(2)

MANUSCRIPT OF JUNE 4, 1992

Unifying the Landmark Developments in Optimal Bounding Ellipsoid Identification

J.R. Deller, Jr., M. Nayeri, and M.S. Liu

Michigan State University
Department of Electrical Engineering
CSSP Group: Speech Processing & Adaptive Signal Processing Laboratories
East Lansing, MI 48824-1226 USA

email: deller@ee.msu.edu phone: (517) 353-8840 FAX: (517) 353-1980

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of the United States under Grant No. MIP-9016734 and by the United States Office of Naval Research under Grant No. N00014-91-J-1329. JD was also supported by an Ameritech Fellowship during the period of this work.

Abstract

A quite general class of Optimal Bounding Ellipsoid (OBE) algorithms including all methods published to date, can be unified into a single framework called the *Unified OBE (UOBE)* algorithm. UOBE is based on generalized weighted recursive least squares in which very broad classes of "forgetting factors" and data weights may be employed. Different instances of UOBE are distiguished by their weighting policies and the criteria used to determine their optimal values.

A study of existing OBE algorithms, with a particular interest in the tradeoff between algorithm performance interpretability and convergence properties, is presented. Results suggest that an intepretable, converging UOBE algorithm will be found. In this context, a new UOBE technique, the *set membership stochastic approximation (SM-SA)* algorithm is introduced. SM-SA possesses interpretable optimization measures and known conditions under which its *estimator* will converge.

1 Introduction

Set-membership-based (SM) system identification algorithms offer an interesting alternative to conventional techniques. SM methods have been receiving increasing attention internationally as is evident from the collection of papers in this volume. Recent reviews of this field are found, for example, in [1]–[3]. This paper is restricted to the class of algorithms known as *optimal bounding ellipsoid (OBE)* algorithms which follow from a *bounded error* constraint.

In this paper we initially formulate a very broad class of OBE algorithms, including all methods published to date, into a general framework called the *Unified OBE (UOBE)* algorithm. We then exploit the UOBE formalism to explore some interesting connections which exist among existing OBE algorithms. A particular concern will be the pursuit of an OBE algorithm which has both well-understood convergence properties, and an intuitively meaningful optimization criterion. These two desirable properties have yet to be combined into a single OBE algorithm.

2 A Unified OBE (UOBE) Algorithm

2.1 The Bounded Error Problem and the UOBE

The bounded error identification problem is as follows: Assume that we are observing some physical system which is generating sequence $\{\boldsymbol{y}(\cdot)\} \in \mathcal{C}^k$ in response to input $\{\boldsymbol{u}(\cdot)\} \in \mathcal{C}^l$. $\{\boldsymbol{u}(\cdot)\}$ is a realization of an ergodic, wide sense stationary stochastic process. Both input and output sequences are measurable. We assume the existence of a "true" regression model of the form

$$\boldsymbol{y}(n) = \boldsymbol{\Theta}_*^H \boldsymbol{x}(n) + \boldsymbol{\varepsilon}_*(n) \tag{1}$$

in which $\boldsymbol{x}(n)$ is an m-vector of known functions,

$$\boldsymbol{x}(n) = \begin{bmatrix} \varphi_1[\boldsymbol{y}(n-1), \boldsymbol{y}(n-2), \ldots, \boldsymbol{y}(n-p), \boldsymbol{u}(n), \boldsymbol{u}(n-1), \ldots, \boldsymbol{u}(n-q)] \\ \varphi_2[\boldsymbol{y}(n-1), \boldsymbol{y}(n-2), \ldots, \boldsymbol{y}(n-p), \boldsymbol{u}(n), \boldsymbol{u}(n-1), \ldots, \boldsymbol{u}(n-q)] \\ \vdots \\ \varphi_m[\boldsymbol{y}(n-1), \boldsymbol{y}(n-2), \ldots, \boldsymbol{y}(n-p), \boldsymbol{u}(n), \boldsymbol{u}(n-1), \ldots, \boldsymbol{u}(n-q)] \end{bmatrix}, \tag{2}$$

and where $\{\boldsymbol{\varepsilon}_*(\cdot)\} \in \mathcal{C}^k$ is a realization of a zero-mean, second moment ergodic, complex vector-valued random sequence whose vector components are independent. The matrix $\boldsymbol{\Theta}_* \in \mathcal{C}^{m \times k}$ parameterizes the model. At time n we wish to use the observed data on $t \in [1, n]$ to deduce an estimated model of the same form. The parameter estimate is denoted by $\boldsymbol{\Theta}(n)$ and the residual process by $\epsilon(\cdot, \Theta(n))$. The dependence of the residual upon the parameter estimates is highly significant, so it is shown explicitly.

As the basis for the unified algorithm, we recall the identification algorithm variously known as *weighted recursive least squares (WRLS)* (e.g. [6],[7]), *weighted sequential least squares* (e.g. [8]), *weighted sequential regression* (e.g. [9]), and other names. We shall use the name "WRLS" throughout. The WRLS algorithm is used to sequentially compute the weighted least square error estimate,

$$\Theta(n) = \arg\min_{\Gamma} \frac{1}{n} \sum_{\tau=1}^{n} w_{n,\tau} \parallel \epsilon(\tau, \Gamma) \parallel^2, \qquad \Gamma \in C^{m \times k} \tag{3}$$

where, in the most general case, the data weights $w_{n,\tau}$ may be time-varying (dependent upon both n and τ) in a simple way,

$$w_{n,\tau} = \begin{cases} \alpha_n w_{n-1,\tau} & \tau \leq n-1 \\ \beta_n & \tau = n \end{cases} \tag{4}$$

where the sequences $\{\alpha_n\}$ and $\{\beta_n\}$ will be specified below (for the present, they should just be regarded as sequences of finite numbers). In these terms, the WRLS relations are[1]

$$\begin{aligned} C_\alpha^{-1}(n) &= C^{-1}(n)/\alpha_n & (5) \\ C^{-1}(n) &= C_\alpha^{-1}(n-1) - \beta_n \frac{C_\alpha^{-1}(n-1)x(n)x^H(n)C_\alpha^{-1}(n-1)}{1+(\beta_n/\alpha_n)G(n)} & (6) \\ \Theta(n) &= \Theta(n-1) + \beta_n C^{-1}(n)x(n)\epsilon^H(n, \Theta(n-1)) & (7) \end{aligned}$$

with $C(0) = 0$ and where $G(n) \stackrel{\text{def}}{=} x^H(n)C^{-1}(n-1)x(n)$. From (4) we note that the number α_n effectively *scales* all previous weights at time n to (in conventional applications of WRLS) decrease the influence of the corresponding data on the estimates. Accordingly, $\{\alpha_n\}$ is often called a sequence of *forgetting factors*, and in many cases the sequence is taken to be a constant which is smaller than, but close to, unity. Either $\{\alpha_n\}$ or $\{\beta_n\}$, or both, may be omitted (set to unity), but we will have use for both sequences in this work. The matrix $C(n)$, usually called the *covariance* matrix[2], is by definition the sum of the weighted outer products,

$$C(n) = \sum_{\tau=1}^{n} w_{n,\tau} x(\tau) x^H(\tau) = \alpha_n C(n-1) + \beta_n x(n) x^H(n). \tag{8}$$

The recursions above theoretically provide an estimate $\Theta(n)$ which is equivalent to the solution of

[1]For completeness, we also note that the WRLS algorithm can be implemented in a different form using QR-decomposition (e.g. [9]), and the QR form has been employed with some advantages in some of the SM-based algorithms to be discussed below (e.g. [10] - [13]). Because our purpose here is to relate many existing developments, we shall focus on the more conventional approach represented by (6) and (7).

