning" in the names of these schemes. Specifically, we have not proved that the control law converges to the optimal control

law. We have only shown that the cost incurred is optimal (i.e., $\lim (1/N)\sum_1^N y^2(t) = \sigma^2$ a.s.).

The question has recently been addressed for the scheme of § 8.5.1 in [107].

THEOREM 8.7 (Becker, Kumar and Wei). *Consider the assumptions of Theorem* 8.4 *along with the additional assumption*:

$$(a_0+c_0)+(a_1+c_1)z+\cdots+(a_n+c_n)z^n \quad \textit{and} \quad b_0+b_1z+\cdots+b_nz^n$$

have no common factors and $|a_n+c_n|+|b_n|>0$. *Then* $\lim \hat{\theta}(t) = k(a_0+c_0,\ a_1+c_1, \cdots, a_n+c_n, b_0, b_1, \cdots, b_n)$ *where k is a random scalar.*

Proof. Since $u(t)$ is chosen to render $\phi^T(t)\hat{\theta}(t)=0$, it follows that $\phi(t)$ and $\hat{\theta}(t)$ are orthogonal. However the recursion $\hat{\theta}(t+1)=\hat{\theta}(t)+(\text{scalar})\ \phi(t)$ shows that $\hat{\theta}(t+1)-\hat{\theta}(t)$ is parallel to $\phi(t)$. These two facts together show that the $\hat{\theta}(t+1)-\hat{\theta}(t)$ is orthogonal to $\hat{\theta}(t)$. Thus the jumps $((\hat{\theta}(t+1)-\hat{\theta}(t)))$ in the parameter estimates are always orthogonal to their value $(\hat{\theta}(t))$ before the jump. Pythagoras' theorem now shows that $\|\hat{\theta}(t)\|$ is an increasing quantity.

By using the Schwarz inequality, the results of [75] can be refined to show that the stochastic Lyapunov function $\|\hat{\theta}(t)-\theta^0\|^2$ converges a.s. This shows, first, that $\{\hat{\theta}(t)\}$ converges to a sphere of random radius with center at θ^0. Second, it also shows that $\{\|\hat{\theta}(t)\|\}$ is bounded, and therefore also converges, since its geometric properties of the preceding paragraph have already shown that it is increasing. Thus $\{\hat{\theta}(t)\}$ also converges to a random sphere centered at the origin.

Since two spheres in Euclidean space (here $\mathbb{R}^{2n+1}$) can intersect either at a point (when they are tangential to each other) or in a hypersphere of dimension $2n$, we wish to show that the latter cannot happen. This is done by showing that there is a subsequence $\{\hat{\theta}(t_k)\}$ which converges to the line connecting the origin and θ^0, i.e., by showing that $\theta_i^0\hat{\theta}_p(t_k)-\theta_p^0\hat{\theta}_i(t_k)\to 0$ for every $i=1,2,\cdots,2n$.

This latter is done by showing that $\lim (1/N)\sum_{t=1}^N (\theta_i^0\hat{\theta}_p(t)-\theta_p^0\hat{\theta}_i(t))^2=0$ a.s. for $i=1,2,\cdots,2n$.

This last part of the proof is illustrated by using an example. Suppose $y(t+1)=a_0y(t)+a_1y(t-1)+b_0u(t)+b_1u(t-1)+w(t+1)+c_0w(t)$. Since the control law is

$$u(t)=-\frac{1}{\hat{b}_0(t)}(\hat{a}_0(t)y(t)+\hat{a}_1(t)y(t-1)+\hat{b}_1(t)u(t-1))$$

(where $\hat{\theta}(t)=:(\hat{a}_0(t), \hat{a}_1(t), \hat{b}_0(t), \hat{b}_1(t))$), the closed loop system is

$$y(t+1)=\left(a_0-b_0\frac{\hat{a}_0(t)}{\hat{b}_0(t)}\right)y(t)+\left(a_1-b_0\frac{\hat{a}_1(t)}{\hat{b}_0(t)}\right)y(t-1)+\left(b_1-b_0\frac{\hat{b}_1(t)}{\hat{b}_0(t)}\right)u(t-1)$$
$$+w(t+1)+c_0w(t).$$

Since

$$\lim\frac{1}{N}\sum_1^N y^2(t+1)=\lim\frac{1}{N}\sum_1^N w^2(t+1)=\sigma^2 \quad \text{a.s.},$$

it can be shown that

$$\lim\frac{1}{N}\sum_1^N\left[\left(a_0-b_0\frac{\hat{a}_0(t)}{\hat{b}_0(t)}\right)y(t)+\left(a_1-b_0\frac{\hat{a}_1(t)}{\hat{b}_0(t)}\right)y(t-1)\right.$$
$$\left.+\left(b_1-b_0\frac{\hat{b}_1(t)}{\hat{b}_0(t)}\right)u(t-1)\right]^2=0 \quad \text{a.s.}$$

Since $\{\hat{b}_0(t)\}$ is bounded, it follows that

$$\lim \frac{1}{N}\sum_1^N [(a_0\hat{b}_0(t) - b_0\hat{a}_0(t))y(t) + (a_1\hat{b}_0(t) - b_0\hat{a}_1(t))y(t-1) + (b_1\hat{b}_0(t) - b_0\hat{b}_1(t))u(t-1)]^2 = 0.$$

