ADAPTATION AND TRACKING IN SYSTEM IDENTIFICATION

Lennart Ljung
Department of Electrical Engineering
Linköping University
S-581 83 Linköping, Sweden

Abstract

This article gives a survey of basic techniques to derive and analyse algorithms for tracking time-varying systems. Special attention is paid to how different assumptions about the true system affect the algorithms. Explicit and semi-explicit expressions for the means square errors are derived, which clearly demonstrate the character of the trade-off between tracking ability and noise sensitivity.

1 PROBLEM FORMULATIONS

The efficient extraction dynamical properties is the topic of System Identification. Clearly, such issues are of major importance in many applications. A specific, but common, case is where the system's properties vary with time. It will then be the task of the identification algorithm to adapt itself so that it can appropriately track the system dynamics. This leads to the world of Recursive Identification. Many issues around such adaptive algorithms have been discussed in the control and signal processing literature. We may list the following items

—Given a model structure, develop a structure for the recursive identification algorithm. (1)

—Given a recursive identification algorithm, analyse its behaviour when the true system does not change with time. (2)

—Given a model structure and a description of how the true system changes, derive the optimal tracking algorithm (i.e. optimal choices of adaptation facilities) (as well as approximative variants). (3)

—Given an algorithm, with particular choices of adaptation facilities, as well as a description of how the true system changes, calculate expressions for the tracking error. (4)

Ljung and Söderström (1983) gives a comprehensive treatment of problem (1), as well of (2) in the case the adaptation gain tends to zero. Problem (2), in the case of non-decreasing gain, is treated, e.g. in Kushner and Huang (1981), Macchi and Eweda (1983). Problem (4) for some special cases of algorithms and system changes is dealt with in, e.g. Benveniste and Ruget (1982), Benveniste (1987), Eweda and Macchi (1985), Widrow et al (1976), Gardner (1984). In the present paper we shall focus on the issues (3) and (4).

1.1 Model Structures

There exist many possible parameterizations of the dynamical properties of systems. We shall not go into these details here, but following Ljung (1987) we consider a direct parameterization of the predictors as

$$\hat{y}(t \mid \theta) \tag{5}$$

Here $\hat{y}(t \mid \theta)$ is an arbitrary function of the parameter vector θ and of past data, $y(t-1), \ldots y(0), u(t-1), \ldots u(0)$. ($y(t)$ and $u(k)$ are the output and input, respectively at time k). One should think of $\hat{y}(t \mid \theta)$ as the prediction, or "guess" of the output at time t, based on the model θ, and based on observations of previous inputs and outputs.

A particularly simple structure is obtained when $\hat{y}(t \mid \theta)$ is a linear (or affine) function of θ:

$$\hat{y}(t \mid \theta) = \theta^T \varphi(t) \tag{6}$$

This is known as a *linear regression.*

Based on the general model structure (5) we can also give a general answer to problem (1), in the spirit of Ljung and Söderström (1983): The archetypical structure for parameter adjustments will be

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \gamma(t) R^{-1}(t) \psi(t) \varepsilon(t) \tag{7}$$

where

$$\varepsilon(t) = y(t) - \hat{y}(t), \tag{8}$$

$\hat{y}(t)$ and $\psi(t)$ are recursively computed approximations of

$$\hat{y}(t) \approx \hat{y}(t \mid \hat{\theta}(t-1)) \tag{9}$$

$$\psi(t) \approx \psi(t \mid \hat{\theta}(t-1)) \triangleq \frac{d}{d\theta} \hat{y}(t \mid \theta) \mid_{\theta = \hat{\theta}(t-1)} \tag{10}$$

(For (6), we have without approximation

$$\hat{y}(t) = \hat{\theta}^T(t-1)\varphi(t), \qquad \psi(t) = \varphi(t) \tag{11}$$

The matrix $R(t)$ and the possible scalar $\gamma(t)$ ("the gain") are the "adaptation facilities", alluded to earlier. The effects of these on the properties of $\hat{\theta}$, will be of main concern in this article.

1.2 Description of System Changes

The predictor structure (5) and the algorithm (7) are "the real things" for the user, together with the observed data stream. However, in order to make rational choices of $\gamma(t)$ and $R(t)$, and in particular in order to perform some quantitative analysis, it is necessary to introduce some assumptions about the true system.

Let $\theta_0(t)$ denote an "ideal value" of the parameter θ at time t. We may think of

$$\hat{y}(t \mid \theta_0(t))$$

as the "true predictor" at time t, but it may also be the best approximation of the true predictor, that is available within the model structure.

Remark. If "some true parameters" $\theta_0(t)$ are time varying, it may be that the ideal predictor expressed as in (5), actually would depend on a sequence of recent θ_0-values. We disregard such technicalities for the moment.

□

We may now introduce some idea of how $\theta_0(t)$ varies with time. Conceptually we then have

$$\theta_0(t) = f(\theta_0(t-1), t, \omega(t), \xi(t)) \tag{12}$$

Here $\{\omega(t)\}$ would be a sequence of random vectors and $\{\xi(t)\}$ would be a sequence of external indicators (perhaps partly observable) that affect the system.