[2]Though it is more properly a *normal* matrix [6].

the *normal equations* (e.g. [6]),

$$C(n)\Theta(n) = C_{xy}(n) \tag{9}$$

with

$$C_{xy}(n) \stackrel{\text{def}}{=} \sum_{\tau=1}^{n} w_{n,\tau} x(\tau) y^H(\tau). \tag{10}$$

It is not widely appreciated that all reported OBE algorithms can be unified into a general framework which we shall call the *Unified OBE (UOBE)* algorithm. Particular algorithms are distinguished by specifying the optimization strategy for determining the sequences of weights $\{\alpha_n\}$ and $\{\beta_n\}$.

Let us initially present the UOBE framework, then enumerate the particular algorithms. UOBE algorithms arise from a bounded error constraint:

$$\| \epsilon_*(n) \|^2 < \gamma_n, \tag{11}$$

where $\{\gamma_n\}$ is a known positive sequence. At time n, a *set* of system parameters, say $\Omega(n)$, can be found which are consistent with the observations and this sequence of bounds. The exact set is difficult to describe and track, but, in conjunction with WRLS processing, $\Omega(n)$ can be shown to be contained in a superset of the form (e.g. [3] - [5])

$$\bar{\Omega}(n) = \left\{ \Theta \mid \ \mathrm{tr}\{[\Theta - \Theta(n)]^H \frac{C(n)}{\kappa(n)} [\Theta - \Theta(n)]\} < 1 \right\} \tag{12}$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace of a matrix, $\Theta(n)$ is the WRLS parameter estimate at time n using weights $\{w_{n,\tau},\ \tau \in [1,n]\}$, $C(n)$ is the weighted covariance matrix, and $\kappa(n)$ is the scalar quantity

$$\kappa(n) \stackrel{\text{def}}{=} \mathrm{tr}\{\Theta^H(n) C(n) \Theta(n)\} + \sum_{\tau=1}^{n} w_{n,\tau} \left[\gamma_\tau - \| y(\tau) \|^2\right] . \tag{13}$$

$\bar{\Omega}(n)$ is a hyperellipsoid in $\mathcal{R}^{2mk}$, with its center at $\Theta(n)$. By examining a single output – say $y_i(\cdot)$, the i^{th} component of $y(\cdot)$ – we see that a common "ellipsoid matrix" $C(n)/\kappa(n)$ is shared by each of the individual outputs, but that each is centered on a different parameter estimate represented by column i of $\Theta(\cdot)$. We conclude therefore that under bounded error constraints, a hyperellipsoid can be associated with a WRLS recursion and conversely.

The weights $\{w_{n,\tau},\ \tau \in [1,n]\}$ directly control the size, orientation, and location of the ellipsoid in the parameter space at time n. However, because of the structure of the WRLS recursions, in moving from time $n-1$ to n, we are not free to alter the set of weights beyond that which can be accomplished using the numbers α_n and β_n. This is evident in the recursions (6) and (7). At most,

Editors: This figure placed here for reviewers' convenience. Also included on a separate sheet per instructions.

At time n,

1. In conjunction with the incoming data set $(\boldsymbol{y}(n), \boldsymbol{x}(n))$, find optimal values of α_n and/or β_n, say α_n^* and/or β_n^*. Optimality criteria are described in the text;
2. If optimal positive (and sometimes further constrained) values α_n^* and/or β_n^* do not exist, then discard the data set (set $\beta_n^* = 0$ and / or $\alpha_n^* = 1$);
3. Update $\boldsymbol{C}(n)$, $\boldsymbol{\Theta}(n)$, and $\kappa(n)$ using (6), (7), and a recursion for $\kappa(\cdot)$ described in Lemma 1.

Figure 1: General steps of the UOBE algorithm.

therefore, we have two free parameters with which to control $\bar{\Omega}(n)$. All existing UOBE algorithms differ only in their sequences $\{\alpha_n\}$ and $\{\beta_n\}$, and the optimization criterion used to determine them. The central objective of the general UOBE algorithm is to employ the weights α_n and/or β_n in the context of WRLS estimation to sequentially minimize the ellipsoid size in some sense. A significant benefit is that often no weights exist which can minimize the ellipsoid, indicating that the incoming data set is uninformative in the SM sense and need not be processed.

All UOBE algorithms adhere to the three general steps displayed in Fig. 1. (For details of initialization and other nuances of the specific algorithms, the reader is referred to original papers cited above and below.) Having established the basic framework of UOBE, we next consider the optimization process.

2.2 Optimization

Within the UOBE framework in Fig. 1, the different algorithms are distinguished by their sequences $\{\alpha_n\}$ and $\{\beta_n\}$, in conjunction with the optimization criterion employed in selecting them. Three optimization criteria have been used. The first two involve set measures on the ellipsoid $\bar{\Omega}(n)$ and are clearly interpretable with minimal explanation, while the third will require some further elaboration. The criteria are:

Optimization Criterion 1 Minimize the the determinant of the inverse ellipsoid matrix,

$$\mu_v\{\bar{\Omega}(n)\} \stackrel{\text{def}}{=} \det\ \{\kappa(n)\boldsymbol{C}^{-1}(n)\} \tag{14}$$

(henceforth $\mu_v(n)$ for simplicity);

Optimization Criterion 2 Minimize the trace of the inverse ellipsoid matrix,

$$\mu_t\{\bar{\Omega}(n)\} \stackrel{\text{def}}{=} \text{tr}\ \{\kappa(n)C^{-1}(n)\} \tag{15}$$

(henceforth $\mu_t(n)$.); and,

Optimization Criterion 3 Minimize the parameter $\kappa(n)$.