Since $\|\hat{\theta}(t)\|^2 = \sum_{p=1}^t \|\hat{\theta}(p) - \hat{\theta}(p-1)\|^2$ (by Pythagoras' theorem), we can replace $\hat{a}_0(t)$, $\hat{a}_1(t)$, $\hat{b}_0(t)$, $\hat{b}_1(t)$ by $\hat{a}_0(t-4)$, $\hat{a}_1(t-4)$, etc., to give

$$\lim \frac{1}{N}\sum_1^N [(a_0\hat{b}_0(t-4) - b_0\hat{a}_0(t-4))y(t) + (a_1\hat{b}_0(t-4) - b_0\hat{a}_1(t-4))y(t-1) + (b_1\hat{b}_0(t-4) - b_0\hat{b}_1(t-4))u(t-1)]^2 = 0.$$

Denote the square root of the expression in the above summand by $x(t)$. Now denote by $z(t)$ the expression in $x(t)$ where we change (only) each of $y(t)$, $y(t-1)$, $u(t-1)$ to $y(t-1)$, $y(t-2)$, $u(t-2)$ respectively. It can be shown that $(1/N)\sum_1^N x^2(t) \to 0$ implies $(1/N)\sum_1^N z^2(t) \to 0$.

By using a local convergence theorem for martingales, it can be shown that if $b_0x(t) + b_1z(t) =: \alpha(t)y(t) + \beta(t)y(t-1) + \gamma(t)y(t-2) + \eta(t)u(t-1) + \mu(t)u(t-2)$, then $\lim (1/N)\sum_1^N (\alpha^2(t) + \beta^2(t) + \gamma^2(t) + \eta^2(t) + \mu^2(t)) = 0$.

By using some algebra and condition (i) we get what we want. □

Some points need to be made. It is *not* true that $\lim \hat{\theta}(t) = (a_0 + c_0, \cdots, a_n + c_n, b_0, \cdots, b_n)$. In fact, in [107] it is shown that such a limiting value can have *zero* probability. Thus the parameter estimates *do converge*, but *not* to their true values (with regard to the model of Case 1, § 8.5.3).

However the control law is (with $\hat{\theta}(t) =: (\hat{\alpha}_0(t), \cdots, \hat{\alpha}_n(t), \hat{b}_0(t), \cdots, \hat{b}_n(t))^T$), $u(t) = -(1/\hat{b}_0(t))[\hat{\alpha}_0(t)y(t) + \cdots + \hat{\alpha}_n(t)y(t-n) + \hat{b}_1(t)u(t-1) + \cdots + \hat{b}_n(t)u(t-n)]$ and it *is* true that

$$\left(\frac{\hat{\alpha}_0(t)}{\hat{b}_0(t)}, \cdots, \frac{\hat{\alpha}_n(t)}{\hat{b}_0(t)}, \frac{\hat{b}_1(t)}{\hat{b}_0(t)}, \cdots, \frac{\hat{b}_n(t)}{\hat{b}_0(t)}\right)$$

converges to

$$\left(\frac{a_0 + c_0}{b_0}, \cdots, \frac{a_n + c_n}{b_0}, \frac{b_1}{b_0}, \cdots, \frac{b_n}{b_0}\right).$$

Thus the parameters of the control law do converge to the *optimal* values and self-tuning *does* occur. So here we finally have a demonstration of the result of Example 8.2 of § 8.3.1 and a justification of the description of the adaptive regulator as *self-tuning.*

An important point to note is that this proof does not rely on a "persistency of excitation" assumption. In fact, as we now demonstrate, a persistency of excitation condition does *not* hold. Let $\{N_k\}$ be a subsequence along which $\lim (1/N_k)\sum_1^{N_k} \phi(t)\phi(t)^T$ has a limit. We show that this limit is not positive definite. To see this, let $\hat{\theta}(\infty)$ be the limit of the parameter estimates. Then

$$\hat{\theta}(\infty)^T\left[\lim \frac{1}{N}\sum_1^{N_k} \phi(t)\phi(t)^T\right]\hat{\theta}(\infty) = \lim \frac{1}{N_k}\sum_1^{N_k} \hat{\theta}(\infty)^T\phi(t)\phi(t)^T\hat{\theta}(\infty)$$

$$= \lim \left[\frac{1}{N_k} \sum_1^{N_k} \hat{\theta}(t)^T \phi(t) \phi(t)^T \hat{\theta}(t) \right.$$

$$+ \frac{1}{N_k} \sum_1^{N_k} (\hat{\theta}(\infty) - \hat{\theta}(t))^T \phi(t) \phi^T(t) (\hat{\theta}(\infty) - \hat{\theta}(t))$$

$$\left. + \frac{2}{N_k} \sum_1^{N} \hat{\theta}(t)^T \phi(t) \phi^T(t) (\hat{\theta}(\infty) - \hat{\theta}(t) \right].$$

Since $\hat{\theta}(t)^T \phi(t) = 0$ by the specification of the control law and since $\hat{\theta}(t) \to \hat{\theta}(\infty)$, we see that

$$\hat{\theta}(\infty)^T \left[\lim \frac{1}{N_k} \sum_1^{N_k} \phi(t) \phi(t)^T \right] \hat{\theta}(\infty) = 0$$

showing the *singularity* of $\lim (1/N_k) \sum_1^{N_k} \phi(t) \phi(t)^T$. So we see that one should not *assume* persistency of excitation conditions to hold in *adaptive control.*

Convergence results such as the above need to be obtained in more general situations where least squares estimates are used, etc.

