INFORMATION BOUNDS, CERTAINTY EQUIVALENCE AND LEARNING IN ASYMPTOTICALLY EFFICIENT ADAPTIVE CONTROL OF TIME-ININVARIANT STOCHASTIC SYSTEMS

BY

TZE LEUNG LAI

TECHNICAL REPORT NO. 21
February 1991

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS87-15614

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Information Bounds, Certainty Equivalence and Learning in Asymptotically Efficient Adaptive Control of Time-Invariant Stochastic Systems

By

Tze Leung Lai and Zhiliang Ying

Technical Report No. 21

February 1991

Prepared Under the Auspices

of

National Science Foundation Grant DMS87-15614

Also supported by The National Security Agency Grant MDA 904-89-H-2040 and issued as Technical Report No. 16, Department of Statistics, Stanford University.

Department of Statistics

Stanford University

Stanford, California
Information Bounds, Certainty Equivalence and Learning in Asymptotically Efficient Adaptive Control of Time-Invariant Stochastic Systems

Tze Leung Lai

Abstract

In adaptive control of stochastic systems with unknown but time-invariant parameters, a fundamental difficulty is the apparent dilemma between the need for information to learn about the unknown system parameters and the control objective of driving the system outputs towards certain target values. Although this difficulty can in principle be resolved by a Bayesian formulation, the dynamic programming equations involved are usually prohibitively difficult to handle, both computationally and analytically. Herein we show how relatively simple and yet asymptotically efficient control rules can be developed by perturbing suitably chosen certainty-equivalence rules with probing signals. In this connection, asymptotic lower bounds on the minimal amount of information needed for efficient control are also developed to keep the probing perturbations within these bounds.
INFORMATION BOUNDS, CERTAINTY EQUIVALENCE AND
LEARNING IN ASYMPTOTICALLY EFFICIENT ADAPTIVE
CONTROL OF TIME-ININVARIANT STOCHASTIC SYSTEMS

Tze Leung Lai
Department of Statistics, Stanford University, Stanford, CA 94305

1. INTRODUCTION

The past two decades have witnessed important advances in the subject of adaptive
control of stochastic systems with unknown but time-invariant parameters. A fundamental
difficulty in this area is the apparent dilemma between the need for information to learn
about the unknown system parameters and the control objective of driving the system
outputs towards certain target values. This difficulty has been more or less completely
resolved in several prototypical problems. Three such problems will be discussed herein.
They are (i) the multi-armed bandit problem in §2, (ii) the multi-period control problem
in stochastic regression models in §3, and (iii) adaptive control of an ARMAX system in
§4.

In principle, given a prior distribution of the unknown system parameters and the joint
probability distribution of the sequence of random variables that determine the stochastic
system, one can formulate an adaptive control problem as a dynamic programming prob-
lem in which the “state” is the conditional distribution of the original system state and
parameter vector given the past observations (cf. [1]). However, the dynamic programming
equations are usually prohibitively difficult to handle, both computationally and analyti-
cally. Moreover, it is often not possible to specify a realistic probability law for all the
random variables involved and a reasonable prior distribution for the unknown parameter
vector.

Instead of the Bayesian approach, a much more practical and intuitively appealing
alternative is the “certainty-equivalence” approach that first finds the optimal solution
when the system parameters are known and then replaces the unknown parameters by
their sample estimates at every stage. This approach incorporates sequential learning,
in the form of updating the parameter estimates based on all the available data at every
stage, into the control scheme that tries to mimic the optimal rule assuming known system
parameters. The idea is particularly attractive when the optimal control law assuming
known system parameters has a simple recursive form that can be implemented in real time,
as is the case for the three classical problems to be discussed in the subsequent sections. If one can also develop real-time recursive algorithms for updating the parameter estimates which are used to replace the unknown parameters in the "fictitious" optimal rule, then the certainty-equivalence rule will indeed satisfy real-time computational constraints.

Instead of trying to solve a difficult optimization problem that represents ignorance of the actual parameter values in the form of a prior distribution as in the Bayesian approach, the certainty-equivalence approach combines a much simpler optimization problem assuming known parameter values with a recursive estimation problem to learn about the parameter values sequentially from the data. The performance of this approach, therefore, depends critically on how well the parameter values are estimated. In view of this, it is therefore highly desirable to estimate the parameters as efficiently as possible. However, there are also practical computational constraints here for on-line implementation. In the case of ARMAX systems to be discussed in §4, this poses formidable problems since the statistically efficient estimation procedures involve nonlinear estimating equations which have to be solved off-line and which also require storing all the data as they accumulate over time. §4 shows, however, that it is possible to develop recursive estimators that are asymptotically as efficient as the off-line efficient estimators by a parallel implementation of a few well known recursive identification algorithms.

Clearly how well the parameter values can be estimated depends not only on the choice of efficient estimation procedures but also on how much information about the parameters there is in the data. In statistical estimation theory, one usually summarizes the amount of information in the form of Kullback-Leibler information numbers or Fisher information matrices, and represents one's uncertainty about the estimates by using confidence intervals or Bayesian posterior distributions. These basic concepts provide important tools in assessing the amount of information about the unknown parameters in the data and the amount of uncertainty in the sequential parameter estimates for the certainty-equivalence approach. In particular, when the amount of information is inadequate and the parameter estimates are highly uncertain, one should introduce "probing signals" to enhance the information content of the data instead of strictly following the certainty-equivalence rule which replaces the unknown parameters in the fictitious optimal rule by dubious estimates. On the other hand, when the standard errors of the parameter estimates are relatively small, the certainty-equivalence input should provide a good approximation to the optimal input that assumes known parameter values. This obvious idea plays an important
role in the development of asymptotically efficient adaptive control rules that perturb the certainty-equivalence approach with probing signals in §§2-4.

A basic issue concerning this information enhancement of the certainty-equivalence approach is the maximal proportion of probing signals that can be introduced so that the resultant control rule may still attain full asymptotic efficiency. In this connection the notion of "asymptotic efficiency" of a control rule has also to be precisely defined. For each of the adaptive control problems considered in §§2-4, we address these issues by considering the fictitious situation which assumes knowledge of some crucial parameter(s) so that there is negligible conflict between estimation and control. The performance of asymptotically optimal solutions in this much more tractable situation then sets a lower bound that we can hope to achieve even when all the parameters are unknown, and asymptotic efficiency can therefore be defined as achieving such asymptotic performance. More importantly, the difference in performance between this fictitious situation assuming knowledge of some (but not all) parameters and the ideal situation in which all parameters are known provides an asymptotic lower bound on the amount of information about the unknown parameters needed for optimal control. By keeping the probing perturbations within these bounds, we show in §§2-4 that suitably chosen certainty-equivalence rules with suitable perturbations are indeed asymptotically efficient.

2. INFORMATION BOUNDS AND DYNAMIC ALLOCATION INDICES IN THE MULTI-ARMED BANDIT PROBLEM

Let \( \Pi_j, j = 1, \ldots, k \), denote statistical populations (treatments, manufacturing processes, etc.) specified, respectively, by univariate density functions \( f(x; \theta_j) \) with respect to some nondegenerate measure \( \nu \), where \( f(\cdot; \cdot) \) is known and the \( \theta_j \) are unknown parameters belonging to some set \( \Theta \). Assume that \( E_\theta |X| = \int_{-\infty}^{\infty} |x| f(x; \theta) d\nu(x) < \infty \) for all \( \theta \in \Theta \). How should we sample \( x_1, \ldots, x_N \) sequentially from the \( k \) populations in order to maximize, in some sense, the expected value of the sum \( S_N = x_1 + \cdots + x_N \)? This is the classical "multi-armed bandit problem", with specified horizon \( N \), in the statistics and engineering literature. The name derives from an imagined slot machine with \( k \geq 2 \) arms. When an arm is pulled, the player wins a random reward. For each arm \( j \), there is an unknown probability distribution \( \Pi_j \) of the reward, and the player's problem is to choose \( N \) pulls on the \( k \) arms so as to maximize the total expected reward. There is an apparent dilemma between the need to learn from all populations about their parameter values ("information") and the objective of sampling only from the best population ("control").
An *adaptive allocation rule* $\varphi$ is a sequence of random variables $\varphi_1, \cdots, \varphi_N$ with values in the set $\{1, \cdots, k\}$ and such that the event $\{\varphi_i = j\}$, $j = 1, \cdots, k$, belongs to the $\sigma$-field $\mathcal{F}_{i-1}$ generated by the previous observations $\varphi_1, x_1, \cdots, \varphi_{i-1}, x_{i-1}$. Letting $\mu(\theta) = \int_{-\infty}^{\infty} xf(x; \theta) d\nu(x)$ and $\theta = (\theta_1, \cdots, \theta_k) \in \Theta^k$, it follows that for every $n \leq N$

$$E_{\theta} S_n = \sum_{i=1}^{n} \sum_{j=1}^{k} E_{\theta} \{E_{\theta}(x_i I_{\{\varphi_i = j\}} \mid \mathcal{F}_{i-1})\} = \sum_{j=1}^{k} \mu(\theta_j) E_{\theta} T_N(j), \quad (2.1)$$

where

$$T_n(j) = \sum_{i=1}^{n} I_{\{\varphi_i = j\}} \quad (2.2)$$

denotes the number of observations that $\varphi$ samples from $\Pi_j$ up to stage $n$. Hence, the objective of maximizing $E_{\theta} S_N$ is equivalent to that of minimizing the regret

$$R_N(\theta) = N \mu^*(\theta) - E_{\theta} S_N = \sum_{j: \mu(\theta_j) < \mu^*(\theta)} (\mu^*(\theta) - \mu(\theta_j)) E_{\theta} T_N(j), \quad (2.3)$$

where $\mu^*(\theta) = \max_{1 \leq j \leq k} \mu(\theta_j)$. In particular, the usual Bayesian formulation of the multi-armed bandit problem, stated in the form of maximizing $\int E_{\theta} S_N dH(\theta)$, can be restated in the more convenient form of minimizing the Bayes risk $\int R_N(\theta) dH(\theta)$, where $H$ is a prior distribution on $\Theta^k$.

In principle, one can use dynamic programming to study the problem of minimizing $\int R_N(\theta) dH(\theta)$. For the case where $k = 2$ and $\Theta$ has two elements, which we shall denote by $a, b$ with $\mu(a) > \mu(b)$, Feldman [2] found by this approach that for the prior distribution which assigns probability $p$ to the parameter vector $\theta = (a, b)$ and probability $1 - p$ to the vector $(b, a)$, the allocation rule that chooses $\Pi_1$ or $\Pi_2$ at stage $i + 1$ according as $p_i \geq \frac{1}{2}$ or $p_i < \frac{1}{2}$ is Bayes, where $p_i$ denotes the posterior probability in favor of the vector $(a, b)$ at the end of stage $i$ ($p_0 = p$). For the case of $k = 2$ Bernoulli populations, Fabius and van Zwet [3] and Berry [4] studied the dynamic programming equations analytically and obtained several interesting results about the Bayes rules with respect to general priors. Beyond the two-point priors considered by Feldman, Bayes rules are usually described only implicitly by the dynamic programming equations, whose numerical solution is too complicated for practical implementation when $N$ is large.