Criteria 1 and 2 were first suggested by Fogel and Huang [4], while criterion 3 was used by Dasgupta and Huang [14]. In these original papers, single output (real, scalar sequence $y(\cdot)$) systems were considered. In the single output case in which $\bar{\Omega}(n)$ is clearly intepretable as an hyperellipsoid ellipsoid in $\mathcal{R}^m$, $\mu_v(n)$ is proportional to the square of the volume of the ellipsoid, while $\mu_t(n)$ is proportional to the sum of squares of its semi-axes. The same two measures are meaningful in the *multiple* output case, since they result in the minimization of the volume or trace of the *common* ellipsoid shared by all the outputs (see discussion below (13)). Of course, an important feature of any optimization criterion is that it be readily intepretable as a desirable objective. The volume and trace criteria apparently have this property. Accordingly, UOBE algorithms following these criteria will often be referred to as *interpretable* algorithms in the following. Criterion 3, however, has been the subject of some controversy with regard to its meaningfulness and intepretability, as we discuss later in the paper. This criterion has been used in conjunction with a specific weighting strategy to achieve a rigorous proof of convergence in a certain sense. The apparent need to trade interpretability of a UOBE algorithm for proof of convergence will be one of the central themes of the remaining parts of the paper.

Having established the optimization criteria, let us now focus on the weight sequences. For any of the criteria above, there is only one quantity to be optimized at time n which in each case is dependent upon both α_n and β_n. However, the numbers α_n and β_n are essentially independent of one another, so that any attempt to optimize one of the criterion measures with respect to *both* α_n and β_n results in an infinity of solutions which is resolved by arbitrarily choosing a value of either weight. Accordingly, we may either tie the weights together through some functional relation, optimize over only one weight and choose the other according to some predetermined purpose, or simply eliminate the "unused" weight altogether by setting it to unity (a special case of the second strategy). We shall adopt the policy of writing the weights α_n and β_n as functions of a single parameter to be optimized at time n, say λ_n, so that (in conventionally abusive notation)

$$\alpha_n = \alpha_n(\lambda_n) \tag{16}$$

$$\beta_n = \beta_n(\lambda_n) \tag{17}$$

where $\{\alpha_n\}$ and $\{\beta_n\}$ should now be considered to be sequences of *functions* whose properties will be specified later. So that we have sufficient generality for our purposes, it is important to note that these functions need *not* depend on λ_n. For example, $\{\alpha_n\}$ may be chosen *independently* of the optimization in which case, at time n,

$$\alpha_n(\lambda_n) = \alpha_n, \quad \text{a constant (function) independent of } \lambda_n. \tag{18}$$

In this special case, it is true that

$$\frac{\partial}{\partial \lambda_n}\alpha_n(\lambda_n) = 0. \tag{19}$$

We shall refer to λ_n as a "weight" at time n, since if $\alpha_n(\lambda_n) = 1$ and $\beta_n(\lambda_n) = \lambda_n$, then λ_n is simply the weight associated with the standard WRLS recursion with no forgetting factor. This general setup will allow us to embrace all UOBE algorithms in a single theoretical framework.

We now turn to the problem of optimization of the identification at time n according to the criteria stated above. The following results generalize and unify all optimization procedures found in the literature.

Theorem 1 *Each of the functions in the sequences $\{\alpha_n(\lambda_n)\}$ and $\{\beta_n(\lambda_n)\}$ are assumed positive and are chosen such that, for each n, $q_n(\lambda_n) = \beta_n(\lambda_n)/\alpha_n(\lambda_n)$ is a continuous, one-to-one mapping*

$$q_n : (0, a_n) \longrightarrow (0, \infty) \tag{20}$$

where $a_n > 0$. Then, if it exists, a weight λ_n^* which minimizes

1. *the volume measure $\mu_v(n)$ is the unique positive root in λ_n of the equation $F_v(q_n(\lambda_n)) = 0$ on the interval $(0, a_n)$, where,*

$$F_v(s) = a_2 s^2 + a_1 s + a_0 \tag{21}$$

with
$$\begin{aligned} a_2 &= \{(mk-1)\gamma_n G^2(n)\}, \\ a_1 &= \{(2mk-1)\gamma_n + \| \epsilon(n, \Theta(n-1)) \|^2 - \kappa(n-1)G(n)\} G(n), \\ a_0 &= mk\left[\gamma_n - \| \epsilon(n, \Theta(n-1)) \|^2\right] - \kappa(n-1)G(n); \end{aligned}$$

2. *the trace measure $\mu_t(n)$ is the unique positive root in λ_n of the equation $F_t(q_n(\lambda_n)) = 0$ on the interval $(0, a_n)$, where*

$$F_t(s) = b_3 s^3 + b_2 s^2 + b_1 s + b_0 \tag{22}$$

with
$$\begin{aligned} b_3 &= \gamma(n)G^2(n)(G(n) - I(n-1)H(n)), \\ b_2 &= 3\gamma(n)G(n)[G(n) - I(n-1)H(n)], \\ b_1 &= H(n)G(n)I(n-1)\kappa(n-1) - 2H(n)I(n-1)\left[\gamma(n) - \| \epsilon(n, \Theta(n-1)) \|^2\right] \\ &\quad - G(n) \| \epsilon(n, \Theta(n-1)) \|^2 + 3\gamma(n)G(n), \\ b_0 &= \gamma(n) - \| \epsilon(n, \Theta(n-1)) \|^2 - H(n)I(n-1)\kappa(n-1), \end{aligned}$$

where $H(n) \stackrel{\text{def}}{=} \boldsymbol{x}^T(n)\boldsymbol{C}^{-2}(n-2)\boldsymbol{x}(n)$, *and* $I(n) \stackrel{\text{def}}{=} tr\ \boldsymbol{C}(n)$.

Before sketching the proof of this important result, we remark that it is unnecessary to actually *solve* for the root of $F_v(q_n(\lambda_n)) = 0$ or $F_t(q_n(\lambda_n)) = 0$ to determine whether a positive root exists. Based on generalizations of previous work [3] - [5], it can be shown that necessary and sufficient tests for the *existence* of the positive root are $a_0 < 0$ and $b_0 < 0$ in the volume and trace minimization cases, respectively. This fact will play an important role in future discussions.

Sketch of Proof: Write the (optimal) normal equations (9) at time n in the form

$$\left[\alpha_n(\lambda_n^*)\boldsymbol{C}(n-1) + \beta_n(\lambda_n^*)\boldsymbol{x}(n)\boldsymbol{x}^H(n)\right]\boldsymbol{\Theta}(n) = \alpha_n(\lambda_n^*)\boldsymbol{C}_{\boldsymbol{xy}}(n-1) + \beta_n(\lambda_n^*)\boldsymbol{x}(n)\boldsymbol{y}^H(n). \tag{23}$$