Caines and Lafortune [108] consider a stochastic approximation based scheme where a randomized control law is used, i.e., noise is injected into the system. By verifying that a "persistency of excitation" type condition holds, an auxiliary identification algorithm (running in parallel with the adaptive control algorithm) is shown to provide estimates which converge to the true values (an indirect or explicit scheme is used). Chen [109] also proves the strong consistency of a randomized control scheme with a modified least squares type parameter estimator as in [103]. Chen and Caines [110] reconsider the problem of [108] and show that one does not really need an additional parameter estimator in parallel.

8.7. Other proposed schemes. We now examine briefly various extensions of the basic self-tuning regulator (STR) which have been proposed to cope with various practical problems.

8.7.1. If the system is nonminimum phase, i.e., $b_0 + b_1 z + \cdots + b_n z^n$ has roots on or inside the unit circle, then we have seen in § 8.2.4 that the standard minimum variance regulator can have severe practical problems.

Åström and Wittenmark [111] consider a self-tuning scheme based on the *constrained* minimum variance regulator of Peterka [91], see Theorem 8.1 of § 8.2.5.

Clarke and Gawthrop [112] have proposed a generalization of the basic STR which incorporates models with nonzero steady state offset value (i.e., steady state output nonzero when input zero), tracking and an ability to cope with some nonminimum phase systems. The heart of the approach is the choice of $u(t)$ to minimize the d-step finite horizon cost criterion:

$$E\left[\left[\sum_{i=0}^{m} p_i y(t+d-i) - r_i y^*(t+d-i) \right]^2 + \sum_{i=0}^{m} q_i u(t-i)^2 \middle| \mathcal{F}_t \right].$$

Here $\{p_i, r_i, q_i\}$ are weighting coefficients and $y^*(t)$ can be a desired reference trajectory. By varying the coefficients one can obtain control laws which are stable for some nonminimum phase systems. It is also shown that the desired control law may be an equilibrium point (rest point) of the scheme. Gawthrop [113] and Clarke and Gawthrop [114] deal with generalizations of this.

Koivo [115] considers the multivariable version of this algorithm which in turn generalizes the multivariate version of the self-tuning regulator of Borisson [116].

Gawthrop [105] and Chen [117] analyze these schemes, the latter making some assumptions on the behaviour of the closed loop system. Goodwin, Ramadge and Caines [75] also allow for multivariable systems.

Wellstead, Edmunds, Prager and Zanker [118] propose a scheme based on the assignment of poles/zeros rather than on a cost criterion. This scheme is also based on practical considerations, the desired goal being the ability to deal with nonminimum phase systems, unknown delay d etc. It is shown that these schemes also have the property that the desired control law is an equilibrium point. Gawthrop [119] brings attention to some similarities between one of the schemes of [118] and the self-tuning controller [112]. Wellstead and Sanoff [120] extends [118], while Wellstead and Zanker [121] treats the tracking problem. Allidina and Hughes [122] presents a cost criterion approach. Åström and Wittenmark [123] contains an extensive treatment of the pole-zero assignment problem and the tracking problem.

Another approach is to use a cost criterion of the form $\lim (1/N)\sum_1^N y^2(t)+\rho u^2(t)$ where $\rho>0$ weights the control used. One can attempt to solve on-line, for every current set of estimated parameters, either a Riccati equation or perform a spectral factorization to obtain a control input which is optimal for the estimated parameters, see Åström, Borisson, Ljung and Wittenmark [124]. Mandl [125], [126] examines this scheme in a state space format and proves asymptotic optimality. The chief restriction is that the state is assumed to be completely observed and *only* the control gain matrix is unknown. [78] removes the latter restriction, but only allows the unknown parameter value to lie in a *finite* parameter set. The approach is based on § 7.3.

Grimble [127] proposes a different scheme, which is easier to implement, and which in contrast to some other schemes possesses the property that for the limiting values of some of the adjustable parameters (control weighting term etc.) the control law described in § 7.2.5 is obtained. See also Grimble [128].

Yet another approach is given by Kumar and Moore [129], [130], see also Clarke and Gawthrop [147].

8.7.2. In practical applications, the self-tuning regulator is implemented *not* on *constant* and unknown systems, but rather on *time varying* and unknown systems. It is hoped, in such cases, that the rate of convergence of the parameter estimates will be rapid in comparison with the rate of change of the system. To "keep up" with the changing system, one can make various modifications to the recursive least squares parameter estimation scheme. One can use a moving "window" of time, see Goodwin and Payne [87]. Alternatively, one can geometrically (exponentially) "forget" past observations on the system, i.e., one chooses $\hat{\theta}(t)$ so that it minimizes

$$\sum_{n=0}^{t} \lambda^{t-n}(y(t)-\phi^T(t)\theta)^2$$

over all θ. The factor λ, $0<\lambda\leqq 1$, is called the "exponential forgetting factor". The case $\lambda=1$ is the case that has been studied so far in this paper. Even if $\lambda<1$, one can obtain recursive schemes for the parameter estimates. The only change from § 8.1.2 is that

$$\hat{\theta}(t+1)=\hat{\theta}(t)+R_\lambda^{-1}(t)\phi(t)[y(t+1)-\hat{\theta}(t)^T\phi(t)]$$

where

$$R_\lambda(t)=\lambda R_\lambda(t-1)+\phi(t)\phi^T(t).$$

Alternatively, one can also obtain a recursive expression

$$R_\lambda^{-1}(t+1)=\frac{1}{\lambda}R_\lambda^{-1}(t)-\frac{1}{\lambda}\frac{R_\lambda^{-1}(t)\phi(t+1)\phi(t+1)^TR_\lambda^{-1}(t)}{\lambda+\phi^T(t+1)R_\lambda^{-1}(t)\phi(t+1)}.$$

