Model Order, Convergence Rates and Information Content in Noisy Partial Realizations

Victor E. DeBrumert and A. A. (Louis) Beext

†School of Electrical Engineering and Computer Science
The University of Oklahoma
Norman, OK 73019 USA

‡Bradley Department of Electrical Engineering
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061 USA

Abstract

We show that convergence rates of recursive algorithms for parameter estimation from noisy partial realizations depend on the structure of the chosen model. The model is analyzed by considering information unique to the parameterization -- the sensitivity and interconnectedness of the model parameters. The convergence analysis is independent of model order.

§1. Introduction

In previous papers (4,6,7), we developed output-error identification algorithms whose gradients could be computed quickly and efficiently in closed form. These algorithms hinge on the fact that the gradients are ARMA auto- and cross-covariances. These algorithms are deterministic, in the sense that the same measurements are used repeatedly in the recursive algorithms. We now extend the algorithms to the case where many different noisy realizations of the same partial realization are used in the identification. The relative convergence rates in this stochastic case are similar to the previously explored convergence. The convergence rate is predicted by a conjecture, which is based on a measure determined from the Fisher information matrix of the model parameters. We have called the measure (denoted by τ), the 'relative convergence time constant.' This information measure combines the ideas of model parameter sensitivity and parameter interconnectedness. The measure depends on the structure of the chosen model. In this paper, we also show how the order of the chosen model affects the convergence rates. The model parameter sensitivity is shown to have an important impact on initial convergence. The parameter conditioning has great impact only when the estimates approach the solution point. A model which is highly sensitive to fluctuations in its parameters converges more quickly than a model less sensitive to its parameters when the initialization is started close to the true parameters (the solution point).

In §2, we introduce the stochastic algorithms. In §3 we discuss the convergence behavior of these algorithms. In §4 we examine the convergence behavior when the model is of a different order than the identified system. We conclude in §5.

§2. The Identification Algorithms

In this section, a gradient algorithm (1,2,13) is discussed which is suitable for identifying the system parameters of a stable, but otherwise arbitrary, state-space structure from a set of noisy impulse response measurements. The specific equations for various state-space structures are given in (4). First, we make a few definitions:

1. y_n, the noisy impulse response measurements at time n
2. h_n(θ), the impulse response at time n corresponding to parameterization θ
3. w_n, a positive weight at time n (0 < w_n ≤ 1)
4. ŷ_n = y_n - h_n(θ), the impulse response error at time n
5. the error criterion:
   \[ J_N = \frac{1}{2} \sum_{n=1}^{N} w_n \| θ_n - θ \|^2 \]
6. θ_n, the (Np x 1) vector of identified parameters at time n
7. Φ_n, the vector of parameter gradients at time n (also called the 'information vector' 14)

We then can define the following Gauss-Newton algorithm (13):

1. \[ θ_{n+1} = θ_n + Φ_n (y_n - h_n(θ_n)) \]
2. \[ Φ_n = \frac{P_n \Phi_n}{\Phi_n P_n \Phi_n - w_n} \]
3. \[ P_n^{+} = P_n \Phi_n P_n \Phi_n - w_n \]
4. \[ P_n \frac{∂ h_n(θ_n)}{∂ θ_{n+1}} = 1 \]
5. \[ \frac{∂ H(θ_n, θ_{n+1})}{∂ θ_{n+1}} = 0 \]

Adjustments can be made to the above algorithm when numerical accuracy is paramount. The most commonly used modification is the square root algorithm due to Potter (13). This algorithm is based on the Cholesky factorization of the positive definite, Hermitian matrix P (given in item 3, above). Other factorizations of P exist, and of course may be used. However, the improvements require more computations than the basic algorithm given above.

General Canonical-Form State-Space Algorithms

Consider the stable single-input, single-output state-space system given by the 4-tuple (A,b,c,d) and...
with transfer function \( H(z) = \frac{c(z-A)^{-1}b+d}{z^2-A_1z-A_0} \). The partial derivatives needed in the gradient computation step (4, above) are then recognized as \( (5) \):

\[
\begin{align*}
\frac{\partial H(z)}{\partial b_j} &= (z-A)^{-1}; \\
\frac{\partial H(z)}{\partial A_j} &= 1; (z-A)^{-1} \quad (1) \\
\frac{\partial H(z)}{\partial a_k} &= \frac{\partial H(z)}{\partial b_j} \frac{\partial b_j}{\partial a_k} \\
\frac{\partial H(z)}{\partial \omega_k} &= \frac{\partial H(z)}{\partial b_j} \frac{\partial b_j}{\partial \omega_k}
\end{align*}
\]

where \( 1 \) is the \( n \)th Euclidean basis vector, i.e., an \( (n \times 1) \) vector whose components are all zero except for the \( n \)th element, which is equal to 1.