Instead of considering the finite-horizon problem of maximizing the expected value of $S_N = \sum_{i=1}^{N} x_i$, one can study the discounted problem that maximizes the expected value
of the series $\sum_1^\infty \beta^{i-1}x_i$, where $0 < \beta < 1$. Major advances in the discounted multi-armed bandit problem were made in the seventies by Gittins and Jones [5], Gittons [6] and Whittle [7]. Their results show, for general $k \geq 2$, that under the assumption of independent prior distributions $G_j$ on $\theta_j$ ($j = 1, \ldots, k$), the optimal solution to the Bayes problem of maximizing

$$\int \cdots \int E_{\theta}(\sum_{i=1}^\infty \beta^{i-1}x_i) dG_1(\theta_1) \cdots dG_k(\theta_k)$$

is the “index rule” described below. To begin with, suppose that $\theta$ is a random variable having distribution $G$ and that conditional on $\theta$, $Y_1, Y_2, \cdots$ are i.i.d. with common density $f(x; \theta)$ with respect to $\nu$. Let $G^{[r]}$ denote the conditional distribution of $\theta$ given $Y_1, \ldots, Y_r$. (Thus, $G_j^{[r]}$ is the posterior distribution of the parameter $\theta_j$ of population $\Pi_j$ given $r$ observations $Y_1, \ldots, Y_r$ from $\Pi_j$.) Set $G^{[0]} = G$. As is well known, the conditional distribution of $(\theta, Y_{r+1}, Y_{r+2}, \cdots)$ given $(Y_1, \ldots, Y_r)$ can be described by saying that $\theta$ has distribution $G^{[r]}$ and that $Y_{r+1}, Y_{r+2}, \cdots$ are i.i.d. with common density $f(x; \theta)$ given the value of $\theta$. Therefore, for $r > m$,

$$E[Y_m|Y_1, \ldots, Y_r] = \int \mu(\theta)dG^{[r]}(\theta) = E[\mu(\theta)|Y_1, \ldots, Y_r].$$

(2.5)

The Gittins index $M(G)$ of the distribution $G$ on $\theta$ is defined as the infimum of the set of solutions $M$ of the equation

$$\sup_{r \geq 0} E \left\{ \sum_{i=0}^{r-1} \beta^i E[\mu(\theta)|Y_1, \ldots, Y_i] + M \sum_{i=r}^\infty \beta^i \right\} = M \sum_{i=0}^\infty \beta^i,$$

where $\sup_{r \geq 0}$ is over all stopping times $r$. The index rule $\varphi^*$ is to sample at stage $n + 1$ from the population $\Pi_j^*$ that has the largest Gittins index at the end of stage $n$, i.e.,

$$M(G_j^{[T_n(j^*)]}) = \max_{1 \leq j \leq n} M(G_j^{[T_n(j)]}).$$

(2.7)

For fixed $j$, the equation (2.6) defining the Gittins index of the posterior distribution $G = G_j^{[T_n(j)]}$ arises from the following (fictitious) optimization problem, cf. [6], [7]. Suppose that the largest mean, $\max_{i \neq j} \mu(\theta_i)$, of the remaining $k - 1$ populations and its population label $i^* \neq j$ are both known. Since $\theta_j$ is unknown, we have only to decide whether $\mu(\theta_j)$ exceeds $M = \max_{i \neq j} \mu(\theta_i)$, in which case it is profitable to sample from $\Pi_j$, or $\mu(\theta_j) < M$, in which case we should switch to sampling from $\Pi_{i^*}$. Initially, in ignorance of $\theta_j$, we should
clearly sample from $\Pi_j$ to learn its value and continue learning until there is enough
evidence to reject the hypothesis $\mu(\theta_j) \geq M$, at which point we should switch to sampling
from $\Pi_\star$ (which does not add any information about $\mu(\theta_j)$). The left hand side of (2.6)
represents the expected discounted reward for such strategy with an optimal choice of the
stopping rule. The right hand side of (2.6) represents the expected discounted reward of
the strategy that samples from $\Pi_\star$ at all stages.

Using diffusion approximations to analyze the optimal stopping problem (2.6) in the
case where $f(x; \theta)$ belongs to an exponential family of densities

$$f(x; \theta) = \exp(\theta x - \psi(\theta)) \quad (2.8)$$

with respect to some measure $\nu$, Chang and Lai [8] found simple asymptotic approx-
imations to Gittins indices as the discount factor $\beta$ approaches 1. These asymptotic
approximations turn out to be certain upper confidence bounds for the $\mu(\theta_j)$. Moreover,
it is natural to expect that replacing $1 - \beta$ by $1/N$ in these approximations should
also lead to asymptotically optimal solutions to the finite-horizon problem of maximizing

$$\int \cdots \int \hat{E}_\theta S_N dG_1(\theta_1) \cdots dG_k(\theta_k).$$

Lai [9] has shown that this is indeed the case, assuming
the $G_i$ to put all its mass on an open interval $A \subset \Theta \hat{=} \{ \theta : \int e^{\theta x} d\nu(x) < \infty \}$ with end-
points ($-\infty \leq a_1 < a_2 \leq \infty$) such that

$$\inf_{a_1 - r < \theta < a_2 + r} \psi''(\theta) > 0, \quad \sup_{a_1 - r < \theta < a_2 - r} \psi''(\theta) < \infty \quad (2.9)$$

and $\psi''$ is uniformly continuous on $(a_1 - r, a_2 + r)$ for some $r > 0$.

In particular, if $Y_1, Y_2, \ldots$ are normally distributed, then $\Theta$ is the entire real line and we
can take $A = \Theta$.

More generally, Lai [9] considers general prior distributions $H$ on $A^k$ (not necessarily
assuming independent $\theta_1, \ldots, \theta_k$) and shows that index rules in which the dynamic allo-
cation indices are certain upper confidence bounds for $\theta_j$ (noting that $\mu(\theta) = \psi'(\theta)$ is an
increasing function for the exponential family (2.8)) provide asymptotically optimal solutions
to the finite-horizon problem of maximizing $\int E_\theta S_N dH(\theta)$. Moreover, these index
rules can be regarded as natural perturbations of obvious certainty-equivalence rules to
ensure that there is enough information to learn about all the population parameters, as
will be explained below. For fixed $j$, based on the observations $Y_1, \ldots, Y_n$ from $\Pi_j$, the
maximum likelihood estimator of $\theta_j$, known to belong to $A$, is given by

$$\hat{\theta}_{j,n} = \mu^{-1}(S_n/n), \quad \text{if } \mu(a_1) \leq S_n/n \leq \mu(a_2),$$
\[ = a_2, \text{ if } S_n/n > \mu(a_2), \]
\[ = a_1, \text{ if } S_n/n < \mu(a_1), \]  

where \( S_n = Y_1 + \cdots + Y_n \). Let \( N > 1 \) and let \( g \) be a nonnegative function on \((0, \infty)\) satisfying the following assumptions:

\[ \sup_{t \geq a} \frac{g(t)}{t} < \infty, \text{ for all } a > 0, \]  
\[ g(t) \sim \log t^{-1}, \text{ as } t \to 0, \]  
\[ g(t) \geq \log t^{-1} + \xi \log \log t^{-1}, \text{ as } t \to 0, \]

for some \( \xi \). Based on the \( n \) observations \( Y_1, \cdots, Y_n \), define the “upper confidence bound”

\[ U_{j,n}(g, N) = \inf \{ \theta \in A : \theta \geq \hat{\theta}_{j,n} \text{ and } I(\hat{\theta}_{j,n}, \theta) \geq n^{-1}g(n/N) \} \]

(\( \inf \theta = \infty \)), where \( I(\theta, \lambda) \) is the Kullback-Leibler information number given by

\[ I(\theta, \lambda) = (\theta - \lambda)\psi'(\theta) - (\psi(\theta) - \psi(\lambda)). \]

A particular choice of \( g \) satisfying (2.11)-(2.13) derived in [9] from an approximating continuous-time optimal stopping problem for Brownian motion is \( g_0(t) = h_0^2(t)/2t \), where

\[ h_0(t) = (t^{-1} - 1)^{1/2} \{0.63883 - 0.40258(t^{-1} - 1)\}, \text{ if } 0.86 < t \leq 1, \]
\[ = -0.5759t^2 + 0.2987t + 0.4034, \text{ if } 0.28 < t \leq 0.86, \]
\[ = -1.58137t + 1.53343t^{1/2} + 0.07327, \text{ if } 0.01 < t \leq 0.28, \]
\[ = \{t[2\log t^{-1} - \log \log t^{-1} - \log 16\pi + 0.99232 \exp(-0.03812t^{-1/2})]\}^{1/2}, \]

if \( 0 < t \leq 0.01 \).

These upper bounds \( U_{j,n}(g, N) \), which we shall simply denote by \( U_{j,n} \) for given \( N \) and \( g \), will serve as dynamic allocation indices in the index rule below.

If the values \( \theta_1, \cdots, \theta_k \) were known, the optimal rule would obviously be to sample from the population with the largest \( \theta_j \). In ignorance of \( \theta_1, \cdots, \theta_k \), the certainty-equivalence approach is to estimate them at stage \( n \) with the estimators \( \hat{\theta}_{1,n}, \cdots, \hat{\theta}_{k,n} \) and to sample at stage \( n + 1 \) from the population with the largest \( \hat{\theta}_{j,n} \). The difficulty with this “play-the-leader” rule is that we may have sampled too little from an apparently inferior population to get a reliable estimate of its parameter and may thereby miss the actually
superior population. Instead of sampling at stage \( n + 1 \) from the population with the largest \( \hat{\theta}_{j,n}(x) \), Lai [9] proposes the following simple modification:

\[
\text{Sample at stage } n + 1 \text{ from the population } \Pi_j \text{ with the largest upper confidence bound } U_{j,n}(x),
\]

where \( U_{j,r} \) is defined in (2.14) and \( n \geq k \). (During the first \( k \) stages, we sample once from each population.) To explain the heuristic idea behind this approach, we first note that the upper confidence bound \( U_{j,r} \) inflates the estimator \( \hat{\theta}_{j,r} \) by an amount which decreases with the number \( r \) of observations already taken from the population. Thus, \( U_{j,r} \) depends not only on the estimator \( \hat{\theta}_{j,r} \) but also on the sample size \( r \), and comparing the \( k \) populations on the basis of \( U_{j,r}(x) \) involves not only the parameter estimates but also the sample sizes of all populations.

EXAMPLE. Suppose that \( Y_1, Y_2, \ldots \) are i.i.d. observations from a normal population \( \Pi_1 \) with mean \( \theta \) and variance 1. Here \( \mu(\theta) = \theta \), \( I(\theta, \lambda) = (\theta - \lambda)^2 / 2 \) and the confidence bound (2.14) reduces to

\[
U_{1,n}(g, N) = S_n / n + (2n^{-1}g(n/N))^{1/2}.
\]

For an example of \( g \) satisfying (2.11)-(2.13), consider a nonnegative continuous function on \((0,1]\) having the asymptotic expansion

\[
g(t) = \log t^{-1} - \frac{1}{2} \log \log t^{-1} - \frac{1}{2} \log 16\pi + o(1), \quad \text{as } t \to 0,
\]

The asymptotic expansion (2.19) first arose in the following one-armed bandit problem considered by Chernoff and Ray [10] and by Chernoff [11]. Suppose that an experimenter can choose at each stage \( n(\leq N) \) between sampling from the normal population \( \Pi_1 \) with unknown mean \( \theta \) and sampling from another normal population \( \Pi_2 \) with known mean 0. Assuming a normal prior on \( \theta \), the Bayes procedure (to maximize the expected sum of \( N \) observations) samples from \( \Pi_1 \) until stage

\[
T^* = \inf\{n \leq N : n^{-1} \sum_{i=1}^{n} Y_i + a_{n,N} \leq 0\},
\]

and then takes the remaining \( N - T^* \) observations from \( \Pi_2 \), where \( a_{n,N} \) are positive constants. Writing

\[
t = n/N, \quad w(t) = (Y_1 + \cdots + Y_n) / N^{1/2}, \quad \delta = \theta N^{1/2},
\]

8
and treating $0 < t \leq 1$ as a continuous variable for large $N$, we can approximate the Bayes stopping time (2.20) by $N\tau(h)$, where $\tau(h) = \inf\{t \in (0, 1) : w(t) + h(t) \leq 0\}$ is the optimal stopping rule in the following continuous-time stopping problem: Assuming a flat prior for the drift coefficient $\delta$ of a Wiener process $w(t)$, find the stopping rule $\tau \leq 1$ to maximize $\int_{-\infty}^{\infty} E_\delta(\delta\tau)d\delta$. Using an asymptotic analysis of the free boundary problem associated with this continuous-time optimal stopping problem, Chernoff and Ray [10] found that as $t \downarrow 0$

$$h(t) = \left\{2t \log t^{-1} - \frac{1}{2} \log \log t^{-1} - \frac{1}{2} \log 16\pi + o(1)\right\}^{1/2}.$$ 

Therefore, letting $h(t)/t^{1/2} = (2g(t))^{1/2}$, $g(t)$ satisfies the asymptotic expansion (2.19). From (2.21), it follows that the $a_{n,N}$ in (2.20) can be approximated by $(2n^{-1}g(n/N))^{1/2}$. With this approximation, the Bayes procedure for the one-armed bandit problem can be described in the form:

At stage $n + 1$ sample from $\Pi_1$ or $\Pi_2$ according as

$$U_{1,T_{1n}(1)}(g,N) > 0 \text{ or } U_{1,T_{1n}(1)}(g,N) \leq 0.$$ 

The boundary (2.16) is obtained by approximating the optimal stopping boundary $h(t)$ that has been evaluated numerically, cf. [9].