During time intervals when the system under control is not changing, one would like to keep $\lambda \approx 1$ whereas if the system is changing, then one wants to keep $\lambda < 1$. Some practical long term problems (burst, blow up) can result from the choice of a constant forgetting factor, see Fortescue, Kershenbaum and Ydstie [131] and [123], Sanoff and Wellstead [144] and Saelid and Foss [145]. [131] proposes an "adaptive" selection of $\lambda(t)$ so that a measure of "information content" is kept constant. Latawiec and Chyra [132] also discuss this problem and offer some solutions. Lozano L. [133] obtains a bound on the asymptotic variance of the error in the parameter estimates. Zarrop [134] investigates the rate at which the forgetting factor sequence $\{\lambda(t)\}$ should converge to 1 so that asymptotic consistency of estimates can still be obtained.

A different approach to the situation of time varying parameters is taken by Caines [142] and Chen and Caines [143]. [142] analyzes the situation where the parameters form a converging martingale, while [143] analyzes the situation where the parameters constitute a uniformly bounded martingale difference sequence plus a constant.

Before leaving the topic of this section, we mention the work of Kalman [135], who as early as 1957, made a very strong case for considering schemes of this sort, and actually built a computer to implement them. Least squares estimates, forgetting factors, deadbeat control laws, efficient recursive least squares estimates, etc. all are ingredients of his scheme. See also the work of Peterka [136] and Peterka and Åström [137].

9. Conclusions. Clearly, much has been done and still much more remains to be done. For the Bayesian problems, efficient computational methods or analytic solutions to new problems are still needed. In the area of "dual control" of linear quadratic (Gaussian) systems, one needs approximations for which rigorous bounds on the quality of the approximations are available. For the adaptive control of Markov chains, one is faced with spaces of huge cardinality when the state spaces, control spaces etc. are large but finite. Further, efficient schemes to implement the algorithms are needed as well as studies of rates of convergence. In the self-tuning area, we still do not have theoretical tools to analyze all the schemes which have been proposed; in fact the original self-tuning regulator has yet to be fully analyzed. Rates of convergence have not been adequately established yet. The problem of *robustness* of the *adaptive* scheme has not been rigorously examined. An analysis of the steady state of the self-tuning regulator with a forgetting factor or periodic resetting is not available.

Clearly much more is needed in the way of theory.

Acknowledgments. The author is grateful to A. Becker, W. Lin, S. Mitter, S. Sastry, U. Shaked, P. Varaiya, J. Walrand and R. Weber for useful discussions. The author also wishes to acknowledge the special role played by H. Kushner in the development of this paper. Without his original suggestion and subsequent encouragement this paper would not have been written. The author also wishes to thank P. Varaiya and J. Walrand for their friendly hospitality during his visit to the University of California, Berkeley and S. Mitter during his visit to M.I.T.

REFERENCES

[1] D. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[2] R. BELLMAN, *Adaptive Control Processes: A Guided Tour*, Princeton Univ. Press, Princeton, NJ, 1961.

[3] E. B. DYNKIN, *Controlled random sequences*, Theory Prob. Appl., 10 (1965), pp. 1-14.

[4] M. AOKI, *Optimal control of partially observable Markovian systems*, J. Franklin Institute, 280 (1965), pp. 367-386.