Now divide through by $\alpha_n(\lambda_n^*)$ to yield

$$\left[\boldsymbol{C}(n-1) + q_n(\lambda_n^*)\boldsymbol{x}(n)\boldsymbol{x}^H(n)\right]\boldsymbol{\Theta}(n) = \boldsymbol{C}_{\boldsymbol{xy}}(n-1) + q_n(\lambda_n^*)\boldsymbol{x}(n)\boldsymbol{y}^H(n) \tag{24}$$

where $q_n(\cdot)$ is defined in the theorem. This shows that an identical estimate (and concomitant optimization problem) results if the covariance matrix is unweighted and the outer product is weighted by $q_n(\lambda_n^*)$. In principle, then, to obtain the desired estimate, $\boldsymbol{\Theta}(n)$, the dependence of $q_n(\lambda_n^*)$ upon λ_n^* is superfluous and we can optimize over, say, $p_n \stackrel{\text{def}}{=} q_n(\lambda_n)$, ignoring the dependence upon λ_n. For this simple case, it has been proven in [3],[5] that

$$\frac{\partial \mu_v}{\partial p_n} = K(p_n)F_v(p_n) \tag{25}$$

where $K(p_n) > 0$ for all $p_n > 0$, and where $F_v(p_n)$ is as defined in (21). Moreover, $F_v(p_n)$ has at most one positive root, say p_n^*, which, if it exists, corresponds to a minimum of $\mu_v(n)$ since

$$\left.\frac{\partial^2 \mu_v}{\partial p_n^2}\right|_{p_n^*} = K(p_n^*)\left.\frac{\partial F_v}{\partial p_n}\right|_{p_n^*} + \left.\frac{\partial K}{\partial p_n}\right|_{p_n^*} F_v(p_n^*) = K(p_n^*)\left.\frac{\partial F_v}{\partial p_n}\right|_{p_n^*} > 0. \tag{26}$$

Now return to the case in which $\{p_n\}$ represents a sequence functions of parameter λ_n, $\{q_n(\lambda_n)\}$. It follows immediately from (25) that, at time n,

$$\frac{\partial \mu_v}{\partial \lambda_n} = K(q_n(\lambda_n))F_v(q_n(\lambda_n))\frac{\partial q_n}{\partial \lambda_n}. \tag{27}$$

Because of the assumed monotonicity of $q_n(\lambda_n)$, the only zeros of the derivative occur when $F_v(q_n(\lambda_n)) = 0$. Since there is a unique root in p_n, *viz.* p_n^*, and since $q_n(\cdot)$ is an invertible function, there is a unique root in λ_n, *viz.* $\lambda_n^* = q_n^{\text{inv}}(p_n^*)$, where $q_n^{\text{inv}}(\cdot)$ is the inverse mapping of

$q_n(\cdot)$.

Similar analysis for the trace case follows from the work in [3] in which $\alpha_n(\lambda_n)$ is taken to be independent of λ_n. □

Theorem 1 does not embrace optimization by minimization of $\kappa(n)$. We shall find this criterion to be problematic from several points of view. In the present situation, the fact that $\kappa(n)$ cannot be expressed as a function of $q_n(\lambda_n)$ alone precludes the derivation of a general result like those in Theorem 1. However, we provide a result which will be useful in future discussions:

Theorem 2 *Consider the optimization problem posed above. If it exists, the optimal weight λ_n^* which minimizes $\kappa(n)$ is a root of the equation*

$$F_\kappa(s) = \left\{(\alpha_n(s) + \beta_n(s)G(n))^2\kappa(n-1) - \beta_n^2(s)G(n) \parallel \varepsilon(n, \Theta(n-1)) \parallel^2\right\}\alpha_n'(s) + \left\{(\alpha_n(s) + \beta_n(s)G(n))^2\gamma_n - \alpha_n^2(s) \parallel \varepsilon(n, \Theta(n-1)) \parallel^2\right\}\beta_n'(s). \quad (28)$$

where α_n' and β_n' indicate derivatives.

This inelegant result will simplify to useful quadratics in two special cases in the paper. It is proved by taking the total derivative with respect to λ_n of the recursive expression for $\kappa(n)$ found in the following lemma. This lemma has been proven for the case $\alpha_n(\lambda_n) = 1$, $\beta_n(\lambda_n) = \lambda_n$ in [5] and for $\alpha_n(\lambda_n) = \alpha_n$ (independent of λ_n), $\beta_n(\lambda_n) = \lambda_n$ in [3]. The generalization given here follows from similar analysis.

Lemma 1 *The sequence $\kappa(n)$ can be computed recursively using*

$$\kappa(n) = \alpha_n(\lambda_n^*)\kappa(n-1) + \beta_n(\lambda_n^*)\gamma_n - \frac{\alpha_n(\lambda_n^*)\beta_n(\lambda_n^*) \parallel \varepsilon(n, \Theta(n-1)) \parallel^2}{\alpha_n(\lambda_n^*) + \beta_n(\lambda_n^*)G(n)} \quad (29)$$

with $\alpha_0(\lambda_0^*)\kappa(0) \stackrel{\text{def}}{=} 0$.

In conjunction with these general optimization results, we present the following corollary which asserts some remarkable facts about the quantities upon which the various UOBE algorithms are based. Once again, we see $\kappa(n)$ to have exceptional behavior relative to the more interpretable criteria:

Corollary 1 *Consider a UOBE algorithm in which volume or trace is to be minimized. Let $q_n(\lambda_n)$ be as described in Theorem 1 for each n. Then, the following are independent of the choices of function sequences $\{\alpha_n(\lambda_n)\}$ and $\{\beta_n(\lambda_n)\}$:*

1. *the sequence of measures of optimality ($\{\mu_v(n)\}$ or $\{\mu_t(n)\}$);*

2. the data points selected (times for which there exists $\lambda_n^* > 0$);

3. the parameter matrix estimate, $\boldsymbol{\Theta}(n)$.

However, in a UOBE algorithm with κ minimization, none of these items is independent of the sequences $\{\alpha_n(\lambda_n)\}$ and $\{\beta_n(\lambda_n)\}$.

Sketch of Proof: Consider first the volume and trace cases. The independence of $\boldsymbol{\Theta}(n)$ follows from the fact that there is a unique root (if any), $p_n^* = q_n(\lambda_n^*)$, of either (21) or (22), which does not depend on functions α_n or β_n. Therefore, $\boldsymbol{\Theta}(n)$ is given by (24) regardless of the choices of α_n and β_n. However, consider $\boldsymbol{C}(n)$ for a particular choice of $\alpha_n(\lambda_n)$ as written in the brackets on the left side of (23). It follows that

$$\frac{\boldsymbol{C}(n)}{\alpha_n(\lambda_n^*)} = \boldsymbol{C}(n-1) + q_n(\lambda_n^*)\boldsymbol{x}(n)\boldsymbol{x}^H(n). \tag{30}$$