Computation of the gradients necessary in the algorithm can be accomplished via 2 methods. The method most consistent with use in a recursive algorithm is based on the fact that the gradients are (in general) ARMA cross-covariances. As has been previously shown (3.5,8), computations of the form

\[
f_{\nu} = \frac{1}{2\pi f_j} \int \frac{x(z)}{z} \frac{dz}{z} \quad (2)
\]

may be determined in closed form in an efficient manner. Alternatively, the gradients can be computed by implementing the partial derivative systems, each driven by a unit sample sequence as its input. This method is numerically superior to the ARMA cross-covariance method in that the identified parameters are used directly in the gradient computation. However, the numerical accuracy is countervaled by its speed disadvantage, as the filtering method can be significantly slower than the direct computation indicated in equation (2).

When the impulse response measurement data is corrupted by additive Gaussian noise, the Hessian of \( J_n \) is equal to \( F \) (14,15), and so the estimate of the Hessian \( F^{-1} \) is thus an approximation to the parameter Fisher information matrix. Thus, the use of the term "information vector" for the gradient vector \( \Phi \).

53. Algorithmic Convergence Behavior

The deterministic algorithms use only one (possibly noisy) data record to determine the model parameter estimates. Thus, the convergence is to a set of model parameters which represents a minimum of \( J_n \) (11). Repeated noisy impulse response measurements are made and used in the stochastic algorithms, and so convergence is to a set of model parameters which represents a minimum of \( J_n \) over all measurement sets. For output error methods in the presence of additive Gaussian noise, these converged-to-parameters are the true system parameters (if enough measurement data is taken and used), and so the noise is averaged out in this case (12). The stochastic algorithms are also biased if the additive noise is not Gaussian.

We base our convergence results on the shape of the criterion \( J_n \) so that we are not considering stochastic convergence in this paper. First, we show that \( J_n \) is nearly quadratic in the parameters, and that the ellipsoid of the quadratic function is defined by the Fisher information matrix \( F \). We have assumed the algorithm weights \( w_k = 1 \) for all \( k \).

Proof: Write the Taylor series of the criterion \( J_n \) about the true parameters

\[
J_n(\theta+\Delta \theta) = J_n(\theta) + \sum_{n=1}^{N_p} \frac{\partial J_n}{\partial \theta_i} \Delta \theta_i + \frac{1}{2} \sum_{n=1}^{N_p} \sum_{m=1}^{N_p} \frac{\partial^2 J_n}{\partial \theta_i \partial \theta_m} \Delta \theta_i \Delta \theta_m + h.o.t
\]

Since \( h_n(\theta) = y_n \) at the true parameters, we have

\[
\Delta J_n = \frac{1}{2} \Delta \theta^T F \Delta \theta \quad (3)
\]

with

\[
F = \sum_{n=1}^{N_p} \frac{\partial h_n(\theta)}{\partial \theta_i} \frac{\partial h_n(\theta)}{\partial \theta_j}
\]

As the number of impulse response measurements \( N \) approaches \( \infty \), the matrix becomes the parameter information matrix whose \( n \)th element is

\[
f_{\nu} \int_{-\infty}^{\infty} \frac{\partial h_n(z)}{\partial \omega_k} \frac{\partial h_n(z^{-1})}{\partial \omega_l} \frac{dz}{z}
\]

We claim that the local convergence properties of the identification procedure depend on this information matrix. This matrix depends on the structure of the identified model, e.g., where the model parameters live in state-space. The rate of convergence appears to depend on two system properties indicated by the information content of the parameterization:

1. The model's sensitivity to its parameters
2. The interconnectedness of the model parameters

As we have seen, the sensitivity \( S \) of the model to its parameters is given by the trace of \( F \). The interconnectedness of the model parameters is indicated by the condition number \( \kappa_F \) of \( F \) (6). This interconnectedness is an indication of the disparity of information available from the system model. For the deterministic algorithms, the interpretation of these model information parameters in a visual sense can be quite enlightening. The trace of \( F \) details the change in attitude of \( J_n \). The condition number \( F \) determines the contour shape of \( J_n \) near the minimum.

We have previously combined these two property measurements into a single measure, which we have called the "relative algorithm time constant" (7):

\[
\tau = \frac{x_0(f)}{b(f)}
\]

When \( \tau \) is small, the model parameter estimates should converge quite fast, while for large \( \tau \) the estimate convergence should be slow. The term relative is included because the measure \( \tau \) does not indicate the absolute convergence rate of the various model parameterizations, but the relative order of convergence.