Typical specific assumptions include:

Random walk:

$$\theta_0(t) = \theta_0(t-1) + \omega(t) \tag{13}$$

$$\omega(t) \in N(0, R_1(t))$$

$\omega(t), \omega(s)$ independent (meaning that $\omega(t)$ is white, Gaussian noise with possibly time varying covariance matrix $R_1(t)$).

Jump changes:

$$\theta_0(t) = \theta_0(t-1) + \omega(t)$$

$$\omega(t) = \begin{cases} 0 & \text{with probability } 1-\mu \\ v & \text{with probability } \mu \end{cases} \tag{14}$$

where v is a random variable with some distribution.

Markov chain:

$\theta_0(t)$ varies as a Markov chain among a finite number of vectors

$S(t)$: Transition probability matrix. (15)

Knowledge-based descriptions

An interesting possibility, especially for dynamics in large systems, is to express possible changes and relationships between sequences of changes in (12) as rules or predicates in a knowledge-base. The interaction between (12) and the algorithm (7) will then be of mixed numeric/symbolic character.

1.3 Coupling Between Algorithm and Assumed System Behaviour

This article is basically concerned with the interaction between assumed model behaviour (12) and the generic algorithm (7), in particular regarding the choices of adaptation mechanisms $R(t)$ and $\gamma(t)$. In Section 2 we ask the following questions

- Suppose a description like (12) is entirely known, what is then the optimal algorithm?

- Suppose a description like (12) is known, up to a number of parameters, how could then a suitable algorithm be developed?

In Section 4 we ask the questions

- Consider a given algorithm of the type (7) with given choices of $\gamma(t)$ and $R(t)$, how does then the error $\hat{\theta}(t)-\theta_0(t)$ behave under various assumptions about (12)?

In the development we mostly focus on the simple case of (6) and the descriptions (13) and (14). Markov chain descriptions (15) of the parameters are dealt with in Millnert (1987), while the intriguing case (15) will have to mature until another occasion.

2 ALGORITHMS

2.1 Optimal Algorithms: Linear regression and random walk parameters

We study now the combination of a linear regression (6) and the random walk model (13):

$$\theta_0(t) = \theta_0(t-1) + w(t) \tag{16}$$

$$y(t) = \theta_0^T(t)\varphi(t) + e(t) \tag{17}$$

We here assumed $\{e(t)\}$ to be white Gaussian noise with variance $R_2(t)$, while $\{w(t)\}$ is white Gaussian noise with covariance matrix $R_1(t)$. It is then well known, see, e.g. Section 2.3 in Ljung and Söderström (1983) that the estimate $\hat{\theta}(t)$ that minimizes

$$P(t) = E(\hat{\theta}(t) - \theta_0(t))(\hat{\theta}(t) - \theta_0(t))^T \tag{18}$$

(even in a matrix sense) is given by the Kalman filter:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + L(t)\varepsilon(t) \tag{19}$$

$$\varepsilon(t) = y(t) - \varphi^T(t)\hat{\theta}(t-1) \tag{20}$$

$$L(t) = \frac{P(t-1)\varphi(t)}{R_2(t) + \varphi^T(t)P(t-1)\varphi(t)} \tag{21}$$

$$P(t) = P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{R_2(t) + \varphi^T(t)P(t-1)\varphi(t)} + R_1(t) \tag{22}$$

We may note that $\hat{\theta}(t)$ is the conditional expectation of $\theta_0(t)$, given the observations $\{\varphi(k), y(k)\}, k \leq t$.

Note also that if the variability of the true parameter $\theta_0(t)$, i.e. the matrix $R_1(t)$ is known, (19)-(22) is the optimal algorithm also for, say, abrupt changes. (Take $R_1(t) = 0$ except when a jump occurs, take then $R_1(t) = R_1$). However, this requires the time instants for the jumps to be known, not-too-realistic an assumption.

2.2 Optimal Algorithms: Non-linear regressions and random walk with known characteristics

Consider now the more general case of ha model structure (5), together with the random walk model (13):

$$\begin{aligned} \theta_0(t) &= \theta_0(t-1) + w(t) \\ y(t) &= \hat{y}(t \mid \theta_0(t)) + e(t) \end{aligned} \tag{23}$$

Suppose that we have an approximation $\theta_*(t)$ of $\theta_0(t)$ available. We can then write, using the mean value theorem

$$\hat{y}(t \mid \theta_0(t)) = \hat{y}(t \mid \theta_*(t)) + (\theta_0(t) - \theta_*(t))^T\psi(t, \zeta(t)) \tag{24}$$

where $\zeta(t)$ is a value "between" $\theta_*(t)$ and $\theta_0(t)$. Here $\psi(t,\theta)$ is the gradient of $\hat{y}(t \mid \theta)$, as defined in (10). Normally, $\psi(t,\zeta(t))$ would not be known, but we may assume that an approximation