For $j = 1, \ldots, k$, let $\tilde{\theta}_j = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{j-1}, \tilde{\theta}_{j+1}, \ldots, \tilde{\theta}_k)$ and let $\tilde{\theta}_j^* = \max_{i \neq j} \tilde{\theta}_i$. For a prior distribution $H$ of $\theta$, let $H_j$ denote the marginal distribution of the $(k - 1)$-dimensional vector $\tilde{\theta}_j$ and let $H^{(j)}(\theta|\tilde{\theta}_j) = P\{\theta_j \leq \theta|\tilde{\theta}_j\}$ denote the conditional distribution function of $\theta_j$ given $\tilde{\theta}_j$. Suppose that the prior distribution $H$ on $A_k$ satisfies the following four conditions for some $\rho > 0$ and every $j = 1, \ldots, k$:

(a) $\int |\tilde{\theta}_j|dH(\tilde{\theta}) < \infty$;

(b) for every fixed $\tilde{\theta}_j \in A_{k-1}$, $H^{(j)}(\theta|\tilde{\theta}_j)$ has a positive continuous derivative $h_j(\theta; \tilde{\theta}_j)$ for $\theta \in (\tilde{\theta}_j^* - \rho, \tilde{\theta}_j^* + \rho) \cap A$;

(c) $\int_{A_{k-1}} \sup_{\theta \in (\theta_j^* - \rho, \theta_j^* + \rho) \cap A} h_j(\theta; \tilde{\theta}_j) dH_j(\tilde{\theta}_j) < \infty$;

(d) $h_j(\theta; \tilde{\theta}_j)/h_j(\theta_j^*, \tilde{\theta}_j) \to 1$ as $\theta \to \theta_j^*$, uniformly in $\tilde{\theta}_j \in B^{k-1}$,

for every compact subset $B$ of $A$. Under the mild assumptions (2.11)-(2.13) on $g$, Lai [9] has established that the index rule (2.17) has the following asymptotically minimal order of magnitude for the Bayes regret $\int_{A_k} R_N(\theta)dH(\theta)$:

$$\left\{\frac{1}{2} \sum_{j=1}^{k} \int_{A_{k-1}} h_j(\theta_j^*; \tilde{\theta}_j) dH_j(\tilde{\theta}_j) \right\}(\log N)^2 \sim \inf_{\phi} \int_{A_k} R_N(\theta)dH(\theta) \text{ as } N \to \infty, \quad 23$$

9
where \( \inf_{\varphi} \) is taken over all adaptive allocation rules \( \varphi \) and \( R_N(\varphi) \) is defined in (2.3).

Replacing \( N \) by \( 1/(1 - \beta) \) in the index rule (2.17) also gives an asymptotically optimal solution to the discounted Bayes problem of maximizing \( \int_{A^*} E_\theta \left( \sum_{i=1}^{\infty} \beta^{i-1} x_i \right) dH(\theta) \) as \( \beta \uparrow 1 \), cf. [8]. The index rule (2.17) also provides an asymptotically optimal solution to the following version of the multi-armed bandit problem first considered by Robbins [12] without assuming a finite horizon \( N \) or discount factor \( \beta \): Find an adaptive allocation rule such that \( n^{-1} E_\theta S_n \) approaches \( \mu^*(\theta) = \max_{1 \leq i \leq k} \mu(\theta_i) \) as rapidly as possible as \( n \to \infty \) for all \( \theta \in \Theta^k \). In particular, for the case \( k = 2 \), Robbins [12] introduced the following class of simple adaptive allocation rules satisfying

\[
\lim_{n \to \infty} n^{-1} E_\theta S_n = \mu^*(\theta) \quad \text{for all} \quad \theta \in \Theta^k.
\] (2.24)

Let \( a_1 = 1 < a_2 < \cdots \) and \( b_1 = 2 < b_2 \leq \cdots \) be two fixed, disjoint, increasing sequences of positive integers such that \( a_n/n \to \infty \) and \( b_n/n \to \infty \) as \( n \to \infty \). At stage \( n \), sample from \( \Pi_1 \) if \( n \in \{a_1, a_2, \cdots\} \), sample from \( \Pi_2 \) if \( n \in \{b_1, b_2, \cdots\} \), and if \( n \not\in \{a_1, a_2, \cdots, b_1, b_2, \cdots\} \), sample from \( \Pi_1 \) or \( \Pi_2 \) according as the mean of all previous observations from \( \Pi_1 \) exceeds or does not exceed the mean of all previous observations from \( \Pi_2 \). From the strong law of large numbers, it follows easily that this procedure attains (2.24).

The sequences \( a_n \) and \( b_n \) above are assumed to be prescribed in advance, and a natural question is how to choose them so that \( n^{-1} E_\theta S_n \) approaches \( \mu^*(\theta) \) (or equivalently, \( n^{-1} R_n(\theta) \) approaches 0) as rapidly as possible. To answer this question, Lai and Robbins [13] first came up with the following asymptotic lower bound for the regret of adaptive allocation rules so that the phrase "as rapidly as possible" can be given a more precise meaning. It is shown in [13] that for every allocation rule satisfying \( R_n(\theta) = o(n^a) \) for all \( a > 0 \) and \( \theta \in \Theta^k \),

\[
\liminf_{n \to \infty} R_n(\theta)/\log n \geq \sum_{j: \mu(\theta_j) < \mu^*(\theta)} (\mu^*(\theta) - \mu(\theta_j))/I(\theta_j, \lambda)
\] (2.25)

for all \( \theta_1, \cdots, \theta_k, \lambda \in \Theta \) such that \( \mu(\lambda) > \mu^*(\theta) \), where

\[
I(\theta, \lambda) = \int_{-\infty}^{\infty} \{ \log(f(x; \theta)/f(x; \lambda)) \} f(x; \theta) d\nu(x).
\] (2.26)

For the exponential family of densities (2.8), consider the allocation rule \( \varphi(\theta) \) that samples at stage \( n + 1(> k) \) from the population \( \Pi_j \) with the largest upper confidence bound
\[ R_n(\theta) \sim \left\{ \sum_{j: \theta_j < \theta^*} (\mu(\theta^*) - \mu(\theta_j))/I(\theta_j, \theta^*) \right\} \log n, \quad (2.27) \]

where \( \theta^* = \max_{1 \leq j \leq k} \theta_j \), cf. Theorem 4 of [14]. Note that \( \mu = \psi' \) is an increasing function and that (2.26) reduces to (2.15) in this case. Hence the index rule \( \varphi(g) \) attains the asymptotic lower bound (2.25) for the regret. Such allocation rules are called “asymptotically efficient” by Lai and Robbins [13] who also show how to construct upper confidence bounds to form asymptotically efficient index rules for several other parametric families of distributions. Their construction has been extended from the i.i.d. to Markov sequences by Anantharaman, Varaiya and Walrand [15] and to the setting with switching costs by Agrawal, Hedge and Teneketzis [16].

The asymptotic lower bound (2.25) can be interpreted as (and follows from) that the minimal expected sample size (information) from an inferior population \( \Pi_j \) is at least as large as \((1 + o(1))(\log n)/I(\theta_j, \lambda)\) if the rule does not perform too badly at every \( \theta \), cf. [13]. In other words, we need to sample at least that many observations on the average from the inferior population \( \Pi_j \) for the rule to have a regret \( R_n(\theta) = o(n^a) \) for every fixed \( \theta \) and \( a > 0 \). In particular, for the exponential family (2.8), this implies that for an inferior population \( \Pi_j \) we need approximately \((\log n)/I(\theta_j, \theta^*)\) observations from \( \Pi_j \) to be reasonably confident that it is not a contender. Of course, since \( \theta_j \) and \( \theta^* \) are unknown, we have to see if there is enough information to adequately estimate \( I(\theta_j, \theta^*) \) \( \sim \) Lai and Robbins [13] found that replacing the maximum likelihood estimate \( \hat{\theta}_{i, T_n(i)} \) for each population parameter by a perturbation (“upper confidence bound”)

\[ U_{i, T_n(i)}(g, n) = \inf\{ \theta : \theta \geq \hat{\theta}_{i, T_n(i)} \text{ and } T_n(i)I(\hat{\theta}_{i, T_n(i)}, \theta) \geq g(T_n(i)/n) \} \]

provides a simple way to get around these difficulties. Such upper confidence bounds turn out also to provide approximations to Gittins indices for the discounted Bayes problem and also to give asymptotically optimal solutions to the finite-horizon Bayes problem, as explained above.
3. LEAST SQUARES ADAPTIVE PREDICTORS AND MULTI-PERIOD ADAPTIVE CONTROL IN STOCHASTIC REGRESSION MODELS

The “multi-period control problem under uncertainty” in the econometrics literature is to choose successive inputs $u_1, \ldots, u_n$ in the linear regression model

$$y_i = \alpha + \beta u_i + \epsilon_i \quad (i = 1, \ldots, n), \quad (3.1)$$

where $\alpha$ and $\beta$ are unknown parameters and the random disturbances $\epsilon_1, \epsilon_2, \ldots$ are i.i.d. with mean 0 and variance $\sigma^2$, so that the outputs $y_1, \ldots, y_n$ are as close as possible in some sense to a given target value $y^*$. A Bayesian formulation is often used so that the problem becomes that of minimizing

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E_{\alpha, \beta} [\sum_{i=1}^{n} (y_i - y^*)^2] d\pi(\alpha, \beta) = n\sigma^2 + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \beta^2 E_{\alpha, \beta} [\sum_{i=1}^{n} (u_i - \theta)^2] d\pi(\alpha, \beta), \quad (3.2)$$

where $\pi$ is a prior distribution of the unknown parameters $\alpha, \beta$ and $\theta = (y^* - \alpha)/\beta$, cf. [17], [18], [19]. The $\epsilon_i$ are often assumed to be normally distributed, and (3.2) can in principle be minimized by using dynamic programming. However, because of the complexities in the numerical solution of the dynamic programming equations and the analytical difficulties in studying the Bayes rules, not much is known about their performance.

A first departure from the Bayesian approach was due to Aoki [20]. Assuming that the sign of $\beta$ is known, say $\beta > 0$, he proposed the use of a stochastic approximation scheme of the form

$$u_{t+1} = u_t - a_t (y_t - y^*), \quad (3.3)$$

where $\{a_t\}$ is a sequence of positive constants such that $\sum_1^\infty a_t = \infty$ and $\sum_1^\infty a_t^2 < \infty$. This approach has the property that $u_t \to \theta$ a.s. (almost surely). Another non-Bayesian approach is the certainty-equivalence rule proposed by Anderson and Taylor [21]. If $\alpha$ and $\beta(\neq 0)$ are both known, the optimal choice of $u_t$ is clearly at $\theta = (y^* - \alpha)/\beta$. Without assuming $\alpha$ and $\beta$ to be known, suppose that bounds $K_1, K_2$ are known such that $K_1 < \theta < K_2$. Assuming the $\epsilon_i$ to be normally distributed, the maximum likelihood estimator of $\theta$ at stage $t \geq 2$ is

$$\hat{\theta}_t = K_2 \land \{\hat{\beta}_t^{-1} (y^* - \hat{\alpha}_t) \lor K_1\}, \quad (3.4)$$

where $\hat{\beta}_t = (\sum_{1}^{t} (u_i - \bar{u}_t) y_i) / (\sum_{1}^{t} (u_i - \bar{u}_t)^2)$ and $\hat{\alpha}_t = \bar{y}_t - \hat{\beta}_t \bar{u}_t$ are the least squares estimates of $\beta$ and $\alpha$, and $\bar{u}_t = t^{-1} \sum_{1}^{t} u_i$, $\land$ and $\lor$ denote minimum and maximum.
respectively. The initial values $u_1$ and $u_2$ are distinct but otherwise arbitrary numbers between $K_1$ and $K_2$, and for $t \geq 2$, the certainty-equivalence rule sets $u_{t+1} = \bar{\theta}_t$.