Finally, note that the measure \( \tau \) does not rule out the idea that low sensitivity structures are good identifier models. Quite the contrary, it spells out the relationship between the parameter interconnectedness and the sensitivity of the model parameters. It is quite possible for a model with low sensitivity to have a low relative
time constant, if its parameters are relatively unconnected \( \phi(f) \). These ideas can be incorporated into areas such as circuit testing and system fault detection to improve system design and test. An analytical tool is presented which can be of aid in the design of systems with a twist towards improving the chances that a fault or construction error can be detected, while not selecting a high sensitivity structure (an undesirable selection). Potential uses here include the detection of space-structure integrity (9):

One of the key unanswered questions in parameter estimation is that of parameter sensitivity to changes in the input-output data and numerical accuracies. There has been very little work on this topic and yet the resolution of the question on which algorithm is the most effective in a numerical sense depends on the sensitivity. It would be foolish in system design work to select the design which is most sensitive to component changes or digital word length.

... Research on parameter sensitivity has been limited ... The sensitivity question will be placed in the category of future research problems for the identification of large space structures.

We have begun work on identifying model structures which possess low parameter sensitivities and good identification properties. Two structures which seem promising are the balanced state-space coordinate system and the pole/zero lattice state-space structure. Since both of these model structures are not canonical forms, some revision of the algorithms described in this paper are necessary. We have formalized the above in the following conjecture:

Let \( S_1 \) and \( S_2 \) be two different parameterizations of the same system of known order. The Gauss–Newton algorithm presented above is used to identify the corresponding parameters of both \( S_1 \) and \( S_2 \), and the optimization criterion \( J_p \) possess two local minima and \( N \) is large. Then for the same impulse response measurement data and equivalent local parameter covariance and initializations, the parameters of \( S_1 \) will converge to the true parameters before those of \( S_2 \) if and only if \( \pi_1 < \pi_2 \).

The localness of the initial parameter guesses depends on the number of impulse response measurements \( N \) used in the identification, as well as the size of the higher order terms in the equation (3).

Examples of Stochastic Algorithm Convergence

We examine a 3rd-order low-pass system. Each impulse response measurement consists of 6 samples, and 60 independent measurement sets were taken, each measurement set corrupted by additive noise to achieve a 40 dB SNR. Figure 1 shows the relative convergence of 3 different model structures. After 70 iterations, the DGHR model clearly outperforms the Dill model. Both of these structures have converged more quickly and more accurately than the PRL identifier structure. However, in the first 20 iterations or so, the Dill model structure is as good as, or better than, the DGHR structure.

We conjecture that the steep error curves of the more sensitive Dill model allow a quicker convergence to a neighborhood of the true system, but that the less sensitive DGHR structure has inherently better noise rejection. The PRL structure not only has low sensitivity, but it also has moderate parameter conditioning \( \phi(f) \), so that its performance is markedly inferior to the other two identifier structures in both the initial and latter convergence. The convergence is as conjectured.

5.4. Choice of Model Order

The arguments (and the conjecture) given above assume that the model order is exactly equal to the identified system order. What happens when the model order is too high or too low? In the former case, the model is overparameterized, but the true system parameters do belong to the model set of parameters. Thus, in this case, an exact identification up to a pole/zero cancellation is possible. In the former (and more interesting) situation, the model is underparameterized and the model parameter set does not encompass the true system.

Overparameterized Models

The Hessian matrix of an overparameterized identifier structure is singular at the convergence point (13). Thus, the approximation to the Hessian inverse computed by the algorithm will grow without bound. The information matrix \( F \) is singular, and so \( J_p \) may either fail to have any extrema, or it may have an infinite number of minima (10). However, in the algorithms presented here, a minimum point always exists, and the converged-to system should be the least-square solution. In fact, the pole/zero cancellation pairs which describe equivalent systems mean that there are an infinite number of minima. We should also note that \( J_p \) now has local minima, which of course interfere with convergence to the global minimum. In most test cases, one (or more) of the identified model structures converged to local minima. However, judicious choice of initial settings can result in convergence to a global minimum. The behavior of each identified structure in reference to local minima is uncertain at this time.

Underparameterized Model

The true system looks like a "noisy" realization to the underparameterized, reduced-order model, even when the impulse response measurements are exact. In this more interesting case of reduced-order modelling, one does not expect that the different structures will converge to equivalent systems because the model parameters can have a quite different sensitivity to measurement variations.

In running several examples, we have found that the different identifier structures do indeed converge to different systems. We computed the relative time constant \( \pi \) for these converged-to systems. We found that the fastest convergence rates (and the systems with the lowest absolute error \( J_p \)) belong to the structure with the lowest \( \pi \). We thus find that the relative time constant
still indicates important information about the convergence rates.

55. Conclusions

We have presented recursive partial realization algorithms suitable for the identification of general state-space models. The convergence behavior of these algorithms is described in terms of the identification criterion shape as well as the information content of the chosen model parameterization. This information content is given by the model parameter sensitivity and interconnectedness. This informational description of the model parameters is apparently independent of the order of the chosen model, so that useful information about relative convergence rates can be inferred even when the model order does not match that of the identified system.

References


