THE RECURSIVE IDENTIFICATION PROBLEM IN CONTROL AND SIGNAL PROCESSING APPLICATIONS

LENNART LJUNG

*Division of Automatic Control, Department of Electrical Engineering, Linköping University,
Linköping, Sweden*

Methods for sequential estimation of parameters in models of dynamical systems are reviewed. Particular attention is paid to the roots of the algorithms in off-line identification problems. The role of the prediction of future outputs and its gradient with respect to the parameters is stressed. Some fundamental asymptotic properties of the algorithms are quoted.

1. Introduction

A common problem in many branches of modern science is that of constructing mathematical models of various real-life objects and systems. In control theory applications the term system identification is used for the problem of building mathematical models of dynamical systems from observed values of input and output signals. By the term *recursive identification* is then meant the problem of calculating the models on-line, sequentially, at the same time as the data are collected from the system. The same problem is known as *sequential parameter estimation* in the statistical literature and *adaptive* (or *adaptation*) *algorithms* in the signal processing field.

In this contribution we shall highlight some basic ideas of recursive identification techniques, as well as stress the close relationship to the general system identification problem. The presentation is by necessity a brief one. Further details on the topic can be found in [1], [2], [3].

By recursive identification we mean the following: A set of candidate models $\mathcal{M}$ is given. The set is parametrized by a finite dimensional parameter vector θ , such that a given model in the set is denoted by $\mathcal{M}(\theta)$. Also, a sequence of measurements $z(t)$; $t = 1, 2, \dots$ from the system in question is given, such that $z(t)$ is the measurement vector obtained at time t . The problem is to, at time t , determine a model $\mathcal{M}(\hat{\theta}(t))$ in the set $\mathcal{M}$ that

describes the data $z(j); j = 1, \dots, t$ well, in some sense. The calculation of $\hat{\theta}(t)$ is constrained to be recursive, i.e., it should be calculated using only a fixed, and t -independent amount of memory space and computations. Conceptually we could write

$$\begin{aligned} S(t) &= F(t, S(t-1), z(t)), \\ \hat{\theta}(t) &= H(S(t)) \end{aligned} \quad (1)$$

where $S(t)$ is a finite dimensional auxiliary vector ("an information state"). Usually, as we shall see, the algorithm takes a more specific form.

Several approaches to the construction of such recursive algorithms can be taken. We shall in these notes use an approach where the algorithms are derived from off-line system identification methods. In Section 2 we give a general description of models for dynamical systems. Basic off-line identification methods are treated in Section 3 and recursive identification techniques are displayed in Section 4. Section 5 deals with the asymptotic properties of the obtained estimates.

2. Models

Many different model sets can be used for a given system. It is the task of the user to find a suitable model set that matches the purpose of the modelling. It will be convenient to single out one part of the measurement vector $z(t)$ as an "output" $y(t)$ (which could be equal to $z(t)$) and term the rest an "input" $u(t)$:

$$z^T(t) = (y^T(t), u^T(t)). \quad (2)$$

A typical feature of any model $\mathcal{M}(\theta)$ is that it provides a prediction or "guess" of the next output, given previous measurements:

$$\begin{aligned} \hat{y}(t|\theta) &= g_{\mathcal{M}}(\theta; t, z^{t-1}), \\ z^{t-1} &= (z(t-1), \dots, z(1)). \end{aligned} \quad (3)$$

Here $g_{\mathcal{M}}(\theta; t, \cdot)$ is a function from R^N to R^p where $p = \dim y$ and $N = (t-1)\dim z$. In a common family of models the prediction is computed by linear, finite dimensional filtering:

$$\begin{aligned} \varphi(t+1, \theta) &= F(\theta) \varphi(t, \theta) + G(\theta) z(t), \\ \hat{y}(t|\theta) &= H(\theta) \varphi(t, \theta). \end{aligned} \quad (4)$$

As a simple example we may take a linear regression model

$$\hat{y}(t|\theta) = -a_1 y(t-1) - \dots - a_n y(t-n) + b_1 u(t-1) + \dots + b_m u(t-m) \quad (5)$$

where

$$\theta^T = (a_1, \dots, a_n, b_1, \dots, b_m). \quad (6)$$

With the notation

$$\varphi^T(t) = (-y(t-1), \dots, -y(t-n), u(t-1), \dots, u(t-m)) \quad (7)$$

we could rewrite (5) in the following convenient form

$$\hat{y}(t|\theta) = \theta^T \varphi(t). \quad (8)$$

The general descriptions (3) and (4) include models much more sophisticated than (8). Let us here only mention an ARMA model of a signal $y(t)$:

$$y(t) + ay(t-1) = e(t) + ce(t-1). \quad (9)$$

Here $\{e(t)\}$ is white noise, and we have confined ourselves to a first order model for simplicity.