$$\psi(t) \approx \psi(t, \zeta(t)) \tag{25}$$

is available. Introduce the known variable

$$z(t) = y(t) - \hat{y}(t \mid \theta_*(t)) + \theta_*^T(t)\psi(t)$$

Subject to the approximation (25) we can then rewrite (23) as

$$\begin{aligned} \theta_0(t) &= \theta_0(t-1) + w(t) \\ z(t) &= \theta_0^T(t)\psi(t) + e(t) \end{aligned} \tag{26}$$

and we are back to the situation of Section 2.1. A natural choice $\theta_*(t)$ of a good approximation of $\theta_0(t)$ would be the previous estimate $\theta_*(t) = \hat{\theta}(t-1)$. We then obtain algorithms of the recursive prediction error type since

$$z(t) - \hat{\theta}^T(t-1)\psi(t) = y(t) - \hat{y}(t \mid \hat{\theta}(t-1))$$

As $\hat{\theta}(t-1)$ comes closer to $\theta_*(t)$, the approximation involved in going from (23) to (26) will become arbitrarily good. This shows that an asymptotic theory of tracking parameters in arbitrary model structures can be devel-

oped from the linear regression case. In the remainder of this article, though, we shall only deal with the latter situation.

2.3 Some Ad-Hoc Variants for the Linear Regression Case

Forgetting factors

A popular approach to deal with time-varying systems is to minimize a weighted criterion

$$V_t(\theta) = \sum_{k=1}^{t} \beta(t,k)(y(t) - \theta^T\varphi(t)) \qquad (27)$$

where more weight is put to recent measurements by

$$\beta(t,k) = \prod_{j=k+1}^{t} \lambda(j); \qquad \lambda(j) \leq 1 \qquad (28)$$

It is well known, e.g. Ljung and Söderström, Section 2.6.2 that this is accomplished by the algorithm (19)-(21) where $L(t)$ is obtained by

$$L(t) = \frac{P(t-1)\varphi(t)}{\lambda(t) + \varphi^T(t)P(t-1)\varphi(t)} \qquad (29)$$

$$P(t) = \frac{1}{\lambda(t)}[P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{\lambda(t) + \varphi^T(t)P(t-1)\varphi(t)}] \qquad (30)$$

We may note that

$$P(t) = [\sum_{k=1}^{t} \beta(t,k)\varphi(k)\varphi^T(k)]^{-1} \qquad (31)$$

Note that this is a special case of (19)-(22), corresponding to the choices

$$\hat{R}_1(t) = (\frac{1}{\lambda(t)} - 1)[P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{\lambda(t) + \varphi^T(t)P(t-1)\varphi(t)}] \qquad (32)$$

$$R_2(t) = \lambda(t)$$

LMS

Widrow's least mean squares algorithm, see e.g. Widrow and Stearns (1986), is a commonly used tool for adaptation. It is given by (19)-(20) with

$$L(t) = \mu\varphi(t) \qquad (33)$$

or, in a normalized variant

$$L(t) = \frac{\mu\varphi(t)}{1 + \mu \mid \varphi(t) \mid^2} \qquad (34)$$

Again, we may verify that (19) + (34) is a special case of the basic algorithm (19)-(22) corresponding to

$$\hat{R}_1(t) = \mu^2 \frac{\varphi(t)\varphi^T(t)}{1 + \mu \mid \varphi(t) \mid^2}$$

$$\hat{R}_2(t) = 1 \qquad (35)$$

$$P(0) = \mu \cdot I$$

2.4 Linear Regression and Random Walk Parameters with Unknown Characteristics

The basic formulation (16)-(17) with the optimal algorithm (19)-(20) is quite powerful. It can deal with both slowly drifting parameters and with sudden changes, by assigning proper values to the covariance matrix $R_1(t)$ and the variance $R_2(t)$. The main shortcoming is then that these values will rarely be known by the user. The remedy is of course to estimate these quantities in one way or another. We shall in this section review a few such methods.

In the first place we can count the methods of Section 2.3 as approaches of this character. See (32) and (35). These ad hoc approaches are thus optimal for some (quite particular) characteristics of the true parameter changes.

Drift

For slowly drifting parameters Isaksson(1987) has devised efficient methods for estimating $R_1(t)$ and $R_2(t)$. These estimates are then used in (19)-(22). The basic idea is to apply ideas from adaptive Kalman filtering. An alternative method is discussed by Wang and Deng (1986).

Another approach would be to use forgetting factor algorithm (29) and adjust the size of the forgetting factor $\lambda(t)$. From (32) we see that this corresponds to estimating the "size" of $R_1(t)$, but neglecting the direction information. Fortesque et al (1981) have devised such a method.

Jump changes

Consider the formulation (14) for sudden changes in the parameters. If v is described as a Gaussian random variable with zero mean and covariance R_1, we can describe $w(t)$ as a sequence of Gaussian random variables with covariances $R_1(t)$, where $R_1(t)$ is either 0 or R_1, but we do not know when. We do know, however that, for N data points, the true sequence $R_1(t)$ is one of 2^N possible combinations of 0 and R_1. In principle, we could run all the 2^N possible versions of (19)-(22), and we would know that the optimal $\hat{\theta}(t)'s$ would be one of the obtained 2^N variants.