Despite its simplicity for implementation, the certainty-equivalence rule is difficult to analyze. A basic difficulty with this approach is that although $\bar{\theta}_t$ may conceivably represent one's best current guess of $\theta$, how good the guess is depends on how the inputs $u_1, \ldots, u_t$ are chosen. In particular, if the inputs $u_i = \bar{\theta}_{i-1}, 1 \leq i \leq t$, tend to cluster around their mean $\bar{u}_t$, then there may not be enough information to give a reliable estimate $\bar{\theta}_t$, even though $\bar{\theta}_t$ may well be one's closest possible guess of $\theta$ at stage $t$. Lai and Robbins [22] subsequently exhibited an event $\Omega_0$ with positive probability on which $\bar{\theta}_t$ is of the wrong sign for all $t$ and $u_t$ does not converge to $\theta$ as $t \to \infty$. There is, therefore, again a dilemma between the desire to set the current input as close to $\theta$ as the data suggest ("control") and the need to set the inputs in a wide enough scatter so as to provide a reliable estimate of $\theta$ ("information"). To resolve this dilemma, we have first to find out how reliable an estimate is really needed and how much information (design scatter) is required to produce such an estimate. A benchmark that gives valuable clues to these questions is provided by studying the "fictitious" situation that assumes $\beta$ to be known.

Suppose that $\beta \neq 0$ is known. Rewriting the regression model (3.1) as

$$\frac{(y_i - y^*)}{\beta} = u_i - \theta + \epsilon_i/\beta,$$

(3.5)

the maximum likelihood estimator of $\theta$ at stage $t$ is

$$\theta^*_t = \bar{u}_t - \beta^{-1}(\bar{y}_t - y^*) = \theta - \beta^{-1}\bar{\epsilon}_t.$$ 

Thus, irrespective of how the inputs $x_i$ are chosen,

$$\theta^*_t - \theta = -\beta^{-1}\bar{\epsilon}_t \sim N(0, \sigma^2/\beta^2).$$

(3.6)

Hence, in the present setting, there is no conflict between estimation and control. In particular, if we set $u_{t+1}$ to be our best current guess $\theta^*_t$ of $\theta$, i.e.,

$$u_{t+1} = \theta^*_t = \bar{u}_t - \beta^{-1}(\bar{y}_t - y^*), \quad t \geq 1,$$

(3.7)

and let $u_1$ = initial guess of $\theta$, then Lai and Robbins [22] showed that $u_t \to \theta$ a.s. and

$$\beta^2 \sum_{t=1}^{n}(u_t - \theta)^2/\log n \to \sigma^2 \text{ a.s.},$$

(3.8)
\[ \beta^2 E_\theta \sum_{t=1}^{n} (u_t - \theta)^2 = \sigma^2 \log n + O(1). \] (3.9)

Let \( R_n = \beta^2 \sum_{t=1}^{n} (u_t - \theta)^2 \) denote the "regret" of an input sequence \( \{u_t : t \leq n\} \), and note that \( E_\theta \{\sum_{t=1}^{n} (y_i - y^*)^2\} = n \sigma^2 + E_\theta (R_n) \). Putting a prior distribution \( \pi \) on \( \theta \), the solution to the Bayes problem of minimizing \( \int_{-\infty}^{\infty} E_\theta (R_n) d\pi(\theta) \) is simply

\[ u_{t+1} = E(\theta|u_1, y_1, \ldots, u_t, y_t) = \text{posterior mean of } \theta; \] (3.10)

in particular, if \( \pi \) is normal, then (3.8) and (3.9) still hold for (3.10), cf. [23].

In this auxiliary problem that assumes knowledge of \( \beta \), we do not have to wrestle with the difficult issue of potential conflict between estimation and control. The auxiliary problem, which we can solve completely, sets a lower bound on the regret \( R_n \) that we hope to achieve even when \( \beta \) is unknown. Lai and Robbins [22], [24] proposed two simple methods to achieve this, under the assumption that the \( \epsilon_i \) are i.i.d. random variables (not necessarily normal) with mean 0 and variance \( \sigma^2 \).

The first method, considered in [24], is adaptive stochastic approximation. The stochastic approximation scheme (3.3) was originally introduced by Robbins and Monro [25] to estimate sequentially the solution \( \theta \), assumed unique, of the equation \( f(u) = y^* \), based on successive observations \( (u_i, y_i) \) from the regression model \( y_i = f(u_i) + \epsilon_i \), where \( f \) is an unknown function satisfying certain assumptions. In the setting of (3.1), \( f(u) = y^* + \beta (u - \theta) \) in which \( \beta \) and \( \theta \) are unknown parameters. If \( \beta(\neq 0) \) is known, then for the choice \( a_t = (\beta t)^{-1} \) in (3.3), the stochastic approximation recursion (3.3) is equivalent to the recursion (3.7), for which (3.8) and (3.9) hold, cf. [24]. Without assuming \( \beta \) to be known, Lai and Robbins [24] suggested using \( a_t = (b_t t)^{-1} \) in (3.3), where \( b_t = b_t(u_1, y_1, \ldots, u_t, y_t) \) is an estimate of \( \beta \) based on the data already observed, and showed that it is possible to choose \( b_t \) such that \( b_t \to \beta \) a.s., which in turn ensures that (3.8) and (3.9) still hold and that

\[ \sqrt{t} (u_t - \theta) \text{ has a limiting normal } N(0, \sigma^2/\beta^2) \text{ distribution as } t \to \infty. \] (3.11)

This approach was subsequently generalized by Wei [26] to the multivariate case in which \( \alpha, y_i, u_i \), and \( \epsilon_i \) in (3.1) are \( k \times 1 \) vectors and \( \beta \) is a \( k \times k \) nonsingular matrix.

The second method, considered in [22], starts with the certainty-equivalence rule (3.7) that assumes \( \beta \) known and then uses a modified least squares estimate \( \hat{\beta}_t \) to replace \( \beta \) in (3.7), leading to the certainty-equivalence rule

\[ u_{t+1} = \bar{u}_t - \hat{\beta}_t^{-1}(\bar{y}_t - y^*), \quad t \geq 1. \] (3.12)
Although this rule appears to be motivated by similar certainty-equivalence principles as in the Anderson-Taylor rule \( u_{t+1} = K_2 \land \{ \tilde{\beta}_t^{-1}(y^* - \tilde{\alpha}_t) \lor K_1 \} \) in (3.4), there are important differences. While the Anderson-Taylor rule applies the certainty-equivalence principle to the regression model (3.1), the rule (3.12) works with the reparametrized model \( y_i = y^* + \beta(u_i - \theta) + \epsilon_i \). Furthermore, while Anderson and Taylor assume known \emph{a priori} bounds \( K_1 < \theta < K_2 \), it is assumed in [22] that upper and lower bounds of the same sign for \( \beta \) are known, say \( 0 < b < \beta < B \). Therefore [22] uses the following truncated version of the least squares estimator of \( \beta \):

\[
\tilde{\beta}_t = B \land \left\{ \left[ \sum_{i=1}^t (u_i - \bar{u}_t)y_i \right] / \left[ \sum_{i=1}^t (u_i - \bar{u}_t)^2 \right] \lor b \right\} \text{ if } \sum_{i=1}^t (u_i - \bar{u}_t)^2 \geq (2B^2)^{-1}s_t^2 \log t,
\]

\[
= \tilde{\beta}_{t-1} \text{ otherwise,}
\]  

(3.13)

where \( s_t^2 = t^{-1} \sum_{i=1}^t (y_i - \bar{y}_t)^2 \) and the initial value \( \tilde{\beta}_1 \) is any constant between \( b \) and \( B \). With this choice of \( \tilde{\beta}_t \), define the certainty-equivalence rule (3.12). It is shown in [22] and [23] that (3.8), (3.9) and (3.11) still hold for the adaptive control rule (3.12)-(3.13).

To extend the above least squares certainty-equivalence approach to the multivariate case in which \( \alpha, y_i, u_i \) and \( \epsilon_i \) in (3.1) are \( k \times 1 \) vectors and \( \beta \) is a \( k \times k \) nonsingular matrix, Lai [27] again reparametrized (3.1) as \( y_i = y^* + \beta(u_i - \theta) + \epsilon_i \) and assumed (by considering \( y_i - y^* \) instead of \( y_i \)) that \( y^* = 0 \). In fact, [27] considers a more general stochastic regression model of the form

\[
Y_i = B(U_i - CX_i) + \epsilon_i,
\]

(3.14)

where

(i) \( Y_i \) is a \( k \times 1 \) vector of outputs at stage \( i \);

(ii) \( C \) is a \( k \times h \) matrix and \( B \) is a \( k \times k \) nonsingular matrix of parameters;

(iii) \( \epsilon_i \) is a \( k \times 1 \) vector of random disturbances such that \( \{ \epsilon_i \} \) is a martingale difference sequence with respect to an increasing sequence of \( \sigma \)-fields \( \{ \mathcal{F}_i \} \) (i.e., \( \epsilon_i \) is \( \mathcal{F}_i \)-measurable and \( E(\epsilon_i | \mathcal{F}_{i-1}) = 0 \) for all \( i \));

(iv) \( U_i \) is a \( k \times 1 \) vector of inputs depending on the current and past observations \( X_1, U_1, Y_1, \cdots, X_{i-1}, U_{i-1}, Y_{i-1}, X_i \);

(v) \( X_i \) is an \( \mathcal{F}_{i-1} \)-measurable \( h \times 1 \) vector of regressors.
The random disturbances $\epsilon_i$ are assumed to satisfy the condition

$$\sup_i E(||\epsilon_i||^\gamma |\mathcal{F}_{i-1}) < \infty \text{ a.s. for some } \gamma > 2,$$

(3.15)

while the regressors $X_i$ are assumed to satisfy the condition

$$X_n = O(1) \text{ and } \lambda_{\min}(\sum_{i=1}^n X_iX_i^T) \to \infty \text{ a.s.},$$

(3.16)

where $\lambda_{\min}$ and $\lambda_{\max}$ denote, respectively, the minimum and maximum eigenvalues of a symmetric matrix. In particular, the case $h = 1 = X_i$ and $C = \theta$ reduces to the model $Y_i = B(U_i - \theta) + \epsilon_i$ considered above.

If $B$ is known, then we can write (3.14) as

$$U_i - B^{-1}Y_i = CX_i - B^{-1}\epsilon_i.$$  

(3.17)

The least squares estimate $C_t$ of $C$ based on $X_1, U_1, Y_1, \cdots, X_t, U_t, Y_t$ can be expressed in the following recursive form:

$$C_t = C_{t-1} + (U_t - B^{-1}Y_t - C_{t-1}X_t)X_t^TP_t,$$

(3.18a)

$$P_t^{-1} = P_{t-1}^{-1} + X_tX_t^T.$$  

(3.18b)

Note that (3.18b) can be expressed in the computationally more convenient form

$$P_t = P_{t-1} - P_{t-1}X_tX_t^T P_{t-1}(1 + X_t^TP_{t-1}X_t).$$

Now suppose that $B$ is unknown but that there exist nonsingular $k \times k$ matrices $B_t$ such that $B_t$ is $\mathcal{F}_t$-measurable and

$$B_t \rightarrow B \text{ w.p.1.}$$

(3.19)

Replacing $B$ by $B_{t-1}$ in (3.18) leads to

$$C_t = C_{t-1} + (U_t - B_{t-1}^{-1}Y_t - C_{t-1}X_t)X_t^TP_t,$$

(3.20a)

$$P_t^{-1} = P_{t-1}^{-1} + X_tX_t^T.$$  

(3.20b)

If $B$ and $C$ are known, then $E(Y_i|\mathcal{F}_{i-1}) = B(U_i - CX_i)$ is the minimum variance predictor of $Y_i$ and the prediction error $Y_i - E(Y_i|\mathcal{F}_{i-1})$ is equal to $\epsilon_i$. When $B$ and $C$ are unknown, replacing $B$ by $B_{t-1}$ and $C$ by $C_{t-1}$ leads to the adaptive predictor $B_{t-1}(U_i -$
with prediction error \( \epsilon_i = Y_i - B_{i-1}(U_i - C_{i-1}X_i) \). Under the assumptions (3.15), (3.16) and (3.19), the cumulative square difference between \( \epsilon_i \) and \( \epsilon_i \) has been shown in [27] to satisfy the basic inequality

\[
\sum_{i=1}^{n} \|\epsilon_i - \epsilon_i\|^2 + \text{tr}\{B(C_n - C)P_n^{-1}(C_n - C)^TB^T\} \
\leq (\sigma^2 + o(1)) \log \det(\sum_{i=1}^{n} X_iX_i^T) + o(\sum_{i=1}^{n} \|U_i - CX_i\|^2) \text{a.s.,}
\]

(3.21)

where \( \sigma^2 = \lim \sup_{i \to \infty} E(\|\epsilon_i\|^2|\mathcal{F}_{i-1}) \).