It is fairly straightforward to show that with $\theta^T = (a, c)$, the predictor corresponding to (9) is given by

$$\hat{y}(t|\theta) + c\hat{y}(t-1|\theta) = (c-a)y(t-1). \quad (10)$$

For more detailed exposé over different model sets and how they relate to the general description (3) we may refer to [3].

3. Identification criteria

From the predicted output $\hat{y}(t|\theta)$ according to the model $\mathcal{M}(\theta)$ and the recorded output $y(t)$ we can compute the prediction error

$$\varepsilon(t, \theta) = y(t) - \hat{y}(t|\theta). \quad (11)$$

The identification problem is to determine a suitable estimate $\hat{\theta}_N$, based on z^N . Almost all identification methods are based on the following idea.

Select $\hat{\theta}_N$, so that the sequence $\varepsilon(t, \hat{\theta}_N)$, $t = 1, \dots, N$, becomes "small". (12)

Two common ways of qualifying what should be meant by "small" are the following ones:

Criterion minimization. Select $\hat{\theta}_N$ so that the criterion

$$V_N(\theta, z^N) = \frac{1}{N} \sum_{t=1}^N |\varepsilon(t, \theta)|^2 \quad (13)$$

is minimized. For the model (8) this approach gives the well-known least squares method.

Of course, other "norms", than the quadratic one in (13) can be chosen. A general criterion would be

$$V_N(\theta, z^N) = \frac{1}{N} \sum_{t=1}^N l(t, \theta, \varepsilon(t, \theta)) \quad (14)$$

where $l(t, \theta, \varepsilon)$ is a suitably chosen scalar-valued function. Within the family (14) we find several well-known methods, such as the maximum likelihood method, corresponding to

$$l(t, \theta, \varepsilon) = -\log f(t, \theta, \varepsilon),$$

where f is the probability density function for the prediction error $\varepsilon(t, \theta)$. Other examples of methods within the family (14) are robust methods, output error methods, etc.

Correlation approach. Select a sequence of vectors $\zeta(t, \theta)$, $t = 1, \dots, N$,

$$\zeta(t, \theta) = h(z^{t-1}, \theta)$$

representing relevant information about the system available at time $t-1$. Then choose $\hat{\theta}_N$ so that $\{\varepsilon(t, \hat{\theta}_N)\}$ is uncorrelated with $\{\zeta(t, \hat{\theta}_N)\}$:

$$\frac{1}{N} \sum_{t=1}^N \zeta(t, \theta) \varepsilon(t, \theta) = 0 \Rightarrow \theta = \hat{\theta}_N. \quad (15)$$

For the model (8) this gives the well-known instrumental variable method. Then $\zeta(t, \theta)$ is often chosen as a function of past inputs, independent of θ . Within the family (15) we also find pseudolinear regressions and model reference techniques.

Now, for a practical identification method we also need a numerical algorithm to actually minimize (13) or solve (15). We shall now briefly discuss some general methods for the first problem.

Except in the case (8), where (13) is quadratic in θ , the minimization is performed by iterative techniques. With i as the iteration number we proceed as follows:

$$\hat{\theta}_N^{(i)} = \hat{\theta}_N^{(i-1)} - \mu_N^{(i)} [R_N^{(i)}]^{-1} V'_N(\hat{\theta}_N^{(i-1)}, z^N). \quad (16)$$

Here V'_N is the gradient of the criterion

$$V'_N(\theta, z^N) = -\frac{1}{N} \sum_{t=1}^N \psi(t, \theta) \varepsilon(t, \theta) \quad (17)$$

where we defined

$$\psi(t, \theta) = \frac{d}{d\theta} \hat{y}(t|\theta). \quad (18)$$

R_N is a matrix that possibly modifies the search direction from the gradient one. Typical choices are:

$$R_N^{(i)} = I \quad \text{or} \quad R_N^{(i)} = |V_N''(\theta_N^{(i-1)}, z^N)|^2 \cdot I \quad (19)$$

(gradient or steepest descent, un-normalized and normalized)

$$R_N^{(i)} = V_N''(\hat{\theta}_N^{(i-1)}, z^N) \quad (\text{Newton}) \quad (20)$$

$$R_N^{(i)} = \frac{1}{N} \sum_1^N \psi(t, \hat{\theta}_N^{(i-1)}) \psi^T(t, \hat{\theta}_N^{(i-1)}) \quad (\text{Gauss-Newton}) \quad (21)$$

The scalar μ is chosen so as to guarantee a decrease in the criterion value.

Similar approaches apply also to the solution of (15).

In addition to these general techniques, special schemes for special model sets have also been devised that utilize some particular structures of $\varepsilon(t, \theta)$ and $\psi(t, \theta)$.