How would we know which one? It is reasonable to assume that it would be the one that produced the smallest sequence of prediction errors $\varepsilon(t) = y(t) - \varphi^T(t)\hat{\theta}(t-1), t = 1,...N$. That would at least be the maximum likelihood estimate among this finite collection of possibilities.

Now, it is of course not feasible to run through all these possibilities in practice. It corresponds to an exponentially increasing tree of choices. Andersson (1985) has formulated a strategy for pruning this tree, so that only an a priori fixed number of possibilities is run in parallel, that works very well for the application to recursive identification. Both jump changes and drift can be handled in that way. A related approach for the Markov chain model (15) is described by Millnert (1987). It is particularly useful when the system dynamic changes refer to change of operating points and the like, so that previous system properties may become relevant again at a later time instant. The corresponding model is then stored away for reuse and slight adjustment, which allows a much faster adaptation.

Another natural way to deal with abruptly changing systems is to devise tools to explicitly detect the change (i.e. to test the hypothesis whether $R_1(t) = 0$ or $R_1(t) = R_1$, in our formulation). This problem has been discussed extensively by Benveniste and Basseville, see e.g. Basseville and Benveniste (1986).

Hägglund (1984) has used particularly devised change detection algorithms to be applied for adjusting gains in recursive identification algorithms.

3 BASIC EXPRESSIONS FOR THE ANALYSIS

3.1 An Exact Expression for the Parameter Error

Let us consider the description (13) for the behaviour of the true system together with the generic parameter estimation algorithms (19)-(20):

$$\theta_0(t+1) = \theta_0(t) + w(t) \tag{36}$$

$$\hat{\theta}(t) = \hat{\theta}(t-1) + L(t)\varepsilon(t) \tag{37}$$

$$\varepsilon(t) = y(t) - \varphi^T(t)\hat{\theta}(t-1) \tag{38}$$

$$y(t) = \varphi^T(t)\theta_0(t) + e(t) \tag{39}$$

Introduce the parameter error

$$\tilde{\theta}(t) = \hat{\theta}(t) - \theta_0(t+1) \tag{40}$$

Remark. The indexing here may seem somewhat peculiar, but it will simplify the expressions to follow. From an expression for the covariance of $\tilde{\theta}(t)$ we can easily derive, e.g. the covariance of $\hat{\theta}(t) - \theta_0(t)$.

□

Then

$$\tilde{\theta}(t) = (I - L(t)\varphi^T(t))\tilde{\theta}(t-1) + L(t)e(t) - w(t) \tag{41}$$

The parameter error thus obeys a linear, time varying difference equation. Notice that the $L(t)$ is always of the form

$$L(t) = \bar{P}(t)\varphi(t) \tag{42}$$

for some matrix $\bar{P}(t)$. Solving (41) gives

$$\tilde{\theta}(t) = \Phi(t,0)\tilde{\theta}(0) + \sum_{k=1}^{t} \Phi(t,k)[\bar{P}(k)\varphi(k)e(k) - w(k)] \tag{43}$$

where

$$\Phi(t,k) = \prod_{j=k}^{t} (I - \bar{P}(j)\varphi(j)\varphi^T(j)) \tag{44}$$

Essentially, (42) and (43) form the basis for all analysis of the performance of the estimator. The difficulty lies in the complicated expression for $\Phi(t,k)$: Its properties depend entirely on the sequence $\{\varphi(t)\}$, but they are inherited in a fairly complicated way.

Notice that so far has no stochastic environment been invoked. The expressions (42), (43) hold for any sequences $\{\varphi(t)\}$, $\{e(t)\}$ and $\{w(t)\}$.

It is also useful at this point to dwell somewhat at the forgetting factor algorithm (29), for which

$$P^{-1}(t) = R(t) = \sum_{k=1}^{t} \beta(t,k)\varphi(k)\varphi^T(k) + \beta(t,0)R(0)$$

Inserting this in (41) gives, after some manipulation

$$\tilde{\theta}(t) = \beta(t,0)R^{-1}(t)R(0)\tilde{\theta}(0) +$$

$$\sum_{k=1}^{T} \beta(t,k)R^{-1}(t)[\varphi(k)e(k) - R(k)w(k)] \tag{45}$$

($\beta(t,k)$ is defined by (28)). This shows that

$$\Phi(t,k) = \beta(t,k)R^{-1}(t)R(k) \tag{46}$$

in this particular case.

3.2 Bounds on $\Phi(t,k)$: Exponential Stability

From (43) we see that an interesting and desirable property of $\Phi(t,k)$ is that it should be exponentially decaying in the sense

$$\| \Phi(t,k) \| \leq C \cdot \lambda^{t-k} \qquad \text{for some} \qquad \lambda \leq 1 \qquad (47)$$

Then the system (41) is exponentially stable and useful conclusions about error propagation, bounds on the errors and the like can be derived.