The inequality (3.21) provides important clues for the construction of asymptotically efficient certainty-equivalence-type adaptive control laws for the regulation problem of choosing the inputs \( U_1, U_2, \ldots \) sequentially so that \( \sum_{i=1}^{n} \|Y_i\|^2 \) is minimized in some sense in ignorance of the parameter matrices \( B \) and \( C \). Minimizing \( E(\sum_{i=1}^{n} \|Y_i\|^2) \) is equivalent to minimizing \( E(R_n) \), where

\[
R_n = \sum_{i=1}^{n} \|B(U_i - CX_i)\|^2 = \sum_{i=1}^{n} \|Y_i - \epsilon_i\|^2
\]

(3.22)

Clearly if \( C \) is known, then the \( \mathcal{F}_{i-1} \)-measurable input \( U_i = CX_i \) (which is equivalent to \( E(Y_i|\mathcal{F}_{i-1}) = 0 \)) minimizes \( E\|Y_i\|^2 \), and the output of this rule is \( \epsilon_i \). The certainty-equivalence rule \( U_i = C_{i-1}X_i \) has output \( Y_i = \epsilon_i \), and therefore its regret is \( R_n = \sum_{i=1}^{n} \|\epsilon_i - \epsilon_i\|^2 \), a bound for which is provided by (3.21).

To apply (3.21), we have to construct estimates \( B_i \) such that \( B_i \to B \) a.s. A simple way of doing this is to apply the method of least squares to “well-excited” blocks of input-output data. To generate such data, we introduce white-noise probing inputs at certain prespecified times to ensure that there is enough information for estimating \( B \). Let

\[
\tau^* = \inf \left\{ i : \lambda_{\text{min}} \left[ \sum_{i=1}^{i} \left( \begin{array}{c} X_i \\ U_i \end{array} \right) \left( \begin{array}{c} X_i^T \\ U_i^T \end{array} \right) \right] \geq \eta \quad \text{and} \quad \left[ \sum_{i=1}^{i} Y_i(\left( \begin{array}{c} X_i \\ U_i \end{array} \right) \left( \begin{array}{c} X_i^T \\ U_i^T \end{array} \right) \right] \left[ \sum_{i=1}^{i} \left( \begin{array}{c} X_i \\ U_i \end{array} \right) \left( \begin{array}{c} X_i^T \\ U_i^T \end{array} \right) \right]^{-1} \text{is nonsingular} \right\}.
\]

(3.23)

where \( \eta \) is a positive constant, and take integers \( 1 < \nu_1 < \nu_2 < \cdots \) Let

\[
J = \{1, \cdots, \tau^*\} \cup \{\tau^* + \nu_1, \tau^* + \nu_2, \cdots\}
\]

(3.24)
be the set of stages when probing inputs are to be introduced. Thus, at stage \( t \in J \), the input

\[
U_t \text{ is a random vector independent of } X_1, U_1, Y_1, \cdots, \\
X_{t-1}, U_{t-1}, Y_{t-1}, X_t \text{ such that the components of } U_t \text{ are independently distributed with mean 0, variance } c, \text{ and } \|U_t\| \leq M,
\]

where \( c \) and \( M \) are positive constants. Note that \( \tau^* < \infty \) a.s. by (3.14), (3.16), (3.25) and the fact that \( B \) is nonsingular. Rewriting (3.14) as \( Y_i = AX_i + BU_i + \epsilon_i \), where \( A = -BC \), let

\[
(\tilde{A}_\nu; \tilde{B}_\nu) = \left\{ \sum_{i \in J} Y_i (X_i^T, U_i^T) \right\} \left\{ \sum_{i \in J} (X_i^T, U_i^T) (X_i^T, U_i^T) \right\}^{-1}, \nu \geq \tau^*,
\]

be the least squares estimate of \((A; B)\) based only on observations at stages \( t \in J \) with \( t \leq \nu \). Then \( \tilde{B}_\nu \to B \) a.s. In fact, letting \( m_\nu \) denote the number of elements in the set \( \{ t \in J : t \leq \nu \} \) and letting \( \tilde{B}_\nu(i,j) \) and \( B(i,j) \) denote the \((i,j)\) elements of \( \tilde{B}_\nu \) and \( B \), Lemma 1 of [27] shows that

\[
\max_{1 \leq i,j \leq k} |\tilde{B}_\nu(i,j) - B(i,j)| = O(m_\nu^{-1/2}(\log m_\nu)^{1/2}) \text{ a.s.}
\]

Suppose that the \( \nu_i \) in (3.24) are so chosen that \( i = o(\log \nu_i) \); for example, take \( \nu_i = \exp((1 + o(1))i^\delta) \) with \( \delta > 1 \). Then for \( \nu_i \leq \nu < \nu_{i+1}, m_\nu = \tau^* + i = o(\log \nu) \). Therefore the number \( m_\nu \) of probing inputs up to stage \( \nu \) diverges to \( \infty \) but within the order of \( o(\log \nu) \) as \( \nu \to \infty \), so the contribution of these probing inputs to the regret (3.22) is of the order \( o(\log n) \) a.s. Define the input \( U_t \) at stage \( t \in J \) by (3.25) ("probing input"), and at stage \( t \notin J \) define \( U_t = C_{t-1} X_t \) (certainty equivalence). For this certainty-equivalence rule with forced learning, \( \sum_{i \in J, i \leq n} \|U_i - CX_i\|^2 = o(\log n) \) a.s., and it follows from (3.21) that

\[
\sum_{i \notin J, i \leq n} \|B(U_i - CX_i)\|^2 = \sum_{i \notin J, i \leq n} \|Y_i - \epsilon_i\|^2 = \sum_{i \notin J, i \leq n} \|\epsilon_i - \epsilon_i\|^2 \\
\leq (\sigma^2 + o(1)) h \log n + o(\sum_{i=1}^n \|U_i - CX_i\|^2) \text{ a.s.},
\]

noting that \( \log \det(\sum_{i=1}^n X_i X_i^T) \leq h(1 + o(1)) \log n \) a.s. by (3.16). Hence

\[
R_n = \sum_{i=1}^n \|B(U_i - CX_i)\|^2 \leq (\sigma^2 + o(1)) h \log n \text{ a.s.}
\]
Note that in the special case \( h = 1 \equiv X_i \), the right hand side of (3.28) is \((1 + o(1))\sigma^2\log n\), which agrees with the asymptotically minimal order (3.8) for the regret for the model (3.1). In fact, for the multivariate model (3.14), a similar Bayesian argument assuming \( B \) to be known shows that the right hand side of (3.28) is also asymptotically minimal under certain additional assumptions, cf. [27].

4. ASYMPTOTICALLY EFFICIENT CERTAINTY-EQUIVALENCE-TYPE ALGORITHMS IN ADAPTIVE CONTROL OF ARMAX SYSTEMS

A widely used stochastic model in the time series and stochastic control literature is the ARMAX system (autoregressive moving average system with exogenous inputs) defined by the linear difference equation

\[
A(q^{-1})y_n = B(q^{-1})u_{n-d} + C(q^{-1})\epsilon_n,
\]

where \( \{y_n\}, \{u_n\} \) and \( \{\epsilon_n\} \) denote the output, input and disturbance sequences, respectively, \( d \geq 1 \) represents the delay and \( A(q^{-1}) = 1 + a_1q^{-1} + \cdots + a_pq^{-p}, \ B(q^{-1}) = b_1 + \cdots + b_kq^{-(k-1)}, \ C(q^{-1}) = 1 + c_1q^{-1} + \cdots + c_hq^{-h} \) are scalar polynomials in the backward shift operator \( q^{-1} \). Throughout the sequel we shall assume that \( b_1 \neq 0 \) and that \( \{\epsilon_n\} \) is a martingale difference sequence with respect to an increasing sequence of \( \sigma \) – fields \( \mathcal{F}_n \) such that \( E(\epsilon_n^2|\mathcal{F}_{n-1}) = \sigma^2 \) (nonrandom) > 0 for all \( n \).

Because of its theoretical interest and practical importance, the problem of determining the inputs \( u_n \), based on current and past observations \( y_n, y_{n-1}, u_{n-1}, \cdots \) (i.e., \( u_n \) is \( \mathcal{F}_n \)-measurable), to keep the outputs \( y_{n+d} \) as close as possible to certain target values \( y_{n+d}^* \) when the system parameters are not known in advance but have to be estimated "on-line" (i.e., during the operation of the system) has been one of the major topics in the subject of stochastic adaptive control. Let \( x_0 = (y_0, \cdots, y_{1-p}, u_0, \cdots, u_{2-d-k}, \epsilon_0, \cdots, \epsilon_{1-h})^T \) denote the "initial condition" of (1.1). The polynomial \( B(z) = b_1 + \cdots + b_kz^{k-1} \) will be called stable if all its zeros lie outside the unit circle. We shall also call two or more polynomials coprime (or relatively prime) if their greatest common divisors have degree 0.

In principle, given a probability distribution of the random sequence \( \{x_0, \epsilon_1, \epsilon_2, \cdots\} \) and a prior distribution \( \pi \) of the unknown parameter vector

\[
\theta = (-a_1, \cdots, -a_p, b_1, \cdots, b_k, c_1, \cdots, c_h)^T,
\]
we can use backward induction to solve the dynamic programming equations defining the inputs $u_1, \ldots, u_{N-d}$ that minimize

$$
\int E_{\theta} \left\{ \sum_{i=d+1}^{N} (y_i - y_i^*)^2 \right\} d\pi(\theta), \quad (4.3)
$$

for every given horizon $N$, where the $y_i^*$ are nonrandom target values for the outputs. Despite the analytical and computational difficulties in the implementation of the Bayesian approach, Bayesian analysis of some very simple examples has provided important insights into the structure of optimal control rules. In particular, Feldbaum [28] and subsequent authors have shown that Bayes rules have the “dual control” function of both probing the system for information about its parameters and trying to drive the outputs towards their target values. Åström [29, p. 478] has provided an interesting numerical example to illustrate this dual control effect in the adaptive control problem (4.3) for the simple ARX model $y_n - \alpha y_{n-1} = \beta u_{n-1} + \epsilon_n$, with i.i.d. zero-mean normal $\epsilon_n$, a normal prior distribution for $\beta$, known value $\alpha = 1$, and $N = 30$, $y_i^* \equiv 0$. The example shows that the Bayes rule takes relatively large and irregular control actions to probe the system when the Bayes estimate $\hat{\beta}_t$ of $\beta$ has poor precision, but when $\hat{\beta}_t$ has high precision the Bayes rule is well approximated by the “certainty-equivalence” rule $\hat{\beta}_t u_t = -y_t$.

Suppose that in Åström’s example, instead of assuming $\alpha$ to be known and $\beta$ to have a normal prior distribution, we assume $\beta \neq 0$ to be known and $\alpha$ to have a normal prior distribution. This problem turns out to be more tractable and can be used to derive lower bounds on the cost of Bayes rules. First note that for any input sequence $\{u_n\}$,

$$
\int_{-\infty}^{\infty} E_{\alpha} \left\{ \sum_{i=2}^{N} (y_i - \epsilon_i)^2 \right\} d\pi(\alpha) = \sum_{i=1}^{N-1} E \{ (\alpha y_i + \beta u_i)^2 \} \\
\geq \sum_{i=1}^{N-1} E \{ [\alpha - E(\alpha|y_1, u_1, \ldots, y_i, u_i)]^2 y_i^2 \}. \quad (4.4)
$$

In particular, for an input sequence $\{u_n\}$ that has the property

$$
\lim_{N \to \infty} N^{-1} \sum_{i=2}^{N} (y_i - \epsilon_i)^2 = 0 \text{ a.s. } [P_{\alpha}] \text{ for every } \alpha, \quad (4.5)
$$

it can be shown that

$$
\sum_{i=1}^{n} \{ \alpha - E(\alpha|y_1, u_1, \ldots, y_i, u_i) \}^2 y_i^2 \\
= (1 + o(1)) \sum_{i=1}^{n} \left\{ \sum_{i=2}^{i} y_{t-1} \epsilon_i / \sum_{i=2}^{i} y_{i-1} \right\}^2 y_i^2 + O(1) \sim \sigma^2 \log n \text{ a.s. } [P_{\alpha}], \quad (4.6)
$$
for every $\alpha$. From (4.4) and (4.6), it follows by an application of Fatou's lemma and Fubini's theorem that

$$\int_{-\infty}^{\infty} E_{\alpha}\{\sum_{i=2}^{N}(y_i - \epsilon_i)^2\} d\pi(\alpha) \geq (1 + o(1))\sigma^2\log N,$$

(4.7)

for all input sequences $\{u_n\}$ satisfying (4.6).