For specific model sets the iterative technique (16), (21) need only be complemented by schemes for computation of $\hat{y}(t|\theta)$ and $\psi(t, \theta)$ for any given value of θ . The general linear model (4) gives, conceptually, a linear filter for this:

$$\begin{aligned} \xi(t+1, \theta) &= A(\theta) \xi(t, \theta) + B(\theta) z(t), \\ \begin{pmatrix} \hat{y}(t|\theta) \\ \text{col} \psi(t, \theta) \end{pmatrix} &= C(\theta) \xi(t, \theta). \end{aligned} \quad (22)$$

Here "col" means that the columns of ψ are stacked on top of each other.

4. Recursive identification

It is clear that the general scheme (16) does in general not comply with the structure (1) for a recursive algorithm. However, we can modify (16) to achieve this. Conceptually, this is obtained by making one additional iteration $i-1 \rightarrow i$ at the same time as the data record is advanced from $N-1$ to N . From (16) we thus obtain

$$\hat{\theta}(t) = \hat{\theta}(t-1) - \mu(t) R^{-1}(t) V_t'(\hat{\theta}(t-1), z^t)$$

where we introduced the notation

$$\hat{\theta}(t) = \hat{\theta}_t^{(t)} \quad \text{etc.}$$

Now, from (17) we see that

$$V_t'(\theta_z^t) = \frac{t-1}{t} V_{t-1}'(\theta, z^{t-1}) - \frac{1}{t} \psi(t, \theta) \varepsilon(t, \theta). \quad (23)$$

Introducing the approximation $\hat{\theta}(t-1)$ that actually minimized $V_{t-1}(\theta, z^{t-1})$, i.e.,

$$V'_{t-1}(\hat{\theta}(t-1), z^{t-1}) = 0 \quad (24)$$

now gives the algorithm

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \gamma(t) R^{-1}(t) \psi(t, \hat{\theta}(t-1)) \varepsilon(t, \hat{\theta}(t-1)) \quad (25)$$

where we introduced

$$\gamma(t) = \mu(t)/t. \quad (26)$$

In case $\psi(t, \theta)$ and $\varepsilon(t, \theta)$ can be computed for any θ using only a fixed amount (t -independent) of past data, then (25) is indeed a recursive algorithm. This is the case for the model (8), where

$$\begin{aligned} \varepsilon(t, \theta) &= y(t) - \theta^T \varphi(t), \\ \psi(t, \theta) &= \varphi(t). \end{aligned} \quad (27)$$

With $\mu = 1$,

$$R(t) = \frac{1}{t} \sum_{k=1}^t \varphi(k) \varphi^T(k) \quad (28)$$

(the Newton or the Gauss-Newton choice), (25)–(28) is the celebrated recursive least squares algorithm. For the gradient choice (19) we obtain the LMS-algorithm of Widrow, well-known from many signal processing applications, e.g., [5].

In general, it is not possible to compute $\varepsilon(t, \hat{\theta}(t-1))$ and $\psi(t, \hat{\theta}(t-1))$ using a fixed amount of data. In such a case these quantities have to be replaced by recursively computed quantities, denoted by $\varepsilon(t)$ and $\psi(t)$. The following simple example shows how to do this:

Consider the model (9). From (10) we easily find how to compute $\varepsilon(t, \theta)$ and $\psi(t, \theta)$ for any θ :

$$\varepsilon(t, \theta) + c\varepsilon(t-1, \theta) = y(t) + ay(t-1), \quad (29)$$

$$\psi(t, \theta) + c\psi(t-1, \theta) = \begin{pmatrix} -y(t-1) \\ \varepsilon(t-1, \theta) \end{pmatrix} \quad (30)$$

Now, the obvious way of computing the approximations $\varepsilon(t)$ and $\psi(t)$ is to replace the nominal value θ in (29)–(30) by the current estimate $\hat{\theta}(t-1)$:

$$\varepsilon(t) + \hat{c}(t-1)\varepsilon(t-1) = y(t) + \hat{a}(t-1)y(t-1), \quad (31)$$

$$\psi(t) + \hat{c}(t-1)\psi(t-1) = \begin{pmatrix} -y(t-1) \\ \varepsilon(t-1) \end{pmatrix}. \quad (32)$$

The complete algorithm for estimating a and c in the ARMA model (9) thus is (31), (32) together with:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \gamma(t) R^{-1}(t) \psi(t) \varepsilon(t). \quad (33)$$

For the Gauss-Newton algorithm we take

$$R(t) = R(t-1) + \frac{1}{t} [\psi(t) \psi^T(t) - R(t-1)]. \quad (34)$$

Obviously, the same idea can be applied to any model set (3):

Derive the equations defining $\varepsilon(t, \theta)$ and $\psi(t, \theta)$ for any nominal value θ . In order to obtain the approximations $\varepsilon(t)$ and $\psi(t)$ replace at each time instant the nominal value in these expressions by the latest available estimate.