Forgetting factor algorithm

For the forgetting factor algorithm, we have from (28) that

$$\beta(t,k) \leq \lambda^{t-k} \qquad \text{where} \qquad \lambda = \max_{kj \leq t} \lambda(j) \qquad (48)$$

In order to establish (47) it is in this case thus sufficient to that $R^{-1}(t)$ is uniformly bounded and that $R(k)$ increases slower than exponentially. The key property is consequently

$$\sum_{k=1}^{t} \beta(t,k)\varphi(k)\varphi^T(k) \geq \varepsilon \cdot I; \qquad \text{some} \qquad \varepsilon \geq 0 \quad (49)$$

which is usually phrased as a *persistence of excitation* condition on the sequence $\{\varphi(k)\}$, or on its components, cf Ljung (1987), Ch. 14.

LMS-algorithm

In an important series of papers Bitmead and Anderson (1980) and Bitmead (1984) have studied conditions for (49) (or a variant thereof) to imply (37) for the LMS algorithm.

Kalman Filter algorithm

We may also note that the observability Gramian of the system (16)-(17) is

$$O(t,k) = \sum_{j=k}^{t} \varphi(j)\varphi^T(j) \qquad (50)$$

so if

$$O(t,k) \geq C \cdot I \qquad \forall\, t,k \;\; | t-k | > N \qquad (51)$$

the the system is uniformly completely observable. From general Kalman filter theory, e.g. Jazwinsky (1970), we then know that the algorithm (19) is exponentially stable, i.e. (47) holds, provided

$$\sum_{j=k}^{t} R_1(j) \geq C \cdot I \qquad \forall\, t,k, \; | t-k | > N \qquad (52)$$

(i.e. the system is uniformly completely controllable from the process noise). Provided (50)-(51) holds we can thus establish exponential stability also for LMS and the forgetting factor algorithm by checking (52) for (35) and (32), respectively.

In summary, then, exponential stability of the error equation (41) is something that essentially follows from conditions like (50), (51) for all variants of the algorithms under discussion here.

Ljung (1986) discussed modifications of (19) that assure exponentially stability of the modified algorithm also when (51) does not hold.

3.3 Asymptotic Expressions for $\Phi(t,k)$

Explicit expressions for Φ are complicated. Also, if the components of $\varphi(t)$ are modelled as stochastic processes, the distribution of Φ will indeed be complex. The interest has therefore focused on asymptotic expressions for $\Phi(t,k) =$ corresponding to the case $\bar{P}(j) \approx \bar{P}$ is a small, almost constant matrix, and the time span $t-k$ is large. The basic result then is as follows: Assume

$$\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t) \to Q \text{ as } N \to \infty. \qquad (53)$$

Then

$$\Phi(t,k) = (I - \bar{P}Q)^{t-k} + g(\bar{P}, t-k) \qquad (54)$$

where

$$g(\bar{P},N) \to 0 \text{ as } \bar{P} \to 0 \text{ and } N \to \infty,$$
$$\text{such that} \qquad N \cdot \bar{P} \to \tau \qquad 0 < \tau < \infty \qquad (55)$$

This also means that

$$\Phi(t,k) = e^{-\bar{P}Q(t-k)} + \tilde{g}(\bar{P}, t-k) \qquad (56)$$

where $\tilde{g}$ has the same properties as g. The idea is of course that the term $\varphi(t)\varphi^T(t)$ is replaced by its average value Q in (44). This should be a reasonable thing to do when Φ comprises the result of many small such factors. There are a couple of different ways to prove (54). We note that the definition (44) implies that (take $k=0$)

$$\Phi(t,0) = \Phi(t-1,0) - \bar{P}(t)\{\varphi(t)\varphi^T(t)\Phi(t-1,0)\} \qquad (57)$$

According to averaging theory, the solution to (57) behaves for small, and/or decreasing values of $\bar{P}(t)$ as the solution to the ordinary differential equation

$$\frac{d}{d\tau}\Phi(\tau,0) = -\bar{P}Q\Phi(\tau,0)$$

which, of course, is (56) without the remainder term. See, e.g. Khazminski (1966), Ljung (1977), Kushner and Clark (1978), Benveniste, et al (1987).

Also, the results that we may develop from (54) are the same as those obtained from weak convergence theory, see, e.g. Kushner and Huang (1981), Benveniste (1987).

We shall in this article, use (54) and (55) in the analysis of $\tilde{\theta}(t)$, and refer to the above mentioned articles for details for further proof.

4 ASYMPTOTIC EXPRESSIONS FOR THE MEAN SQUARE ERROR

Equipped with the result (54) we can now return to an evaluation of the parameter error $\tilde{\theta}(t)$. We shall then deal with the general case (19)-(22), and later dwell on its special cases LMS and forgetting factor algorithm.