Lai [27] has extended the results (4.4), (4.6) and (4.7) above for the problem (4.3) with $y_i^* \equiv 0$ to general ARX models (4.1) (with $C(q^{-1}) = 1$) and unit delay $d = 1$. Specifically, assuming $b_1(\neq 0)$ to be known and putting a truncated normal prior distribution $\pi$ on $\lambda = b_1^{-1}(-a_1, \ldots, -a_p, b_2, \ldots, b_k)^T$, Lai [27] showed that in analogy with (4.7),

$$\int E_{\lambda}\{\sum_{i=2}^{N}(y_i - \epsilon_i)^2\} d\pi(\lambda) \geq (1 + o(1))\sigma^2(p + k - 1)\log N,$$

(4.8)

for all input sequences $\{u_n\}$ satisfying (4.5) and the additional growth condition that $u_n^2 = O(n^\delta)$ a.s. for some $0 < \delta < 1$. The truncated normal prior distribution $\pi$ in (4.8) is the restriction of a standard multivariate normal distribution to the $\lambda$-region defined by

$$A(z) \text{ and } B(z) \text{ are stable, and the polynomials } a_1z^{p-1} + \cdots + a_p \text{ and } z^{k-1}B(z^{-1}) \text{ are relatively prime.}$$

(4.9)

In the case of unit delay $d = 1$, if all the parameters of the system (1.1) are known, then the optimal controller chooses the input $u_t$ at stage $t$ so that $E(y_{t+1}|F_t) = y_{t+1}^*$, and its output at stage $t + 1$ is $y_{t+1}^* + \epsilon_{t+1}$. In view of this, Lai [27] defines the "regret" at stage $N$ of an input sequence $\{u_n\}$ to be

$$R_N = \sum_{i=2}^{N}\{y_i - (y_i^* + \epsilon_i)\}^2 = \sum_{i=2}^{N}\{E(y_i|F_{i-1}) - y_i^*\}^2.$$ 

(4.10)

Note that (4.8) above represents a lower bound for the expected regret in the regulation problem $y_i^* \equiv 0$ within a Bayesian framework. For general delay $d$, the definition (4.10) of "regret" can be extended to

$$R_N = \sum_{i=d+1}^{N}\{E(y_i|F_{i-d}) - y_i^*\} = \sum_{i=d+1}^{N}(y_i - y_i^* - \eta_i)^2,$$

(4.11)

where $\eta_i = y_i - E(y_i|F_{i-d})$. The regret $R_N$, which is 0 for the optimal controller assuming knowledge of all system parameters so that $u_t$ is determined by $E(y_{t+d}|F_t) = y_{t+d}^*$, can be
regarded as the cumulative cost up to stage \( N \) due to lack of knowledge of the system parameters in an adaptive controller. Since \( E\{\sum_{i=d+1}^{N}(y_i - y_i^*)^2\} = E(R_N) + E(\sum_{i=d+1}^{N}\eta_i^2) \), the problem of minimizing the total expected quadratic cost is equivalent to that of minimizing the expected value of the regret \( R_N \).

An input sequence is called "self-optimizing" (or "globally convergent") if

\[
R_n/n \to 0 \text{ a.s.} \tag{4.12}
\]

Since the seminal paper of Åström and Wittenmark [30] on "self-tuning regulators", an active area of research has been to find self-optimizing control schemes for (4.1) that can be implemented in real time. To review the Åström-Wittenmark [30] idea of self-tuning regulators for the stochastic linear system (4.1), first consider the case of unit delay \( (d = 1) \) and white noise \( (C(q^{-1}) = 1) \). Replacing \( b_1 \) in (4.1) by a prior guess \( b \neq 0 \), they proposed to estimate the other parameters at stage \( t \) by \( \alpha_1^{(t)}, \ldots, \alpha_p^{(t)}, \beta_2^{(t)}, \ldots, \beta_k^{(t)} \) that minimize

\[
\sum_{i=1}^{t}(y_i + \alpha_1 y_{i-1} + \cdots + \alpha_p y_{i-p} - bu_{i-1} - \beta_2 u_{i-2} - \cdots - \beta_k u_{i-k})^2, \tag{4.13}
\]

and to determine the input \( u_t \) by the certainty-equivalence rule

\[
u_t = \frac{\alpha_1^{(t)} y_t + \cdots + \alpha_p^{(t)} y_{t-p+1} - \beta_2^{(t)} u_{t-1} - \cdots - \beta_k^{(t)} u_{t-k+1}}{b} \tag{4.14}\]

They showed that if the estimates should converge as \( t \to \infty \), then \( \alpha_1^{(t)}/b, \ldots, \alpha_p^{(t)}/b, \beta_2^{(t)}/b, \ldots, \beta_k^{(t)}/b \) must necessarily converge to the coefficients \( a_1/b_1, \ldots, a_p/b_1, b_2/b_1, \ldots, b_k/b_1 \) in the optimal controller \( u^*_t = \{a_1 y_t + \cdots + a_p y_{t-p+1} - b_2 u_{t-1} - \cdots - b_k u_{t-k+1}\}/b_1 \) that assumes knowledge of the system parameters, and therefore the adaptive controller \( u_t \) "self-tunes" itself in the sense that its defining equation has asymptotically negligible difference from that of \( u^*_t \). By reparametrizing the system (4.1) and using least squares or extended least squares to directly estimate the transformed parameters, they also suggested natural extensions of (4.14) to general delay and colored noise.

An open problem with the Åström-Wittenmark approach is whether with positive probability the parameter estimates may fail to converge. Instead of adhering to a prior guess \( b \) of \( b_1 \) in their rule, an obvious modification is to update this guess with the current and past data. Extending the ideas described in §3 for adaptive control of the stochastic regression model (3.14), Lai and Wei [31] studied this modification of the original Åström-wittenmark scheme (4.13)-(4.14). Specifically, under the assumption that \( A(z) \) is stable,
they introduced occasional blocks of white-noise probing inputs and used only the data from these stages of forced learning to estimate the system parameters, thus obtaining a strongly consistent estimate \( b(t) \) of \( b_1 \). Replacing \( b \) by \( b(t) \) in (4.13)-(4.14) whenever \( b(t) \neq 0 \) and the number of probing inputs up to stage \( t \) does not fall below some threshold \( K_t \) (with \( K_t \to \infty \) and \( K_t = o(\log t) \)) leads to a certainty-equivalence controller, for the system (4.1) with \( d = 1 \) and \( C(q^{-1}) = 1 \), whose regret has the logarithmic order

\[
R_n \leq (1 + o(1))\sigma^2(p + k - 1) \log n \text{ a.s.,} \tag{4.15}
\]

cf. [31]. Note that the right hand side of (4.15) is the same as the lower bound (4.8) for the Bayes regret assuming knowledge of \( b_1 \). Lai [27] calls adaptive regulators having the property (4.15) "asymptotically efficient". As shown in [31], not only does \( b(t) \) converge a.s. to \( b_1 \), but the least squares estimates \( \alpha_1^{(t)}, \ldots, \alpha_p^{(t)}, \beta_2^{(t)}, \ldots, \beta_k^{(t)} \) determined by (4.13) with \( b \) replaced by \( b(t) \) also converge a.s to \( a_1, \ldots, a_p, b_2, \ldots, b_k \).

In the unit-delay, white noise case (i.e., \( d = 1 = C(q^{-1}) \), (4.1) can be written as a stochastic regression model

\[
y_{n+1} = \theta^T \psi_n + \epsilon_{n+1}, \tag{4.16}
\]

where \( \theta = (-a_1, \ldots, -a_p, b_1, \ldots, b_k)^T \) and \( \psi_n = (y_n, \ldots, y_{n-p+1}, u_n, \ldots, u_{n-k+1})^T \). An obvious certainty-equivalence rule in this case is to define \( u_t \) by the equation

\[
\theta^T_t \psi_t = y_t^*, \tag{4.17}
\]

where \( \theta_t \) is the least squares estimator which has the recursive representation

\[
\theta_t = \theta_{t-1} + P_{t-1} \psi_{t-1}(y_t - \theta^T_{t-1} \psi_{t-1}), \tag{4.18a}
\]

\[
P_t = P_{t-1} - P_{t-1} \psi_t \psi^T_t P_{t-1}/(1 + \psi^T_t P_{t-1} \psi_t). \tag{4.18b}
\]

A first difficulty with this straightforward approach is that (4.17) need not be well defined since the coefficient \( b_i^{(t)} \) of \( u_t \) in \( \theta^T_t \psi_t \) may be 0 unless some continuity assumptions are made on the distribution of \( \{\epsilon_t\} \). For the case \( y_t^* \equiv 0 \), Åström and Wittenmark [30] circumvented this difficulty by reparametrizing (4.16) as

\[
y_{n+1} = b_1(u_n - \lambda^T X_n) + \epsilon_{n+1}, \text{ where } X_n = (y_n, \ldots, y_{n-p+1}, u_{n-1}, \ldots, u_{n-k+1})^T,
\]

and replacing the unknown \( b_1 \) by \( b \neq 0 \) while using least squares criterion (4.13) to estimate \( \lambda \). A second difficulty with (4.17) or with (4.14) lies in the convergence analysis of the"
least squares estimator of $\theta$ or $\lambda$. A well known condition that ensures consistency of $\theta_t$ is the "persistent excitation" property:

$$n^{-1} \sum_{i=1}^{n} \psi_i \psi_i^T \text{ converges a.s. to a positive definite matrix.}$$  \hfill (4.19)

However, for the regulation problem $y_t^* \equiv 0$, the self-optimizing property (4.12) precludes persistent excitation for $\{\psi_i\}$, although it does not preclude persistent excitation for $\{X_i\}$, as has been noted by Kumar [1, pp. 372-373] and Lai and Wei [32, pp. 230 and 251].

For the unit-delay colored-noise case, (4.1) can still be written in the form of a stochastic regression model (4.16) with

$$\psi_n = (y_n, \ldots, y_{n-p+1}, u_n, \ldots, u_{n-k+1}, \varepsilon_n, \ldots, \varepsilon_{n-h+1})^T.$$  \hfill (4.20)

However, the regressor $\psi_n$ contains unobservable components $\varepsilon_n, \ldots, \varepsilon_{n-h+1}$. Replacing the unobservable $\varepsilon_i$ by their estimates $\hat{\varepsilon}_i$ in the recursion (4.18) leads to the "extended least squares" algorithm of the form

$$\theta_t = \theta_{t-1} + P_{t-1} \phi_{t-1} (y_t - \theta_{t-1}^T \phi_{t-1}), \quad P_t^{-1} = P_{t-1}^{-1} + \phi_t \phi_t^T,$$  \hfill (4.21)

where

$$\phi_t = (y_t, \ldots, y_{t-p+1}, u_t, \ldots, u_{t-k+1}, \hat{\varepsilon}_t, \ldots, \hat{\varepsilon}_{t-h+1})^T.$$  \hfill (4.22)

The estimates $\hat{\varepsilon}_i$ of $\varepsilon_i$ in (4.22) are given either by the residuals $\hat{\varepsilon}_i = y_i - \theta_t^T \phi_{i-1}$, in which case (4.22) is called the AML algorithm, or by the prediction errors $\hat{\varepsilon}_i = y_i - \theta_{i-1}^T \phi_{i-1}$, in which case (4.22) is called the RML1 algorithm, cf. [33]. Assuming that $\text{Re}(1/C(e^{it}) - 1/2) > 0$ for all $t \in [-\pi, \pi]$, Solo [33] proved the strong consistency of $\theta_t$ under the persistent excitation condition (4.19).