[5] K. ÅSTRÖM, *Optimal control of Markov processes with incomplete state information*, J. of Math. Anal. Appl., 10 (1965), pp. 174-205.
[6] A. N. SHIRYAEV, *Some new results in the theory of controlled random processes*, English translation in Selected Translations in Mathematical Statistics and Probability, 8 (1970), pp. 49-130.
[7] C. STRIEBEL, *Sufficient statistics in the optimal control of stochastic systems*, J. of Math. Anal. Appl., 12 (1965), pp. 576-592.
[8] ———, *Optimal Control of Discrete Time Stochastic Systems*, Springer-Verlag, New York, 1975.
[9] K. HINDERER, *Foundations of Nonstationary Dynamic Programming with Discrete Time Parameter*, Springer-Verlag, New York, 1970.
[10] Y. SAWARAGI AND T. YOSHIKAWA, *Discrete time Markovian decision processes with incomplete state information*, Ann. Math. Stat., 41 (1970), pp. 78-86.
[11] D. RHENIUS, *Incomplete information in Markovian decision models*, Ann. Stat., 2 (1974), pp. 1327-1334.
[12] J. J. MARTIN, *Bayesian Decision Problems and Markov Chains*, John Wiley, New York, 1967.
[13] U. RIEDER, *Bayesian dynamic programming*, Adv. Appl. Prob., 7 (1975), pp. 330-348.
[14] K. M. VAN HEE, *Bayesian control of Markov chains*, Mathematical Center Tracts 95, Mathematisch Centrum, Amsterdam, 1978.
[15] J. C. GITTINS AND D. M. JONES, *A dynamic allocation index for the sequential design of experiments*, in Colloquia Mathematica Societatis Janos Bolyai 9, Progress in Statistics, European Meeting of Statisticians, pp. 241-266, J. Gani, K. Sarkadi and I. Vincze, eds., North-Holland, London, 1972.
[16] J. C. GITTINS AND K. D. GLAZEBROOK, *On Bayesian models in stochastic scheduling*, J. of Appl. Prob., 14 (1977), pp. 556-565.
[17] P. NASH, *Optimal allocation of resources between research projects*, Ph.D. Thesis, Cambridge University, 1973.
[18] P. WHITTLE, *Multi-armed bandits and the Gittins index*, J. Royal Stat. Soc., 42B (1980), pp. 143-149.
[19] K. D. GLAZEBROOK, *Optimal strategies for families of alternative bandit processes*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 858-861.
[20] P. WHITTLE, *Optimization Over Time: Dynamic Programming and Stochastic Control*—I, John Wiley, New York, 1982.
[21] S. ROSS, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.
[22] J. C. GITTINS, *Bandit processes and dynamic allocation indices*, J. Royal Stat. Soc., 41B (1979), pp. 148-177.
[23] K. D. GLAZEBROOK, *On a sufficient condition for superprocesses due to Whittle*, J. of Appl. Prob., 19 (1982), pp. 99-110.
[24] P. WHITTLE, *Arm-acquiring bandits*, Ann. of Prob., 9 (1981), pp. 284-292.
[25] P. VARAIYA, J. WALRAND AND C. BUYUKKOC, *Extensions of the multi-armed bandit problem: The discounted case*, Univ. California, Berkeley, 1983.
[26] M. ROTHSCHILD, *A two-armed bandit theory of market pricing*, J. Economic Theory, 9 (1974), pp. 185-202.
[27] F. P. KELLY, *Multi-armed bandits with discount factor near one: The Bernoulli case*, Ann. Stat., 9 (1981), pp. 897-1001.
[28] K. D. GLAZEBROOK, *On randomized dynamic allocation indices for the sequential design of experiments*, J. Royal Stat. Soc., 42B (1980), pp. 342-346.
[29] J. BATHER, *Randomized allocation of treatments in sequential trials*, Adv. Appl. Prob., 12 (1980), pp. 174-182.
[30] K. D. GLAZEBROOK, *On the evaluation of suboptimal strategies for families of alternative bandit processes*, J. of Appl. Prob., 19 (1982), pp. 716-722.
[31] P. R. KUMAR AND T. I. SEIDMAN, *On the optimal solution of the one-armed bandit adaptive control problem*, IEEE Trans. on Automat. Control, AC-26 (1981), pp. 1176-1184.
[32] D. BLACKWELL, *Discounted dynamic programming*, Ann. Math. Stat., 36 (1965), pp. 226-235.
[33] R. HOWARD, *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA, 1960.
[34] J. K. SATIA AND R. E. LAVE, *Markov decision processes with uncertain transition probabilities*, Oper. Res., 21 (1973), pp. 728-740.
[35] A. M. MAKOWSKI, *Results on the filtering problem for linear systems with non-Gaussian initial conditions*, Proc. 21st IEEE Conference on Decision and Control, 1982, pp. 201-204.
[36] K. J. ÅSTRÖM AND B. WITTENMARK, *Problems of identification and control*, J. Math. Anal. Appl., 34 (1971), pp. 90-113.
[37] A. A. FELDBAUM, *The theory of dual control* IV, Automation and Remote Control, 22 (1961), pp. 109-121.
[38] O. L. R. JACOBS AND J. W. PATCHELL, *Caution and probing in stochastic control*, Internat. J. Control, 16 (1972), pp. 189-199.

[39] E. TSE AND Y. BAR-SHALOM, *Wide-sense adaptive dual control for nonlinear stochastic systems*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 98–108.

[40] ———, *An actively adaptive control for linear systems with random parameters via the dual control approach*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 109–117.

[41] ———, *Generalized certainty equivalence and dual effect in stochastic control*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 817–819.

[42] ———, *Dual effect certainty equivalence, and separation in stochastic control*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 494–500.

[43] B. WITTENMARK, *Stochastic adaptive control methods: a survey*, Internat. J. of Control, 21 (1975), pp. 705–730.

[44] C. J. WENK AND Y. BAR-SHALOM, *A multiple model adaptive dual control algorithm for stochastic systems with unknown parameters*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 703–710.

[45] Y. BAR-SHALOM, *Stochastic dynamic programming: caution and probing*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1184–1194.

[46] J. G. DESHPANDE, T. N. UPADHYAY AND D. G. LAINIOTIS, *Adaptive control of linear stochastic systems*, Automatica, 9 (1973), pp. 107–115.

[47] D. G. LAINIOTIS, *Partitioning: a unifying framework for adaptive systems*, II: *control*, Proc. IEEE, 64 (1976), pp. 1182–1198.

[48] P. L. DERSIN, M. ATHANS AND D. A. KENDRICK, *Some properties of the dual adaptive stochastic control algorithm*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1001–1008.