We shall now assume that $\hat{R}_1(t)$ in the algorithm is chosen as a constant

$$\hat{R}_1(t) = \gamma^2 R_1 \tag{58}$$

where γ serves as a scaling factor. We then have from (20)

$$P(t) = P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{R_2 + \varphi^T(t)P(t-1)\varphi(t)} + \gamma^2 R_1$$

$$L(t) = \frac{P(t-1)\varphi(t)}{R_2 + \varphi^T(t)P(t-1)\varphi(t)} \tag{59}$$

As γ tends to zero, we can, using averaging theory replace $\varphi(t)\varphi^T(t)$ by Q as in Section 3.3, and also disregard the term $\varphi^T(t)P(t-1)\varphi(t)$ in comparison with R_2. As γ tends to zero, $P(t)$ will thus behave as

$$P(t) \equiv \gamma P \tag{60}$$

where P is the solution of

$$PQP = R_2 R_1 \tag{61}$$

The asymptotic gain in (19) will be

$$L(t) = \bar{P}\varphi(t)$$

with

$$\bar{P} = \gamma\frac{P}{R_2} \tag{62}$$

Let us now replace $\Phi(t,k)$ in (43) by $(I-\bar{P}Q)^{t-k}$, using (54). Then we have

$$\tilde{\theta}(t) = (I-\bar{P}Q)\tilde{\theta}(t-1) + \bar{P}\varphi(t)e(t) - w(t) \tag{63}$$

From this we can derive the mean square error

$$\Pi(t) = E\tilde{\theta}(t)\tilde{\theta}(t)$$

as

$$\Pi(t) = (I-\bar{P}Q)\Pi(t-1)(I-\bar{P}Q)^T + R_2^0(t)\bar{P}Q\bar{P} + R_1^0(t) \tag{64}$$

Here $R_2^0(t)$ and $R_1^0(t)$ are the true variances of $e(t)$ and $w(t)$, respectively.

The expression (64) describes the propagation of the mean square error $\Pi(t)$ by a linear difference equation. It is valid for slow drift, as well as for sudden changes in $\theta_0(t)$, i.e. there is no assumption that $R_1^0(t)$ should be small and constant. The only assumption concerns small gain in the recursive identification algorithm, i.e. (58).

If we specialize to constant drift for the true system, i.e.

$$R_1^0(t) \equiv R_1^0 \qquad R_2^0(t) \equiv R_2^0 \tag{65}$$

we find that $\Pi(t)$ converges to the solution $\bar{\Pi}$ of

$$\bar{P}Q\bar{\Pi} + \bar{\Pi}Q\bar{P} = R_2^0\bar{P}Q\bar{P} + R_2^0 \tag{66}$$

Using (62) and (61) we obtain

$$PQ\bar{\Pi} + \bar{\Pi}QP = \gamma R_2^0 R_1 + \frac{1}{\gamma}R_2 R_1^0 \tag{67}$$

$$PQP = R_2 R_1 \tag{68}$$

This expression is the most general expression for the asymptotic mean square error $\bar{\Pi}$ under slow adaptation and steady parameter drift.

Viewing R_1 and R_2 as design variables, we realize from Section 2 that $\bar{\Pi}$ is minimized by the choices

$$R_2 = R_2^0 \qquad \gamma^2 R_1 = R_1^0$$

which gives

$$PQ\bar{\Pi} + \bar{\Pi}QP = \frac{2}{\gamma}R_2^0 R_1^0$$

$$PQP = R_2^0 R_1^0 \tag{69}$$

Let us now apply (67)-(68) to the ad hoc algorithms of Section 2.

The LMS algorithm

Comparing with (33)-(35) we easily find that asymptotic results for the LMS algorithm are obtained by taking

$$\gamma = \mu \qquad R_1 = Q \qquad \bar{P} = I \qquad R_2 = 1$$

which gives

$$Q\bar{\Pi} + \bar{\Pi}Q = \mu R_2^0 Q + \frac{1}{\mu} R_1^0 \tag{70}$$

The forgetting factor algorithm

Compared with (29)-(32) we find that the forgetting factor algorithm with constant γ (close to 1) corresponds to

$$R_2 = 1;\ \gamma = \frac{1}{\gamma} - 1;\ R_1 = \ P = Q^{-1};\ \bar{P} = \gamma Q^{-1}$$

which gives the following explicit expression for the mean square error:

$$\bar{\Pi} = \frac{\gamma}{2} \cdot R_2^0 Q^{-1} + \frac{1}{2\gamma} R_1^0 \tag{71}$$

It is interesting to note that the expression (71) can also be derived by direct manipulations of the expression (45), without reference to explicit averaging results. See Appendix A (omitted in this version).

From expressions like (71) we directly see the character of the trade-off in the choice of adaptation gain γ: Slow adaptation, i.e. a small γ gives a small contribution from the observations noise $R_2^0 Q^{-1}$ but a large contribution from the tracking error R_1^0 and vice versa. The problem clearly is to balance these contributions.

5 EVALUATION OF THE ERROR IN THE FREQUENCY DOMAIN

The expressions for the mean square error that we derived in the previous section are somewhat implicit. In Gunnarsson (1986) and Gunnarsson and Ljung (1987) explicit expressions for the mean square error of a corresponding transfer function estimate were derived. The results can be summarized as follows:

consider a FIR-model, where $\varphi(t)$ contains only lagged inputs:

$$y(t) = \varphi^T(t)\theta = \sum_{k=1}^{d} g_k u(t-k) \tag{72}$$

The corresponding transfer function then is

$$G(e^{i\omega}) = \sum_{k=1}^{d} g_k e^{-ik\omega} = W_d^*(\omega)\theta \tag{73}$$

where

$$W_d(\omega) = [e^{i\omega} .... e^{di\omega}] \tag{74}$$

and where "*" denotes transpose and complex conjugate.