Because of the difficulties in satisfying the persistent excitation condition for consistency of least squares or extended least squares estimates in a closed loop system where the inputs are determined by a certainty-equivalence rule, it was natural that the first major breakthrough in the development of self-optimizing controllers made use of a different recursive identification scheme which does not involve matrix-type analysis and conditions like the persistent excitation property (4.19). Instead of the matrix gain $P_{t-1} = (P_{t-1}^{-1})^{-1}$ in (4.18) or (4.21), Goodwin, Ramadge and Caines [34] used the scalar gain $1/\text{tr}(P_{t-1}^{-1}) = 1/(\text{tr}P_0^{-1} + \sum_1^{t-1} \|\phi_t\|^2)$, which circumvents the difficulties of matrix
analysis. Moreover, instead of convergence analysis of the recursive identification algorithm, they proceeded directly to prove that the certainty-equivalence control rule associated with this scalar-gain algorithm is self-optimizing in the sense of (4.12). An important first step in their approach is to reparametrize (4.1) in the following prediction form. By the division algorithm, there exist polynomials \( F(z) = 1 + f_1 z + \cdots + f_{d-1} z^{d-1} \) and \( G(z) = g_1 + \cdots + g_{p(d)} z^{p(d)-1} \) with \( p(d) = p \vee (h - d + 1) \) such that

\[
C(z) = A(z) F(z) + z^d G(z),
\]

and therefore (4.1) can be rewritten in the form

\[
C(q^{-1}) \{ y_{n+d} - F(q^{-1}) e_{n+d} \} = G(q^{-1}) y_n + B(q^{-1}) F(q^{-1}) u_n,
\]

cf. [35]. Hence, in the case of known system parameters, the optimal \( d \)-step ahead predictor \( \tilde{y}_{n+d} \triangleq E(y_{n+d} | \mathcal{F}_n) \) is given by

\[
\tilde{y}_{n+d} + c_1 \tilde{y}_{n+d-1} + \cdots + c_h \tilde{y}_{n+d-h} = G(q^{-1}) y_n + (BF)(q^{-1}) u_n,
\]

where \( (BF)(z) = B(z) F(z) = b_1 + (bf)_2 z + \cdots + (bf)_{k+d-1} z^{k+d-2} \), and its prediction error is

\[
\eta_{n+d} \triangleq y_{n+d} - \tilde{y}_{n+d} = F(q^{-1}) e_{n+d}.
\]

Let

\[
\tilde{\theta} = (g_1, \cdots, g_{p(d)}, b_1, (bf)_2, \cdots, (bf)_{k+d-1}, -c_1, \cdots, -c_h)^T.
\]

Since \( y_{n+d} = \tilde{y}_{n+d} + \eta_{n+d} \), we obtain from (4.25) the prediction form of (4.1): \( y_{n+d} = \tilde{\theta}^T \tilde{\psi}_n + \eta_{n+d} \), where

\[
\tilde{\psi}_n = (y_n, \cdots, y_{n-p(d)+1}, u_n, \cdots, u_{n-k-d+2}, \tilde{y}_{n+d-1}, \cdots, \tilde{y}_{n+d-1})^T.
\]

First consider the case of unit delay \( d = 1 \). Here \( F(z) = 1, z G(z) = C(z) - A(z) \), \( p(d) = p \vee h \), and \( \eta_t = \epsilon_t \). In the case of known system parameters, the optimal controller chooses \( u_t \) so that \( \tilde{y}_{t+1} = y^*_{t+1} \). Without assuming knowledge of the parameters, Goodwin, Ramadge and Caines [34] therefore replaced the \( \tilde{y}_t \) in \( \tilde{\psi}_n \) by the target value \( y^*_t \) and the unknown parameter vector \( \tilde{\theta} \) in (4.27) by a stochastic gradient estimate of the form

\[
\theta^*_t = \theta^*_{t-1} + (a/r^*_t) \phi^*_{t-1} (y_t - \tilde{\theta}^T_{t-1} \phi^*_{t-1}), \quad r^*_t = r^*_t + \| \phi^*_t \|^2,
\]

25
where

$$\phi^*_n = (y_n, \cdots, y_{n+1-p}, u_n, \cdots, u_{n+1-k}, y^*_n, \cdots, y^*_{n+1-h})^T \text{ for } n \geq h. \quad (4.30)$$

They used the certainty-equivalence rule that defines the input $u_t$ for $t \geq h$ by the equation

$$\theta^*_t \phi^*_t = y^*_{t+1}. \quad (4.31)$$

To ensure that the component $b_{t,1}$ of $\theta^*_t$ is nonzero a.s. so that $u_t$ is well defined by (4.31), they assumed that $(x_0, \epsilon_1, \cdots, \epsilon_n)$ is absolutely continuous with respect to Lebesgue measure for every $n \geq 1$. Under certain additional stability and positive real assumptions on $B(z)$ and $C(z)$ and boundedness assumptions on $\{y^*_t\}$, they used a martingale argument to establish the self-optimizing property (4.12) for the certainty-equivalence rule (4.31). For general delay, they made use of an “interlacing” technique to extend this control scheme.

As mentioned earlier, a long-standing difficulty with certainty-equivalence rules based on standard recursive identification methods like least squares or extended least squares has been that the persistent excitation conditions commonly used in their convergence analysis may fail to hold for self-optimizing systems. For the least squares method in the white-noise, unit-delay case, Lai and Wei [36] succeeded in establishing its strong consistency under a much weaker excitation condition. Specifically, for a general stochastic regression model (4.16) in which $\{\epsilon_n, F_n, n \geq 1\}$ is a martingale difference sequence such that $\sup_n E(|\epsilon_n|^2 | F_{n-1}) < \infty$ a.s. for some $\alpha > 2$ and $\psi_n$ is $F_n$-measurable, they showed that the least squares estimate $\theta_n$ defined recursively by (4.18) converges a.s. to $\theta$ if

$$\lambda_{\min}(\sum_1^n \psi_i \psi_i^T) \rightarrow \infty \text{ and } \log \lambda_{\max}(\sum_1^n \psi_i \psi_i^T) = o(\lambda_{\min}(\sum_1^n \psi_i \psi_i^T)) \text{ a.s.} \quad (4.32)$$

They also gave an example of a stochastic regression model in which $\lambda_{\max}(\sum_1^n \psi_i \psi_i^T)$ is of the same order of magnitude as $\lambda_{\min}(\sum_1^n \psi_i \psi_i^T)$ and $\theta_n$ is inconsistent. For the unit-delay case $d = 1$ in (4.1), under the stability and positive real assumption $\min_{|\epsilon| \leq \frac{\pi}{2}} \Re\{1/C(e^{i\epsilon}) - \frac{1}{2}\} > 0$, Lai and Wei [37] showed that (4.32) is again sufficient to ensure the strong consistency of the AML algorithm (4.21)-(4.22). The relaxation of the excitation condition from the persistent type (4.19) to the weak type (4.32) suggests new possibilities in using the AML algorithm (which reduces to least squares in the white noise case $C(q^{-1}) = 1$) for consistent parameter estimation in an adaptive control environment. For example, for systems with bounded sample mean square inputs and outputs (i.e., $n^{-1} \sum_1^n (y_i^2$ +
\( u^2 = O(1) \text{ a.s.}, \) \( \log \lambda_{\max}(\sum_1^n \psi_i \psi_i^T) \) is of the order \( O(\log n) \), and therefore the excitation condition (4.32) is satisfied if \( \lambda_{\min}(\sum_1^n \psi_i \psi_i^T)/\log n \to \infty \text{ a.s.} \), which can typically be achieved by introducing white-noise probing inputs occasionally so that the total variance \( V_n \) of these probing inputs up to stage \( n \) is of the order \( (\log n)(\log \log n) \). However, if we want to keep the regret \( R_n \) to be no larger than \( O(\log n) \), then we cannot afford to have \( V_n/\log n \to \infty \) but must have at most \( V_n = O(\log n) \). Although this is in conflict with the excitation condition (4.32) for consistency of the AML algorithm \( \theta_n \), the convergence analysis of the least squares algorithm in [36] and of the AML algorithm [37] is fortunately versatile enough to provide two other useful ideas to handle this delicate situation.

The first idea is to obtain bounds for \( \theta_n - \theta \), and the convergence analysis of [37] in fact shows that

\[
||\theta_n - \theta|| = O((\log \lambda_{\max}(P_n^{-1}))^{1/2}/\lambda_{\min}^{1/2}(P_n^{-1})) \text{ a.s.} \tag{4.33}
\]

The second idea is to obtain bounds for the cumulative squared difference \( \sum_1^n (\theta^T \psi_i - \theta_i^T \phi_i)^2 \) between the optimal one-step ahead predictor \( \theta^T \psi_i \) of the response \( y_{i+1} \) (that assumes knowledge of the parameter vector \( \theta \) and initial condition \( x_0 \)) and the adaptive predictor \( \theta_i^T \phi_i \) (that does not assume knowledge of \( \theta \) and \( x_0 \)). This cumulative squared difference is of particular interest in adaptive control applications since the optimal controller assuming knowledge of the system parameters is given by \( \theta^T \psi_t = y_t^*, \) while the certainty-equivalence rule based on the adaptive predictor is given by \( \theta_i^T \phi_i = y_t^* \). It is shown in [37] that for every \( 0 < \delta < 1 \),

\[
\sum_{i=1}^n (\theta^T \phi_i - \theta_i^T \phi_i)^2 I_{\{\phi_i^T P_i \phi_i \leq \delta\}} = O(\log \lambda_{\max}(P_n^{-1})) \text{ a.s.} \tag{4.34}
\]

Making use of (4.33) and (4.34), Lai and Wei [37] proved not only the self-optimizing property (4.12) but also the much stronger conclusion

\[
R_n = O(\log n) \text{ a.s.} \tag{4.35}
\]

for a modification of the certainty-equivalence rule based on the AML algorithm, under the assumptions of boundedness of the target values \( y^*_i \), stability of the open-loop plant (i.e., \( A(z) \) and \( B(z) \) are stable), and positive real assumptions for \( C(z) \). A basic ingredient of this modification is a simple criterion to decide whether information is inadequate for approximating the unobservable \( \theta^T \psi_i \) by \( \theta_i^T \phi_i \). When the data show inadequate information for such approximation, instead of adhering to the certainty-equivalence formula
\( \theta^T \phi_t = y_{t+1}^* \) to determine the output \( u_t \), [37] proposes to introduce a block of white-noise perturbations to improve the information content of the design, in such a way that the number of these perturbations up to stage \( n \) is kept within the order \( O(\log n) \).

Despite the much stronger results (4.15) and (4.35) on least squares and extended least squares certainty-equivalence rules with forced learning in [31] and [36] than the self-optimizing property (4.12) for the stochastic gradient certainty-equivalence rule (4.31), the requirement that \( A(z) \) be stable in [31] and [36] imposes a serious limitation, particularly in light of the fact that (4.31) can be used to stabilize a system that would be unstable in the open loop. Another limitation in the methods and results of [31] and [36] is that they are applicable only to the unit delay case. Lai and Ying [38] recently removed these limitations by a parallel implementation of stochastic gradient and recursive maximum likelihood certainty-equivalence rules and by using occasional white-noise perturbations of the target values instead of the simple white-noise inputs in [31] and [36]. For the case of general delay \( d \) and without assuming \( A(z) \) to be stable, this approach gives the following logarithmic order for the regret:

\[
R_n \leq (1 + o(1))(1 + f_1^2 + \cdots + f_{d-1}^2)\sigma^2(p(d) + h + k + d - 1)(2d - 1)\log n \text{ a.s.} \quad (4.36)
\]

Moreover, if \( \log(1 + \sum_{i=1}^n y_i^*) = o(\log n) \), as in the regulation problem \( (y_i^* \equiv 0) \), we can further strengthen (4.36) into

\[
R_n \leq (1 + o(1))(1 + f_1^2 + \cdots + f_{d-1}^2)\sigma^2(p(d) + d + k - 2)(2d - 1)\log n \text{ a.s.} \quad (4.37)
\]

Note that in the case \( d = 1 \) and \( C(q^{-1}) = 1 \) (so that \( h = 0 \)), (4.37) reduces to (4.15).