[49] H. ROBBINS, *Some aspects of the sequential design of experiments*, Bull. Amer. Math. Soc., 58 (1952), pp. 527–537.

[50] V. BORKAR AND P. VARAIYA, *Adaptive control of Markov chains*, I: *finite parameter set*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 953–958.

[51] P. R. KUMAR AND A. BECKER, *A new family of optimal adaptive controllers for Markov chains*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 137–146.

[52] B. SAGALOVSKY, *Adaptive control and parameter estimation in Markov chains: a linear case*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 414–419.

[53] P. R. KUMAR, *Adaptive control with a compact parameter set*, this Journal, 20 (1982), pp. 9–13.

[54] V. BORKAR AND P. VARAIYA, *Identification and adaptive control of Markov chains*, this Journal, 20 (1982), pp. 470–489.

[55] P. MANDL, *Estimation and control in Markov chains*, Adv. Appl. Prob., 6 (1974), pp. 40–60.

[56] ———, *On the adaptive control of finite state Markov processes*, Z. Wahrsch. Verw. Geb., 27 (1973), pp. 263–276.

[57] ———, *On the control of a Markov chain in the presence of unknown parameters*, Trans. Sixth Prague Conference on Information Theory, Random Processes and Statistical Decision Functions, Prague, 1971, pp. 601–612.

[58] M. KURANO, *Discrete-time Markovian decision processes with an unknown parameter-average return criterion*, J. Oper. Res. Soc. Japan, 15 (1972), pp. 67–76.

[59] M. KOLONKO, *Strongly consistent estimation in a controlled Markov renewal model*, J. Appl. Prob., 19 (1972), pp. 532–545.

[60] ———, *The average-optimal adaptive control of a Markov renewal model in presence of an unknown parameter*, Math. Operations Forsch. Statist. Ser. Optim., 13 (1982), pp. 567–591.

[61] J. P. GEORGIN, *Estimation et contrôles des chaînes de Markov sur des espaces arbitraires*, in Lecture Notes in Mathematics, 636, Springer-Verlag, Berlin, 1978.

[62] V. V. BARANOV, *Recursive algorithms of adaptive control in stochastic systems*, Cybernetics, 17 (1981), pp. 815–824.

[63] M. SCHÄL, *Estimation and control in stochastic dynamic programming: finite and asymptotic results*, Report No. 521, Univ. Bonn, July 1982.

[64] B. L. FOX AND J. E. ROLPH, *Adaptive policies for Markov renewal programs*, Ann. Stat., 1 (1973), pp. 334–341.

[65] B. DOSHI AND S. SHREVE, *Strong consistency of a modified maximum likelihood estimator for controlled Markov chains*, J. Appl. Prob., 17 (1980), pp. 726–734.

[66] M. SATO, K. ABE AND H. TAKEDA, *Learning control of finite Markov chains with unknown transition probabilities*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 502–505.

[67] L. M. LYUBCHIK AND A. S. POZNYAK, *Learning automata in stochastic plant control problems*, Automation and Remote Control, 6 (1974), pp. 777–789.

[68] Y. M. EL-FATTAH, *Recursive algorithms for adaptive control of finite Markov chains*, IEEE Trans. Systems, Man and Cybernetics, SMC-11 (1981), pp. 135–144.

[69] Y. M. EL-FATTAH, *Gradient approach for recursive estimation and control in finite Markov chains*, Adv. Appl. Prob., 13 (1981), pp. 778–803.