The mean square error of the transfer function estimate at frequency ω then is

$$\pi_d(\omega) = W_d^*(\omega)\bar{\Pi}W_d(\omega) \tag{75}$$

where $\bar{\Pi}$ is the mean square error matrix for the parameters, as derived in Section 4.

The key properties to be used are as follows:

If

$$\frac{1}{d} W_d^*(\omega) A W_d(\omega) \to a(\omega) \qquad \text{as} \qquad d \to \infty$$

and

$$\frac{1}{d} W_d^*(\omega) B W_d(\omega) \to b(\omega) \qquad \text{as} \qquad d \to \infty$$

then (under some regularity conditions, see Gunnarsson and Ljung (1987))

$$\frac{1}{d} W_d^*(\omega) AB W_d(\omega) \to a(\omega)b(\omega) \qquad \text{as} \qquad d \to \infty$$

and

$$\frac{1}{d} W_d^*(\omega) A^{-1} W_a(\omega) \to \frac{1}{a(\omega)} \qquad \text{as} \qquad d \to \infty.$$

Also

$$\frac{1}{d} W_a^*(\omega) Q W_d(\omega) \to \Phi_u(\omega) \qquad \text{as} \qquad d \to \infty$$

where $\Phi_u(\omega$ is the spectrum of the input $\{u(t)\}$, and Q is defined by (53).

We are now going to apply these results to (66)-(67) by evaluating

$$\bar{\pi}(\omega) = \lim_{d\to\infty} \frac{1}{d}\pi_d(\omega) = \lim_{d\to\infty} \frac{1}{d} W_d^*(\omega)\bar{\Pi} W_d(\omega) \qquad (76)$$

as the order, d, of the FIR model (72) tends to infinity. For large order models we will thus have that the mean square error of the transfer function estimate at frequency ω is given by

$$\pi_d(\omega) \approx d \cdot \bar{\pi}(\omega). \qquad (77)$$

Introduce

$$p(\omega) = \lim_{d\to\infty} \frac{1}{d} W_d^*(\omega) P W_d(\omega)$$

$$r_1(\omega) = \lim_{d\to\infty} \frac{1}{d} W_d^*(\omega) R_1 W_d(\omega)$$

$$r_1^0(\omega) = \lim_{d\to\infty} \frac{1}{d} W_d^*(\omega) R_1^0 W_d(\omega)$$

From (67) we then find, by applying the limiting procedure to both members:

$$p^2(\omega)\Phi_u(\omega) = R_2 r_1(\omega)$$

or

$$p(\omega) = \sqrt{\frac{R_2 r_1(\omega)}{\Phi_u(\omega)}} \qquad (78)$$

Similarly (66) gives

$$2p(\omega)\Phi_u(\omega)\bar{\pi}(\omega) = \gamma R_2^0 r_1(\omega) + \frac{1}{\gamma} R_2 r_1^0(\omega)$$

or

$$\bar{\pi}(\omega) = \frac{1}{2}\sqrt{\frac{r_1(\omega)R_2}{\Phi_u(\omega)}}\left[\gamma \cdot \frac{R_2^0}{R_2} + \frac{1}{\gamma}\frac{r_1^0(\omega)}{r_1(\omega)}\right] \qquad (79)$$

Expressions (77) and (79) give an explicit and useful description of how the accuracy of the estimate varies with frequency and with the design variables $r_1(\omega)$ and R_2.

It is easy to explicitly minimize (79) with respect to these variables, and this gives, as it should,

$$r_1(\omega) = \frac{1}{\gamma^2} r_1^0(\omega); \qquad R_2 = R_2^0.$$

We also obtain for the LMS-algorithm from (69)

$$\bar{\pi}(\omega) = \frac{1}{2}\left[\gamma \cdot R_2^0 + \frac{1}{\gamma}\frac{r_1^0(\omega)}{\Phi_u(\omega)}\right] \qquad (80)$$

and for the forgetting factor algorithm from (70)

$$\bar{\pi}(\omega) = \frac{1}{2}\left(\gamma \frac{R_2^0}{\Phi_u(\omega)} + \frac{1}{\gamma} \cdot r_1^0(\omega)\right) \qquad (81)$$

The results (79)-(81) thus describe how the basic recursive identification algorithm performs under small gain and under steady parameter drift. Recall that the expression (64) can also deal with abrupt changes of the true system (but still slow adaptation on the part of the algorithm). Hence a frequency-domain theory for transient responses by slowly adapting algorithms can be developed along the same lines as above. Such results are given in Gunnarsson and Ljung (1987).

6 CONCLUSIONS

We have studied the tracking of time-varying systems in this paper. The focus has been on model structures of the linear regression type, but for small mean square errors (and most analytic results deal with such errors) the ideas and techniques apply also to general structures (see (26)).