Since the number $q_n(\lambda_n^*)$ does not depend on the choice of α_n, the right side is invariant with α_n. $\boldsymbol{C}(n)$ must vary with choice of α_n on the left to maintain the equality. A similar analysis pertains to the ratio $\boldsymbol{C}(n)/\beta_n(\lambda_n^*)$. Also, from (29)

$$\frac{\kappa(n)}{\alpha_n(\lambda_n^*)} = \kappa(n-1) + q_n(\lambda_n^*)\gamma_n - q_n(\lambda_n^*)\frac{\| \boldsymbol{\varepsilon}(n, \boldsymbol{\Theta}(n-1)) \|^2}{1 + q_n(\lambda_n^*)G(n)} \tag{31}$$

and a similar argument applies to show that $\kappa(n)$ depends on α_n. The ratio $\kappa(n)/\beta_n(\lambda_n^*)$ is formed to show dependence of $\kappa(n)$ upon β_n. On the other hand, consider the ellipsoid matrix $\boldsymbol{C}(n)/\kappa(n)$. Dividing numerator and denominator by $\alpha_n(\lambda_n^*)$ yields

$$\frac{\boldsymbol{C}(n)}{\kappa(n)} = \frac{\boldsymbol{C}(n)/\alpha_n(\lambda_n^*)}{\kappa(n)/\alpha_n(\lambda_n^*)} = \frac{\boldsymbol{C}(n-1) + q_n(\lambda_n^*)\boldsymbol{x}(n)\boldsymbol{x}^H(n)}{\kappa(n-1) + q_n(\lambda_n^*)\gamma_n - q_n(\lambda_n^*)\frac{\|\boldsymbol{\varepsilon}(n,\boldsymbol{\Theta}(n-1))\|^2}{1+q_n(\lambda_n^*)G(n)}} \tag{32}$$

which reveals that $\boldsymbol{C}(n)/\kappa(n)$, and hence $\mu_v(n)$ and $\mu_t(n)$, depend only on $p_n^* = q_n(\lambda_n^*)$ and not on particular choices of α_n and β_n. By similar means, it can be argued that the quantities $G(n)\kappa(n-1)$ and $H(n)I(n)\kappa(n-1)$, and, hence, a_0 and b_0 of Theorem 1, are independent of α_n and β_n. By the remarks under Theorem 1, it is therefore seen that the selection of points does not depend on $\{\alpha_n(\lambda_n)\}$ nor $\{\beta_n(\lambda_n)\}$.

Now consider the κ minimization policy. We provide a counterexample to the claim that the minimum value of $\kappa(n)$, the estimate $\boldsymbol{\Theta}(n)$, and the selection or rejection of data, are all independent of the choice of function sequences. At time n, for a given $\boldsymbol{C}(n-1)$, $\kappa(n-1)$, and $\boldsymbol{\varepsilon}(n, \boldsymbol{\Theta}(n-1))$, suppose that $\| \boldsymbol{\varepsilon}(n, \boldsymbol{\Theta}(n-1)) \|^2 > \gamma_n - \kappa(n-1)$, but $\| \boldsymbol{\varepsilon}(n, \boldsymbol{\Theta}(n-1)) \|^2 \leq \gamma_n$. We shall show later in the paper that if $\alpha_n(\lambda_n) = 1$ and $\beta_n(\lambda_n) = \lambda_n$, then $\lambda_n^* > 0$ does not exist; whereas if

$\alpha_n(\lambda_n) = 1 - \lambda_n$ and $\beta_n(\lambda_n) = \lambda_n$, then $0 < \lambda_n^* < 1$ may exist[3]. Therefore, under the first choice of functions, the point will certainly be rejected; whereas with the second, it may be accepted. The resulting estimate $\boldsymbol{\Theta}(n)$, and value $\kappa(n)$, can therefore be different under the different choices of α_n and β_n. □

We now turn to the consideration of specific algorithms which have been used in practice. In addition to showing that these methods are quickly unified under the UOBE framework, one of the main themes will be to explore the apparent tradeoff between interpretability and convergence which seems to exist in the currently employed methods. The UOBE paradigm will contribute the understanding of this relationship.

3 The "Landmark" OBE Algorithms

It is the purpose of this section to enumerate instances of the UOBE algorithm which have been used in practice. These algorithms have each arisen for a different reason and the unification of existing methods has not been appreciated nor explored because their original developments seem somewhat disparate. However, in light of the UOBE framework, it is natural to inquire to what extent the various algorithms are truly serving distinctly different purposes. This inquiry is the subject of the next section of the paper.

We shall make no attempt to *formally* reconstruct original developments. Rather, in this section we simply distinguish the methods by specification of their sequences $\{\alpha_n(\lambda_n)\}$, $\{\beta_n(\lambda_n)\}$, and optimization criteria. The reader is referred to the original papers for a clearer understanding of the history and motivations for the different algorithms.

Three principle OBE algorithms have been studied extensively. These are the *Fogel-Huang OBE (F-H/OBE)* algorithm [4] (originally called simply "OBE"), the *set-membership weighted recursive least squares (SM-WRLS)* algorithm [10],[5], and the *Dasgupta-Huang OBE (D-H/OBE)* algorithm [14]. Their differences in terms of the UOBE framework are shown in Table 1. To these three basic versions, we have added a fourth algorithm which has been developed recently by the authors, the *set-membership stochastic approximation (SM-SA)* algorithm [15]. In the ensuing discussion, SM-SA will be found to be related to its predecessors in some interesting ways. We have also noted a heretofore unpublished variation on SM-WRLS, *Dual SM-WRLS*, which will be found to exhibit some useful numerical properties.

We should also remark that our focus in this paper is principally upon the identification of time-invariant systems in which the components of the disturbance vectors $\boldsymbol{\epsilon}_*(n)$ are independent

[3] Because of the weighting strategy, λ_n^* must be constrained in this case to the interval $(0, 1)$.

Editors: This table placed here for reviewers' convenience. Also included on a separate sheet per instructions.

Table 1: Specification of Existing UOBE Algorithms

Algorithm	$\alpha_n(\lambda_n^*)$	$\beta_n(\lambda_n^*)$	Optimization
F-H/OBE	$1/\kappa(n-1)$	λ_n^*/γ_n	$\mu_v(n)$ or $\mu_t(n)$
SM-WRLS	1	λ_n^*	$\mu_v(n)$ or $\mu_t(n)$
Dual SM-WRLS	λ_n^*	1	$\mu_v(n)$ or $\mu_t(n)$
D-H/OBE	$1-\lambda_n^*$	λ_n^*	$\kappa(n)$
SM-SA	$\Lambda_{n-1}^*/(\Lambda_{n-1}^*+\lambda_n^*)$ $\Lambda_n^* \stackrel{\text{def}}{=} \sum_{\tau=1}^{n} \lambda_\tau^*$	$\lambda_n^*/(\Lambda_{n-1}^*+\lambda_n^*)$	$\mu_v(n)$ or $\mu_t(n)$

and each orthogonal sequences. Explicitly adaptive UOBE algorithms have been discussed in [16] - [19]. A discussion of colored noise issues for a D-H/OBE–like algorithm is found in [20], and for a more general class of algorithms in [21],[22].