[70] Y. Z. TSYPKIN, *Adaptation and Learning in Automatic Systems*, Academic Press, New York, 1971.
[71] Y. Z. TSYPKIN, *Foundations of the Theory of Learning Systems*, Academic Press, New York, 1973.
[72] J. NEVEU, *Discrete-Parameter Martingales*, North-Holland, Amsterdam, 1975.
[73] H. ROBBINS AND D. SIEGMUND, *A convergence theorem for nonnegative almost super-martingales and some applications*, in Optimization Methods in Statistics, J. S. Rustagi, ed., Academic Press, New York, 1971, pp. 233-257.
[74] B. T. POLYAK AND Y. Z. TSYPKIN, *Pseudogradient adaptation and training algorithms*, Automation and Remote Control, 23 (1973), pp. 377-397.
[75] G. GOODWIN, P. RAMADGE AND P. CAINES, *Discrete time stochastic adaptive control*, this Journal, 19 (1981), pp. 829-853.
[76] P. R. KUMAR AND W. LIN, *Optimal adaptive controllers for unknown Markov chains*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 765-774.
[77] P. R. KUMAR, *Simultaneous identification and adaptive control of unknown systems over finite parameter sets*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 68-76.
[78] ———, *Optimal adaptive control of linear-quadratic-Gaussian systems*, this Journal, 21 (1983), pp. 163-178.
[79] A. BECKER AND P. R. KUMAR, *Optimal strategies for the N-armed bandit problem*, Math. Research Report No. 81-1, Univ. Maryland Baltimore County, 1981.
[80] J. RIORDAN, *An adaptive automaton controller for discrete time Markov processes*, Automatica, 5 (1969), pp. 721-730.
[81] F. M. D'HULSTER, R. M. C. DEKEYSER AND A. R. VAN CAUWENBERGHE, *Simulations of adaptive controllers for a paper machine headbox*, Automatica, 19 (1983), pp. 407-414.
[82] C. KIPARISSIDES AND S. L. SHAH, *Self tuning and stable adaptive control of a batch polymerization reactor*, Automatica, 19 (1983), pp. 223-224.
[83] G. DUMONT, *Self-tuning control of a chip refiner motor load*, Automatica, 18 (1982), pp. 307-314.
[84] H. BOEHM, *Adaptive control to a dry etch process by microcomputer*, Automatica, 18 (1982), pp. 665-673.
[85] K. Y-J. KO, B. C. MCINNIS AND G. C. GOODWIN, *Adaptive control and identification of the dissolved oxygen process*, Automatica, 18 (1982), pp. 727-730.
[86] T. L. LAI AND C. Z. WEI, *Least squares estimates in stochastic regression with applications to identification and control of dynamic systems*, Ann. Stat., 10 (1982), pp. 154-166.
[87] G. C. GOODWIN AND R. L. PAYNE, *Dynamic System Identification*, Academic Press, New York, 1977.
[88] V. SOLO, *The convergence of* AML, IEEE Trans. Automat. Control, AC-24 (1979), pp. 958-962.
[89] K. J. ÅSTRÖM, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.
[90] ———, *Lectures on the identification problem—the least squares method*, Report 6806, Lund Institute of Technology, 1968.
[91] V. PETERKA, *On steady state minimum variance control strategy*, Kybernetika, 8 (1972), pp. 218-231.
[92] K. J. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica, 9 (1973), pp. 185-199.
[93] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551-575.
[94] ———, *On positive real transfer functions and the convergence of some recursive schemes*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 539-551.
[95] H. KUSHNER, *Convergence of recursive adaptive and identification procedures via weak convergence theory*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 921-930.
[96] H. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.
[97] J.-J. J. FUCHS, *Explicit self-tuning methods*, Proc. IEE, 127 (1980), pp. 259-264.
[98] G. C. GOODWIN, K. S. SIN AND K. K. SALUJA, *Stochastic adaptive control and prediction—the general delay-colored noise case*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 946-949.
[99] J.-J. J. FUCHS, *Indirect stochastic adaptive control: the general delay-white noise case*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 219-223.
[100] ———, *Indirect stochastic adaptive control: the general delay-colored noise case*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 470-472.
[101] H. KUSHNER AND R. KUMAR, *Convergence and rate of convergence of a recursive identification and adaptive control scheme which uses truncated estimators*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 775-782.
[102] R. KUMAR AND J. B. MOORE, *Convergence of adaptive minimum variance algorithms via weighting coefficient selection*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 146-153.
[103] K. S. SIN AND G. C. GOODWIN, *Stochastic adaptive control using a modified least squares algorithm*, Automatica, 18 (1982), pp. 315-321.
[104] Z. YOU-HONG, *Stochastic adaptive control and prediction based on a modified least squares—the general delay-colored noise case*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 1257-1260.

[105] P. J. GAWTHROP, *On the stability and convergence of a self-tuning controller*, Internat. J. Control, 31 (1980), pp. 973-998.

[106] R. KUMAR, *Almost sure convergence of adaptive identification prediction and control algorithms*, LCDS 81-8, Div. Applied Mathematics, Brown Univ., Providence, RI, 1981.

[107] A. BECKER, P. R. KUMAR AND C. Z. WEI, *Adaptive control with the stochastic approximation algorithm: geometry and convergence*, Univ. Maryland Baltimore County Mathematics Research Report No. 83-8, 1983, IEEE Trans. Automat. Control, to appear.

[108] P. E. CAINES AND S. LAFORTUNE, *Adaptive optimization with recursive identification for stochastic linear systems*, McGill Univ., Montreal, 1981.

[109] H. F. CHEN, *Recursive system identification and adaptive control by use of the modified least squares algorithm*, McGill Univ., Montreal, 1983.

[110] H. F. CHEN AND P. E. CAINES, *The strong consistency of the stochastic gradient algorithm of adaptive control*, McGill Univ., Montreal, 1983.

[111] K. ÅSTRÖM AND B. WITTENMARK, *Analysis of a self-tuning regulator for non-minimum phase systems*, Proc. IFAC Symposium on Stochastic Control, Budapest, Hungary, 1974.

[112] D. W. CLARKE AND P. J. GAWTHROP, *Self-tuning controller*, Proc. IEEE, 122 (1975), pp. 929-934.

[113] P. J. GAWTHROP, *Some interpretations of the self-tuning controller*, Proc. IEEE, 124 (1977), pp. 889-894.

[114] D. W. CLARKE AND P. J. GAWTHROP, *Self-tuning control*, Proc. IEEE, 126 (1979), pp. 633-640.

[115] H. N. KOIVO, *A multivariable self-tuning controller*, Automatica, 16 (1980), pp. 351-366.

[116] U. BORISSON, *Self-tuning regulators for a class of multivariable systems*, Automatica, 15 (1979), pp. 209-215.

[117] CHEN HAN-FU, *Self-tuning controller and its convergence under correlated noise*, Internat. J. Control, 35 (1982), pp. 1051-1059.

[118] P. E. WELLSTEAD, J. M. EDMUNDS, D. PRAGER AND P. ZANKER, *Self-tuning pole/zero assignment regulators*, Internat. J. Control, 30 (1979), pp. 1-26.