Applied to the linear model (4) (see also (22)), this gives a general recursive Gauss-Newton algorithm

$$\begin{aligned} \varepsilon(t) &= y(t) - \hat{y}(t), \\ \hat{\theta}(t) &= \hat{\theta}(t-1) + \gamma(t) R^{-1}(t) \psi(t) \varepsilon(t), \\ \xi(t+1) &= A(\hat{\theta}(t)) \xi(t) + B(\hat{\theta}(t)) z(t), \\ \begin{pmatrix} \hat{y}(t+1) \\ \text{col } \psi(t+1) \end{pmatrix} &= C(\hat{\theta}(t)) \xi(t), \\ R(t) &= R(t-1) + \gamma(t) [\psi(t) \psi^T(t) - R(t-1)]. \end{aligned} \quad (35)$$

For a general criterion function $l(\varepsilon(t, \theta))$ as in (14) the factor $\psi(t) \varepsilon(t)$ should be replaced by $\psi(t) l'(\varepsilon(t))$.

The algorithm (35) can be applied to arbitrary model sets, yielding a large family of recursive identification methods. These are well known in special cases. We can also, for special models, use certain approximations for the gradient ψ to yield other families of methods; see, e.g., [3] or [2].

5. Asymptotic properties of the recursive estimates

The recursively computed estimate $\hat{\theta}(t)$ is a fairly complicated function of z^t . Its statistical properties will thus be inherited from those of z^t in a complex manner. It is possible, though, to compute the asymptotic statistical properties of $\hat{\theta}(t)$ as t approaches infinity, for the general algorithm (35) applied to an arbitrary (linear) model set. The result can be quoted as follows:

For any positive definite choice of $R(t)$ in (35), $\hat{\theta}(t)$ converges with probability one to a (local) minimum of

$$V(\theta) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N E |\varepsilon(t, \theta)|^2 \quad (36)$$

as t tends to infinity, if $\gamma(t) \rightarrow 0$.

If θ^* is such a limit point where $\bar{V}(\theta^*)$ is nonsingular, then, for the Gauss–Newton choice (35) and supposing $t \cdot \gamma(t) \rightarrow 1$ as $t \rightarrow \infty$, we have $\sqrt{t}(\hat{\theta}(t) - \theta^*) \in \text{As } N(0, P)$ (i.e. the right hand side converges in distribution to the normal distribution with zero mean and covariance matrix P .) (37)

An explicit expression for P can be given. It is fairly complicated, though. It simplifies in the following special case.

If θ^* is such that $\{\varepsilon(t, \theta^*)\}$ is a sequence of independent random variables with zero mean and covariance matrix I , then $P = [E\psi(t, \theta^*)\psi^T(t, \theta^*)]^{-1}$. (38)

We might add that the results (36)–(38) are the same that hold for the off-line estimate $\hat{\theta}_N$ defined by minimization of (13).

Notice that there is no assumption that the true system can be exactly described within the model set for (36)–(37) to hold. Result (36) thus states that $\mathcal{M}(\hat{\theta}(t))$ converges to the best approximation of the system (if not caught in a non-global local minimum).

Notice also that (38) equals the Cramér–Rao lower bound for the covariance of any estimate, provided the conditions mentioned in (38) hold and the $\varepsilon(t, \theta)$ are gaussian. The recursive identification algorithm (33)–(35) is thus asymptotically efficient. Asymptotically, no other algorithm can perform better for the chosen model set.

Proofs for the statements in this section can be found in [3].

6. Conclusions

In this contribution we have discussed the close relationship between general recursive identification algorithm and the corresponding off-line identification problem. This in turn makes the connection between adaptive algorithms of various kinds and basic identification concepts. This connection is perhaps worthwhile to stress.

We have also shown the fundamental role of prediction for this problem.

References

- [1] L. Ljung, *Convergence of recursive estimators*, in: *Proc. 5th IFAC Symposium on Identification*, Darmstadt, September 1979, Survey paper S8, 131–144.
- [2] L. Ljung, *Analysis of a general recursive prediction error algorithm*, *Automatica*, **17** (1981), 89–99.

- [3] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*, MIT PRESS, Cambridge, MA, 1983.
- [4] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Stat., **22** (1951), 400–407.
- [5] B. Widrow et al., *Adaptive noise cancelling: Principles and Applications*, Proc. IEE-E **63** (1975), 1962–1716.

*Presented to the semester
Sequential Methods in Statistics
September 7–December 11, 1981*