For a random walk model for the true system parameters, the optimal tracking algorithm can be derived from the Kalman filter. Commonly used ad-hoc variants can be interpreted within the same framework as corresponding to specific assumptions about the covariance matrix of the random walk increments.

A number of more or less explicit results about the parameter mean square error have been derived under the assumption of slow adaptation. These results are particularly easy to interpret in the frequency domain, in a form that is asymptotic in the model order.

It should be particularly stressed that the two ad-hoc approaches LMS, and the forgetting factor algorithm are just that - ad hoc - in the case of a time-varying system. The computationally more complex forgetting factor algorithm is not uniformly better, in the slow adaptation case, than LMS neither in dealing with the transient response to an abrupt changes nor in dealing with steady parameter drift. Which algorithm performs better in these cases depend on the covariance matrix of the parameter changes (i.e. $R^0 1(t)$ in our notation). On the other hand - and we did not discuss that explicitly here - the forgetting factor algorithm is superior to LMS sin quickly getting back to good estimates after an abrupt change in the true system, provided it is allowed to increase its adaptation rate and take larger steps.

REFERENCES

Andersson, P. (1985). Adaptive forgetting in recursive identification through multiple models. *Int. J. Control*, vol 42, no 5, pp 1175-1193.

Basseville, M. and A. Benveniste (1986) (Ed). Detection of abrupt changes in signals and dynamic systems, LNCIS 77, Springer Verlag, Berlin.

Benveniste, A. (1987). Design of adaptive algorithms for the tracking of time-varying systems. *Int Journal of Adaptive Control and Signal Processing*, vol 1, pp 3-29.

Benveniste, A., M. Metivier and P. Priouret (1987). *Algorithmes Adaptatifs et Approximations Stochastiques*, Masson, Paris.

Benveniste, A. and G. Ruget (1982). A measure of the tracking capability of recursive stochastic algorithms with constant gains. *IEEE Trans. Automatic Control*, vol. AC-27, pp 639-649.

Bitmead, R. R. (1984). Persistence of excitation conditions and the convergence of adaptive schemes. *IEEE Trans Info Theory*, vol IT-30, pp 183-191.

Bitmead, R. R. and B. D. O. Anderson (1980). Performance of adaptive estimation algorithms in dependent random environments. *IEEE Trans Autom Control*, vol. AC-25, pp 788-794.

Eweda, E. and O. Macchi (1985). Tracking error bounds of adaptive non-stationary filtering, *Automatica*, vol 21, pp 293-302.

Fortesque, T. R. et al (1981). Implementation of self-tuning regulators with variable forgetting factors. *Automatica*, vol. 17, pp 831-835.

Gardner, W. A. (1984). Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis and critique. *Signal Processing*, vol 6, pp 113-133.

Gunnarsson, S. (1986). On the variance of recursive transfer function estimates of systems with FIR-structure. *Proc. 25th IEEE Conference on Decision and Control*, Athens, Greece, pp 2029-2030.

Gunnarsson, S. and L. Ljung (1987). Frequency domain tracking characteristics of adaptive algorithms, Report LiTH-ISY-I-0686, Department of Electrical Engineering, Linköping University, Sweden.

Hägglund, T. (1984) Adaptive control of systems subject to large parameter changes. *Proc. 9th IFAC World Congress*, Budapest Hungary, pp. 993-998.

Isaksson, A. (1987). Identification of time-varying systems through adaptive Kalman filtering. Proc. 10th IFAC World Congress, Munich, July 1987, pp 306-311.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press, New York.

Khasminski, R. Z. (1966). On stochastic processes defined by differential equations with small parameter. *Theory of Probability and Its Applications*, vol. 11, pp. 211-288.

Kushner, H. J. and D. S. Clark (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York.

Kushner, H. J. and H. Huang (1981). Asymptotic properties of stochastic approximations with constant coefficients. *SIAM J. Control and Optimization*, vol. 19, pp 87-105.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control*, vol. AC-22, pp 551-575.

Ljung, L. (1986) Error propagation in adaptation algorithms with poorly exciting signals. *Annales des Telecommunications*, May-June 1986.

Ljung, L. (1987). *System Identification - Theory for the User*, Prentice-Hall, Inc. Englewood Cliffs, N J.

Ljung, L and T. Söderström (1983). *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, Mass.

Macchi, O. and E. Eweda (1983). Second order convergence analysis of stochastic adaptive linear filtering. *IEEE Trans. Automatic Control*, vol. AC-28, pp. 76-85.

Millnert, M. (1987). Identification of ARX models with Markovian parameters. *Int Journal of Control*, vol 45, no 6, pp 2045.

Wang, J. G. and Z. L. Deng (1986). Simulation of a newly designed adaptive controller. IFAC Symposium on Simulation of Control Systems, Vienna, pp. 93-196.

Widrow, B., J. M. McCool, M. G. Larimore and C. R. Johnson Jr (1976). Stationary and non-stationary learning characteristics of the LMS adaptive filter. *Proc. IEEE*, vol 64, pp 1151-1162.

Widrow, B. and S. Stearns (1984). *Adaptive Signal Processing*. Prentice-Hall, Englewood Cliffs, N.J.