4 Discussion and Comparative Analysis of Existing UOBE Algorithms

4.1 Volume and Trace Minimizing Algorithms

F-H/OBE and SM-WRLS. F-H/OBE represents the first major journal paper on the application of ellipsoid algorithms to parametric LP models. The entirely nonintuitive sequences $\{\alpha_n(\lambda_n)\}$ and $\{\beta_n(\lambda_n)\}$ used in F-H/OBE are the consequence of the algorithmic approach taken rather than deliberate choices of the functions (see [4]). In fact, F-H/OBE was developed using a geometric approach which attempts to optimally bound with a new hyperellipsoid, $\bar{\Omega}(n)$, the intersection of the existing ellipsoid, $\bar{\Omega}(n-1)$, and the feasible set implied by the incoming data set. It is interesting to note that, because the weighting sequence $\{\alpha_n(\lambda_n)\}$ is equivalent to the sequence $\{\kappa^{-1}(n)\}$ in F-H/OBE, the ellipsoid matrix at time n, $C(n)/\kappa(n)$, is identical to the scaled covariance matrix $C_\alpha(n) = \alpha_n C(n)$ whose inverse is computed directly in the course of the recursion (6).

As an aside, we note that the volume optimization version of F-H/OBE is "suboptimal" in the sense that it may sometimes result in ellipsoids which are optimal in the prescribed sense, but are are unnecessarily large according to certain simple arguments. Belforte and Bona have suggested a remedy in [23] (see also [24]). As pointed out by Walter and Piet-Lahanier [1], the modified procedure is equivalent to the *ellipsoid with parallel cuts* algorithm developed by researchers working in linear programming.

Even though Fogel and Huang clearly state in their 1982 paper that there is a LSE problem underlying F-H/OBE, the geometric approach tends to draw attention away from its presence. The approach, notwithstanding, however, the similarity of the F-H/OBE equations to "nonadaptive" WRLS (i.e., without the $\{\alpha_n\}$ sequence) is striking, and it has not gone unnoticed in the literature [1],[10],[16]. The paper by Norton and Mo [16] which treats *adaptive* OBE processing uses the WRLS framework and implicitly suggests the basis for the UOBE approach taken here. The key to recognizing the potential for an unlimited variety of UOBE algorithms under the WRLS umbrella, is the recognition of Fogel and Huang's $\kappa^{-1}(n)$ parameter as an unusual "forgetting factor" α_n. Until recently, however, this uniformity of ellipsoid algorithms was not fully appreciated. In the early and mid 1980's, Deller and students (early papers cited, e.g., in [5]) recognized the similarity of F-H/OBE to RLS, and attempted to associate an ellipsoid directly with WRLS rather than conversely. The result is SM-WRLS, which is so-named to emphasize the nature of the approach.

While developed very differently, it can be appreciated that F-H/OBE and SM-WRLS are very similar, the most significant difference being the choice of the "unoptimized" sequence $\{\alpha_n(\lambda_n)\}$. Indeed, since we know that the parameter estimates and minimization measures will be identical in the two cases, there is little in the way of theoretical consideration to commend one over the other. No practical considerations are known which indicate a preference. It is true that modifying the $\{\alpha_n(\lambda_n)\}$ and $\{\beta_n(\lambda_n)\}$ sequences used in F-H/OBE will destroy the original geometrical interpretation of the algorithm. However, it is not clear that this interpretation has any *practical* significance. With this disclaimer, we note that the extensions of SM-WRLS discussed below apply in similar ways to F-H/OBE and other algorithms in this class.

The Dual SM-WRLS algorithm [22] has arisen out of an important practical consideration. It is apparent from (8) that, if $\alpha_n = \alpha_n(\lambda_n)$ tends to be not less than unity, $\{\beta_n = \beta_n(\lambda_n)\}$ must be a generally increasing sequence for incoming data to have any impact on the estimate, particularly as n becomes large. For SM-WRLS, this fact frequently leads to huge numbers in the computations and the potential for numerical instabilities. This problem does not occur when the sequence $\{\alpha_n = \alpha_n(\lambda_n)\}$ is optimized. In fact, in unpublished simulation studies (with the volume criterion) we have found the weight sequence to remain nicely bounded. Interestingly, but not unexpectedly (Corollary 1), if SM-WRLS is run on the same data, first with the β_n weights optimized, then with the α_n weights optimized, identical data are selected in each case by the set-membership considerations, and identical estimates result from the two approaches.

F-H/OBE and SM-WRLS have been successfully applied to the identification of simulated and real systems (e.g. [3],[11],[4],[5]). The recent discovery of the "dual" optimization concept promises to add a new meritorious feature to these algorithms, improved numerical stability. The central benefit of the general classes of algorithms represented by these methods is the meaningfulness of the optimization process. The main deficiency of these groups of algorithms has been their lack of well-understood convergence properties. This problem led to the development of the D-H/OBE algorithm to which we turn below. Before doing so, however, we address the issue of whether volume and trace algorithms converge. An affirmative answer to this question is most desirable because it would combine the desired features of interpretability and convergence.

Convergence of Volume and Trace Algorithms. While our immediate discussion is focusing on existing UOBE algorithms, the work in earlier sections of this paper renders the following applicable to virtually any algorithm which minimizes volume or trace. We restrict our attention to the case in which the components of the disturbance vectors $\boldsymbol{\varepsilon}_*(n)$ are independent and each orthogonal sequences. A discussion of colored noise issues is found in [21],[22].

One of the alluring aspects of having interpreted the general UOBE algorithm as a WRLS algorithm with a bounded error "overlay" is that the convergence properties of the estimate resulting from the basic RLS algorithm ($\alpha_n(\lambda_n) = \beta_n(\lambda_n) = 1$ for all n) are well-known. In the RLS case, if the sequence $\{\boldsymbol{\varepsilon}_*(\cdot)\}$ is wide-sense stationary, second moment ergodic almost surely (a.s.), white noise, then the RLS estimator $\boldsymbol{\Theta}(\cdot)$ will converge asymptotically to $\boldsymbol{\Theta}_*$ a.s. (e.g. [6],[8]). However, this well-known convergence result falls far short of a convergence proof for the UOBE algorithms under consideration which use vastly different data weighting strategies. A simple inclusion of the sequence $\{\alpha_n(\lambda_n) = \alpha\}$ with $0 < \alpha < 1$ (with $\{\beta_n(\lambda_n) = 1\}$), for example, has been shown to lead to inconsistent asymptotic estimates [25]. Likewise, we may even assert a.s. convergence of the RLS estimate, albeit to a bias, when $\{\boldsymbol{\varepsilon}_*(\cdot)\}$ is colored and persistently exciting [26]. Again, while this result does not provide proof of estimator convergence for the present UOBE cases, the UOBE estimate has been found to practically converge, expectedly to a bias [21].

More generally, it would be interesting to have a precise understanding of the asymptotic behavior of the hyperellipsoidal *feasible set*, especially in the case of colored noise. Unfortunately, convergence proofs for the volume and trace minimization algorithms are not known. The original OBE paper by Fogel and Huang [4] is sometimes misunderstood to indicate the convergence of the bounding ellipsoid to a point under ordinary conditions on $\{\boldsymbol{\varepsilon}_*(\cdot)\}$. In fact, the F-H paper only proves this convergence for ordinary RLS so that the fundamental optimization process is not taken into account.