[119] P. J. GAWTHROP, *A comment on "self-tuning pole/zero assignment regulators*, Internat. J. Control, 31 (1980), pp. 999-1002.

[120] P. E. WELLSTEAD AND S. P. SANOFF, *Extended self-tuning algorithm*, Internat. J. Control, 34 (1981), pp. 433-455.

[121] P. E. WELLSTEAD AND P. ZANKER, *Servo self-tuners*, Internat. J. Control, 30 (1979), pp. 27-36.

[122] A. Y. ALLIDINA AND F. M. HUGHES, *Generalized self-tuning controller with pole assignment*, Proc. IEEE, 127D (1980), pp. 13-18.

[123] K. J. ÅSTRÖM AND B. WITTENMARK, *Self-tuning controllers based on pole-zero placement*, Proc. IEEE, 127D (1980), pp. 120-130.

[124] K. J. ÅSTRÖM, U. BORISSON, L. LJUNG AND B. WITTENMARK, *Theory and applications of self-tuning regulators*, Automatica, 13 (1977), pp. 457-476.

[125] P. MANDL, *The use of optimal stationary policies in the adaptive control of linear systems*, Proc. Symposium to Honour Jerzy Neyman, Warsaw, 1974, pp. 223-243.

[126] ———, *Some results in the adaptive control of linear systems*, Trans. Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians, Prague, 1977, pp. 399-410.

[127] M. J. GRIMBLE, *A control weighted minimum-variance controller for non-minimum phase systems*, Internat. J. Control, 33 (1981), pp. 751-762.

[128] ———, *Weighted minimum-variance self-tuning control*, Internat. J. Control, 36 (1982), pp. 597-609.

[129] R. KUMAR AND J. B. MOORE, *Minimum variance control harnessed for non-minimum-phase plants*, Technical Report No. EE8014, July 1980.

[130] ———, *On adaptive minimum variance regulation for non-minimum phase plants*, Automatica, 19 (1983), pp. 449-451.

[131] T. R. FORTESCUE, L. S. KERSHENBAUM AND B. E. YDSTIE, *Implementation of self-tuning regulators with variable forgetting factors*, Automatica, 17 (1981), pp. 831-835.

[132] K. LATAWIEC AND M. CHYRA, *On low frequency and long-run effects in self-tuning control*, Automatica, 19 (1983), pp. 419-424.

[133] R. LOZANO L., *Convergence analysis of recursive identification algorithms with forgetting factor*, Automatica, 19 (1983), pp. 95-97.

[134] M. B. ZARROP, *Variable forgetting factors in parameter estimation*, Automatica, 19 (1983), pp. 295-298.

[135] R. KALMAN, *Design of a self-optimizing control system*, Trans. ASME, 80 (1958), pp. 468-478.

[136] V. PETERKA, *Adaptive digital regulation of noisy systems*, Proc. 2nd IFAC Symposium on Identification and Process Parameter Estimation, Prague, 1970.

[137] V. PETERKA AND K. J. ÅSTRÖM, *Control of multivariate systems with unknown but constant parameters*, Proc. 3rd IFAC Symposium, Hague/Delft, 1973.

[138] U. SHAKED AND P. R. KUMAR, *Minimum variance control of multivariable* ARMAX *systems*, LIDS Report, Massachusetts Institute of Technology, Cambridge, 1984.
[139] M. METIVIER AND P. PRIOURET, *Applications of a Kushner and Clark lemma to general classes of stochastic algorithms*, IEEE Trans. Inform. Theory, IT-30, Part 1 (1984), pp. 140–151.
[140] R. RISHEL, *An exact formula for a linear quadratic adaptive stochastic optimal control law*, Preprint, 1984.
[141] O. HIJAB, *Optimal adaptive control and stabilization of families of linear systems*, Preprint, 1983, Systems and Control Letters, to appear.
[142] P. CAINES, *Stochastic adaptive control: randomly varying parameters and continually disturbed controls*, Proc. 8th IFAC Congress, Kyoto, 1981, pp. 925–930.
[143] H. F. CHEN AND P. E. CAINES, *On the adaptive control of a class of systems with random parameters and disturbances*, Preprint, 1983.
[144] S. P. SANOFF AND P. E. WELLSTEAD, *Comments on "Implementation of self-tuning regulators with variable forgetting factors,"* Automatica, 19 (1983), pp. 345–346.
[145] S. SAELID AND B. FOSS, *Adaptive controllers with a vector variable forgetting factor*, Proc. 22nd IEEE Conference on Decision and Control 3, San Antonio, 1983, pp. 1488–1494.
[146] R. KUMAR, *Simultaneous adaptive control and identification via the weighted least-square algorithm*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 259–263.
[147] D. W. CLARKE AND P. J. GAWTHROP, *Comments on "On adaptive minimum variance regulation for non-minimum phase plants"*, Automatica, 20 (1984), pp. 261.
[148] V. SOLO, *The convergence of an instrumental-variable-like recursion*, Automatica, 17 (1981), pp. 545–547.
[149] H. KUSHNER AND A. SHWARTZ, *An invariant measure approach to the convergence of stochastic approximations with state dependent noise*, this Journal, 22 (1984), pp. 13–27.
[150] H. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, MA, 1984.