Although the extended least squares recursive identification (4.21) has been called "approximate maximum likelihood" (AML) or "recursive maximum likelihood of type 1" (RML1), its statistical properties are actually unrelated to those of the off-line maximum likelihood estimator even when the \( \epsilon_i \) are normal. Linear approximations to the nonlinear estimating equations defining the off-line maximum likelihood estimator in some neighborhood of the true parameter lead to the RML2 algorithm, cf. [39]. For adaptive control in the case of general delay \( d \), Lai and Ying [38] use an extension of the RML2 algorithm to estimate the parameter \( \tilde{\theta} \) in the regression model (4.28) which is the prediction form of the model (4.1). This extension also includes monitoring by a consistent recursive estimator \( \tilde{\theta}_{m_j} \) constructed by the method of moments at stages \( m_1 < m_2 < \cdots \) that will be described later. Letting \( I_j \) be a cube in \( \mathbb{R}^{p(d)+k+d-1+h} \) with center at \( \tilde{\theta}_{m_j} \) so that

\[
P\{ \tilde{\theta} \in I_j \text{ for all large } j \} = 1, \lim_{j \to \infty} \text{ (width of } I_j) = 0, \quad (4.38)
\]

28
as will be specified later, the monitored recursive maximum likelihood algorithm $\theta_{n,M}$ is constrained to lie inside the “confidence interval” $I_j$ for $m_j \leq n < m_{j+1}$. The projection which we use to constrain $\theta_{n,M}$ is taken with respect to the norm induced by the positive definite matrix $P_{n-d}^{-1}$ defined in (4.40d) below, instead of the usual Euclidean norm. For $x \in R^{p(d)+k+d-1+h}$ and $m_j \leq n < m_{j+1}$, let $\pi_n(x)$ denote the unique solution of the quadratic programming problem

$$(\pi_n(x) - x)^TP_{n-d}^{-1}(\pi_n(x) - x) = \min_{y \in I_j}((y - x)^TP_{n-d}^{-1}(y - x)),$$  

(4.39)

i.e., $\pi_n(x)$ is the projection of $x$ into $I_j$ with respect to the norm induced by $P_{n-d}^{-1}$. The choice of a cube for the confidence region $I_j$ implies linear constraints for the quadratic programming problem (4.39), which can be handled by simple computational methods. Define $\theta_{n,M} = (\xi_{n,1}, \ldots, \xi_{n,p(d)}, \xi_{n,1},(\xi_{n,2}, \ldots, (\xi_{n,k+d-1,-\xi_{n,h}, \ldots, -\xi_{n,h}})^T$ for $n > m_1$ by the recursion

\begin{align}
\theta_{n,M} &= \pi_n(\theta_{n-1,M} + P_{n-d}\xi_{n-1}(y_n - \hat{y}_n,M)), \\
\xi_n + \xi_{n-1} + \cdots + \xi_{n,h} &= \phi_{n,M}, \text{ where} \\
\phi_{n,M} &= (y_{n-1}, \ldots, y_{n-p(d)+1}, u_{n}, \ldots, u_{n-k+d-2}, \hat{y}_{n-1,d-1}, \cdots, \hat{y}_{n,1,d-h,M})^T, \\
\hat{y}_{n+d,M} &= \theta_{n,M}^T\phi_{n,M}, \\
P_{n-1}^{-1} &= P_{n-1}^{-1} + \xi_n\xi_n^T + I/n,
\end{align}

(4.40a,b,c,d)

where $P_{m_1}$ is a positive definite matrix and $\theta_{m_1,M}$ represents an initial guess of $\tilde{\theta}$ (e.g., set $\theta_{m_1,M} = \tilde{\theta}_{m_1}$), and $I$ denotes the identity matrix.

Suppose that $C(z)$ is stable and that the $e_n$ satisfy assumption (4.2) and the boundedness condition $\sup_n |e_n| < \infty$ a.s. Defining $\xi_n$ and $P_n$ as in (4.40), assume that

$$\xi_n^TP_n\xi_n \rightarrow 0, \sum_{i=1}^n \|\xi_i\|^2 \leq n^{1+o(1)}, \sum_{j:m_j \leq n} \sum_{r=1}^{h+d-1} (\|\xi_m - r\|^2 + \|\xi_m - r\|) = O(\log n) \text{ a.s.}$$

(4.41)

Let $\sigma^2 = \sigma^2(1 + f_1^2 + \cdots + f_{d-1}^2)$. Then as shown in [38],

$$\sum_{i=1}^n \{E(y_{t+d}|F_t) - \hat{y}_{t+d,M})^2 \leq (1 + o(1))\sigma^2(p(d) + h + k + d - 1)(2d - 1)\log n \text{ a.s.}$$

(4.42)

Note that the right hand side of (4.42) is the same as the logarithmic order of magnitude (4.36) of the regret for adaptive controllers that we want to achieve. Therefore the certainty-equivalence rule based on the monitored recursive maximum likelihood algorithm

$$\theta_{n,M}^T\phi_{n,M} = y_{n+d}^*$$

(4.43)
has the desired order of magnitude (4.36) for its regret if it can be shown that (4.41) holds and if consistent auxiliary estimates can be developed to define the monitoring sets $I_j$. This is the basic idea behind the asymptotically efficient adaptive control scheme in [38].

With probability 1, since $\theta_{n,M} \to \tilde{\theta}$ and $b_1 \neq 0$, (4.43) is well defined for all large $n$. To define strongly consistent estimators $\hat{\theta}_{m_j}$, Lai and Ying [38] introduce blocks of well-excited input-output data from which the $\theta_{m_j}$ are to be constructed. Since $A(z)$ may be unstable, they modify an idea due to Caines and Lafortune [40] to construct such blocks of well excited data. Without assuming $A(z)$ to be stable but assuming that $B(z)$ is stable and $\min_{\|t\| \leq n} \text{Re}\{C(e^{it}) - a/2\} > 0$, Caines and Lafortune [40] made use of the stochastic gradient certainty-equivalence rule (4.31) to stabilize the system but introduced white-noise perturbations $w_t$ into the target values $y_{t+1}^*$ of (4.31) to excite the system. Thus, they replaced the Goodwin-Ramadge-Caines scheme (4.31) for the unit-delay case by

$$\theta_T \phi_t^* = y_{t+1}^* + w_t, \quad \text{where}$$

$$\phi_t^* = (y_t, \cdots, y_{t+1-p}v, u_t, \cdots, u_{t+1-k}, y_t^* + w_{t-1}, \cdots, y_{t+1-h}^* + w_{t-h})^T,$$

and $w_t$ are independent random variables such that $Ew_t^2 = v > 0$ and $\sup_t Ew_t^2 < \infty$. Lai and Ying [38] propose to apply the stochastic gradient scheme only to broken blocks of successive observations for which white-noise dither signals are used as target values in the corresponding certainty-equivalence equation defining the inputs. Specifically, assuming that $\min_{\|t\| \leq n} \text{Re}\{C(e^{it}) - (d - \frac{1}{2})a\} > 0$ for some $a > 0$ and that $\sup_n |e_n| < \infty$ a.s., they define the inputs $u_n$ at stages $n \in J = \bigcup_{i=1}^{\infty} \{n_i, n_i + 1, \cdots, m_i - 1\}$ (representing the broken blocks of consecutive observations) by the equation

$$\theta_{n,G} \phi_{n,G} = w_n \quad \text{if} \quad b_{n,G} \neq 0, \quad \text{and} \quad u_n = w_n \quad \text{otherwise},$$

(4.44)

where $\theta_{n,G}$ is a modified stochastic gradient algorithm estimating $\tilde{\theta}$ defined recursively as follows: Choose $\theta_{0,G} = \cdots = \theta_{n_1+d-1,G}$ such that its component $b_{0,G}$ estimating $b_1$ is nonzero. Let $\phi_{n,G} = (y_n, \cdots, y_{n-p(d)+1}, u_n, \cdots, u_{n-k-d+2}, w_{n-1}, \cdots, w_{n-k})^T$. For $m_{j-1} + d - 1 < n < n_j + d$, define $\theta_{n,G} = \theta_{m_{j-1}+d-1,G}$. For $n_j + d - n \leq m_j + d - 1$ define

$$\theta_{n,G} = \theta_{n-1,G} + (1/r_{n-d})(y_n - w_n)\phi_{n-d,G},$$

(4.45a)

$$r_{n-d} = r_{n-d-1} + \|\phi_{n-d,G}\|^2, \quad r_{n_j-1} = \sum_{i=1}^{j-1} \sum_{t=n_i+d}^{t=n_i+d-1} \|\phi_{n-d,G}\|^2 +$$

$$\sum_{i=1}^{j} \left( \sum_{\nu=0}^{p(d)-1} y_{n_i-\nu}^2 + \sum_{\nu=1}^{k+d-2} u_{n_i-\nu}^2 \right) \log^2 \left( \sum_{i=1}^{j} \left( \sum_{\nu=0}^{p(d)-1} y_{n_i-\nu}^2 + \sum_{\nu=1}^{k+d-2} u_{n_i-\nu}^2 \right) \right).$$

(4.45b)
Assuming furthermore that $B(z)$ is stable and that $z^p A(z^{-1}), z^{k-1} B(z^{-1})$ and $z^h C(z^{-1})$ are coprime polynomials, [38] shows that the input-output data at stages $t \in J$ have the following excitation property:

$$\left( \sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} Z_i U_t^T \right) / \#_j$$ converges a.s. to a nonrandom matrix of full rank $p + k$, where $U_t = (y_{t-1}, \ldots, y_{t-p}, u_{t-d}, \ldots, u_{t-d-k+1})^T$,

$$Z_t = (y_{t-h-1}, \ldots, y_{t-h-k+1}, w_{t-d}, \ldots, w_{t-d-p-k+1})^T, \#_j = \sum_{i=1}^{j} (m_i - n_i). \quad (4.46)$$

In view of (4.46), [38] defines consistent estimates $\lambda_{m_j}$ of $\lambda = (-a_1, \ldots, -a_p, b_1, \ldots, b_k)^T$ by

$$\lambda_{m_j} = (V_j^T V_j)^{-1} V_j^T \sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} Z_t y_t, \text{ where } V_j = \sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} Z_t U_t^T.$$  

In fact, [38] uses (4.46) to show that $\lambda_{m_j} - \lambda = o(\#_j^{-1/2+\delta})$ a.s. for every $\delta > 0$. Using adaptive spectral factorization, [38] then estimates $(c_1, \ldots, c_h)$ with an error of $o(\#_j^{-1/2+\delta})$ a.s. This in turn provides strongly consistent estimates $\tilde{\lambda}_{m_j}$ of $\lambda$ such that $\tilde{\lambda}_{m_j} = \lambda + o(\#_j^{-1/2+\delta})$ a.s. for every $\delta > 0$. Pick any $0 < \delta < 1/2$ and let $I_j$ denote the cube with center $\tilde{\lambda}_{m_j}$ and width $\#_j^{-1/2+\delta}$ Then (4.38) clearly holds.

Suppose that the target value $y_n^*$ at stage $n$ is $\mathcal{F}_{n-d}$-measurable and that $\sum_{i=1}^{n} y_i^2 = O(n)$ and $y_n^* = o((\log n)^\gamma)$ a.s. for some $0 < \gamma < 1$. To define the sequence $n_1 < n_2 < \ldots$ associated with the modified stochastic gradient algorithm $\theta_{n,G}$, take a nondecreasing sequence of constants $K_n \geq d + h$ such that $K_n \to \infty$, $K_n = O((\log n)^{1/2 + (1-\gamma)})$ and $K_{2n} = O(K_n)$. Define inductively the stopping times

$$n_j = \inf\{n > m_{j-1} : \sum_{i=1}^{j-1} (m_i - n_i) \leq K_n \} \quad \text{and} \quad \sum_{r=0}^{p(d)-1} y_{n-r}^2 + \sum_{r=1}^{k+d-2} u_{n-r}^2 \leq K_n / \log^2(K_n + 2),$$

$$m_j = n_j + \left[ K_n \sum_{r=0}^{p(d)-1} y_{n-r}^2 + \sum_{r=1}^{k+d-2} u_{n-r}^2 \log \left( \sum_{i=1}^{j} (2 + \sum_{r=0}^{p(d)-1} y_{n_i-r}^2 + \sum_{r=1}^{k+d-2} u_{n_i-r}^2) \right) \right].$$

Let $J_n = \{ t \in J : t \leq n \}$ and let $\#(J_n)$ denote the number of elements of $J_n$. The adaptive control scheme in [38] defines the input $u_n$ by (4.44) using the modified stochastic
gradient scheme $\theta_{n,G}$ and white-noise perturbation $w_n$ at stage $n \in J$. At stage $n \notin J$, it uses the certainty-equivalence equation (4.43) based on the monitored recursive maximum likelihood algorithm $\theta_{n,M}$ to determine the input $u_n$ if $\#(J_n) > K_n^{1/2}$. If $n \notin J$ and $\#(J_n) \leq K_n^{1/2}$ (signalling too few white-noise excitations), [38] uses a stochastic gradient certainty-equivalence rule to determine the input $u_n$. The regret of the resultant rule is shown in [38] to have the logarithmic order (4.36), and in the case $\log(1 + \sum_{i=1}^{n} y_i^2) = o(\log n)$, to have even the smaller order (4.37). This basically follows from (4.42) since as shown in Lemmas 8 and 9 of [38], $\#(J_n) = O(K_n)$ and $\liminf_{n \to \infty} \#(J_n)/K_n > 0$ a.s.

REFERENCES