Whereas no known convergence proof for either the estimator or the feasible set exists for any volume or trace algorithm, a recent result indicates some theoretical support for the favorable convergence behavior of these methods which is observed in practice. The following has been proven for a specific UOBE class ($\alpha_n(\lambda_n^*)$ sequence chosen arbitrarily, $\beta_n(\lambda_n) = \lambda_n$ optimized) in [21]. By arguments in this paper, the more general result follows:

Theorem 3 *For any UOBE algorithm in which μ_v is minimized, if there exists $\lambda_n^* > 0$, then there also exists a large neighborhood of weights around λ_n^*, say $\mathcal{N}_{\lambda_n^*}$ (including, e.g., all λ_n such that $0 < \lambda_n \leq \lambda_n^*$), such that if $\lambda_n \in \mathcal{N}_{\lambda_n^*}$ is used, $\mu_v(n) < \mu_v(n-1)$.*

Though it has not been formally proven, it is likely that a similar result pertains to trace algorithms as well.

Theorem 3 indicates that the ellipsoid volume will tend to some unspecified size in some unspecified manner. If we consider the ratio

$$\nu(n) \stackrel{\text{def}}{=} \frac{\mu(n)}{\mu(n-1)} \tag{33}$$

(where $\mu(n)$ means either $\mu_v(n)$ or $\mu_t(n)$), for example, the rate at which $\nu(n)$ approaches unity will determine the convergence behavior of the ellipsoid. Suppose, for example, that

$$\nu(n) \sim 1 - \frac{1}{n^2} = \frac{n^2-1}{n^2}. \tag{34}$$

In this case

$$\lim_{n\to\infty} \mu(n) = \mu(1) \lim_{n\to\infty} \prod_{\tau=1}^{n} \frac{\tau^2-1}{\tau^2} = \frac{1}{2}\mu(1) \neq 0. \tag{35}$$

On the other hand, if

$$\nu(n) \sim 1 - \frac{1}{n} = \frac{n-1}{n}, \tag{36}$$

then,

$$\lim_{n\to\infty} \mu(n) = \mu(1) \lim_{n\to\infty} \prod_{\tau=1}^{n} \frac{\tau-1}{\tau} = \mu(1) \lim_{n\to\infty} \frac{1}{n} = 0. \tag{37}$$

This result has not been clearly understood, and its finding offers some hope that a proof of convergence (in some sense) for the volume and trace algorithms may be found in the white noise case.

It has frequently been noted that the hyperellipsoidal bounding sets resulting from UOBE algorithms can be quite "loose" supersets of the exact feasibility sets (polytopes) (e.g. [27],[28]), particularly in "finite" time[4]. However, many simulation studies in the literature (white noise case) have shown the volume of the ellipsoids to become quite small in the "long term." Further, as

[4] Norton has proposed the use of *inner* bounds as a possible remedy for this problem [28],[29]

we and other researchers have demonstrated, the empirical convergence and tracking properties of the UOBE estimator are favorable in spite of the few data used. This is an indication that the *presence* of the ellipsoid and the optimization procedure centered on it, are quite useful for parameter identification, regardless of our present inability to completely understand its behavior in theory. Theorem 3 offers further support for "good behavior" of this class of algorithms.

SM-SA. The SM-SA algorithm provides an interesting "bridge" between the present discussion and that of D-H/OBE to follow. SM-SA represents the authors' pursuit of a converging, interpretable UOBE algorithm. The algorithm is so-named because it is equivalent in form to the so-called ***stochastic approximation (SA)*** algorithm (e.g. [30]) as we discuss below. Our work has shown that a feature which tends to promote convergence of the ellipsoid is the prevention of "drifting" of the covariance matrix toward infinity. In an *unweighted* RLS algorithm, this problem is eliminated by normalizing the covariance matrix to the time n, that is, by replacing $\boldsymbol{C}(n)$ by $(1/n)\boldsymbol{C}(n)$. In principle, if the sequence $\{\boldsymbol{x}(\cdot)\}$ represents a stationary stochastic process with appropriate ergodicity properties, then $(1/n)\boldsymbol{C}(n)$ will tend to $\mathcal{E}\left\{\boldsymbol{x}(n)\boldsymbol{x}^H(n)\right\}$. Clearly, however, this strategy may not work for weights determined by SM considerations as $\boldsymbol{C}(n)$ may grow much faster than n. In the SM-SA approach, SM-WRLS with either volume or trace minimization is modified so that covariance matrix is normalized to the sum of the *weights*, say $\Lambda_n^* \stackrel{\text{def}}{=} \sum_{\tau=1}^{n} \lambda_\tau^*$. Accordingly,

$$\boldsymbol{C}(n) = \frac{\Lambda_{n-1}^*}{\Lambda_n^*}\boldsymbol{C}(n-1) + \frac{\lambda_n^*}{\Lambda_n^*}\boldsymbol{x}(n)\boldsymbol{x}^H(n). \tag{38}$$

Since $\Lambda_n^* = \Lambda_{n-1}^* + \lambda_n^*$, we find that

$$\alpha_n(\lambda_n) = \frac{\Lambda_{n-1}^*}{\Lambda_{n-1}^* + \lambda_n} \tag{39}$$

$$\beta_n(\lambda_n) = \frac{\lambda_n}{\Lambda_{n-1}^* + \lambda_n}. \tag{40}$$

While the *ellipsoid* associated with the SM-SA algorithm has not been proven to converge, the method is, of course, subject to the volume (and trace) contraction rule specified by Theorem 3. However, unlike the F-H/OBE and SM-WRLS algorithms, conditions under which the ***estimator***, $\boldsymbol{\Theta}(n)$, converges to $\boldsymbol{\Theta}_*$ can be clearly stated in this case[5].

Theorem 4 *Sufficient conditions for convergence of the SM-SA estimator in the sense*

$$\lim_{n\to\infty} \boldsymbol{\Theta}(n) \xrightarrow{a.s.} \boldsymbol{\Theta}_* \tag{41}$$

[5]We shall find that these conditions also apply to D-H/OBE.

are

$$\begin{aligned} \beta_n(\lambda_n^*) &\geq 0 \\ \sum_{\tau=1}^{\infty} \beta_\tau(\lambda_\tau^*) &= \infty & (42) \\ \sum_{\tau=1}^{\infty} \beta_\tau^2(\lambda_\tau^*) &< \infty. & (43) \end{aligned}$$

Sketch of Proof: The function sequences $\{\alpha_n(\lambda_n)\}$ and $\{\beta_n(\lambda_n)\}$ can be replaced by the following, and the optimization done over $\bar{\lambda}_n$, without affecting the optimization:

$$\bar{\alpha}_n(\bar{\lambda}_n) = 1 - \bar{\lambda}_n \tag{44}$$

$$\bar{\beta}_n(\bar{\lambda}_n) = \bar{\lambda}_n. \tag{45}$$

This follows immediately from the fact that the ratios $q_n(\lambda_n^*) = \beta_n(\lambda_n^*)/\alpha_n(\lambda_n^*)$ and $\bar{q}_n(\bar{\lambda}_n^*) = \bar{\beta}_n(\bar{\lambda}_n^*)/\bar{\alpha}_n(\bar{\lambda}_n^*)$ must be equal (to, say, p_n^*) according to Theorem 1, from which $\beta_n(\lambda_n^*) = \bar{\beta}_n(\bar{\lambda}_n^*) = p_n^*/(1+p_n^*)$ and $\alpha_n(\lambda_n^*) = \bar{\alpha}_n(\bar{\lambda}_n^*) = 1/(1+p_n^*)$. The WRLS algorithm with the weighting strategy given by (44) and (45), is frequently referred to as the SA algorithm for identifying linear parametric models. The work of Robbins and Monroe [31] and Blum [32] on the SA algorithm results in the sufficient conditions of the theorem. □

Let us henceforth adopt the simpler weighting strategy (44) and (45) for SM-SA. Remarkably, the weighting strategy that emerges here is identical to the Dasgupta-Huang weighting, resulting in the same convex combination of past covariance matrix and incoming outer product. However, this weighting strategy arises in a very different context in which the objective is to minimize the volume or trace of the hyperellipsoid at each step — if such can still be accomplished. Indeed, there is an ellipsoid associated with D-H OBE at each step, but the "usual" measures of its size are ignored in the optimization process – sacrificing interpretability. On the other hand, SM-SA does not inherent the convergence properties of D-H/OBE (described below). Unlike previous intepretable UOBE algorithms, however, SM-SA does have known conditions for estimator convergence. Further, SM-SA also exhibits the desirable property of "covariance boundedness" which we conjecture will be required for a volume or trace algorithm to converge in the set theoretic sense of D-H/OBE.

4.2 D-H/OBE and the Issue of κ Minimization

In the work above, we have discussed the fact that volume and trace UOBE algorithms are intepretable with respect to their principles of operation, but lacking in well understood convergence properties. Some significant progress on the convergence issue is cited, and it seems likely that the elusive convergence proof for at least some classes of volume and trace algorithms will be discovered. In this section, we briefly examine the problem from the "other direction." Given the D-H/OBE algorithm with its desirable convergence proof, can it be shown that this algorithm is actually performing according to "interpretable" principles?

Of the fundamental variations on UOBE, D-H/OBE is the most recent to be published. The technique is unlike all other existing methods in the use of κ minimization. This minimization approach, in conjunction with weighting strategy (44) and (45), provides the means with which to prove asymptotic and exponential convergence of the ellipsoid, and cessation of updating, using Lyaponov theory. From an analytical point of view, the reason for the choice of the κ optimization criterion is that $\kappa(n)$ is a bound on the Lyapunov function used in the minimization at time n, and the convergence of the Lyapunov function is used to prove convergence of the algorithm. Upon convergence, the residuals, $\varepsilon(\cdot, \Theta(\cdot))$ are guaranteed to remain in the "dead zone" indicated by the error bounds, i.e., as $\tau \to \infty$, $\| \varepsilon(\tau, \Theta(\tau-1)) \|^2 < \gamma_\tau$.

From an interpretive point of view, however, diminishing $\kappa(n)$ is not clearly helpful because its magnitude is not clearly related to the "size" of the set $\bar{\Omega}(n)$. Dasgupta and Huang [14] argue simply that $\kappa(n)$ is "a bound on the estimation error," and should be minimized. Norton and Mo [16] dispute this claim writing "[$\kappa(n)$] is not a bound on the parameter error, nor does it bear a simple relation to it."

In this section, we wish to determine whether D-H/OBE is, in fact, performing according to some interpretable principles. To begin, let us use Theorem 2 in conjunction with (44) and (45) to write a quadratic for the optimal root at time n for D-H/OBE,

$$
\begin{aligned}
F_\kappa^{D-H}(s) = {} & s^2 \left[\gamma_n (G(n)-1)^2 - \kappa(n-1)(G(n)-1)^2 + \| \varepsilon(n, \Theta(n-1)) \|^2 (G(n)-1)\right] \\
& + 2s \left[(\gamma_n - \kappa(n-1))(G(n)-1) + \| \varepsilon(n, \Theta(n-1)) \|^2\right] + \left[\gamma_n - \kappa(n-1) - \| \varepsilon(n, \Theta(n-1)) \|^2\right]
\end{aligned}
\tag{46}
$$

For future reference, let us also write a similar expression for UOBE with $\beta_n(\lambda_n) = \lambda_n$ and $\alpha_n(\lambda_n) = 1$. This latter case is similar to the SM-WRLS setup, except that κ minimization is used. For this reason we write "$F_\kappa^{SM-WRLS}(s)$":

$$
F_\kappa^{SM-WRLS}(s) = \gamma_n G^2(n) s^2 + 2\gamma_n G(n) s + (\gamma_n - \| \varepsilon(n, \Theta(n-1)) \|^2). \tag{47}
$$

The reasons for including (47) will become apparent momentarily. Dasgupta and Huang [14] show that an optimal weight in the sense of minimizing $\kappa(n)$ exists iff[6]

$$\| \epsilon(n, \Theta(n-1)) \|^2 > \gamma_n - \kappa(n-1). \tag{48}$$

Accordingly, this simple and computationally inexpensive ($\mathcal{O}(m)$) test may be employed to determine whether the the current data set $(y(n), x(n))$ is useful in the sense of minimizing $\kappa(n)$. Interestingly, the test (48) is tantamount to testing the zero order coefficient of quadratic F_κ^{D-H} for negativity. This is reminiscent of a similar test which can be performed for *any* volume or trace algorithm (see remarks below Theorem 1). However, that checking of the zero order coefficient F_κ^{D-H} should be a sufficient test for an optimal weight is not apparent as it is in the volume or trace cases. In particular, this is because the second order coefficient of F_κ^{D-H} need not be positive. Consequently, Dasgupta and Huang go to some effort to verify (48) as a test, and a set of rules centered on the second order coefficient is presented for finding the optimal weight if the test is met[7]. Let us juxtapose this fact with the following:

Theorem 5 *1. Consider the SM-WRLS algorithm. If $\kappa(n)$ is to be minimized at time n, a necessary and sufficient test for the existence of an optimal (κ-minimizing) weight, say $\lambda^*_{n,\kappa}$, is that the zero order coefficient of (47) be negative:*

$$\| \epsilon(n, \Theta(n-1)) \|^2 > \gamma_n. \tag{49}$$

2. Again consider SM-WRLS. Test (49) is also a sufficient condition for the existence of an optimal volume ($\lambda^*_{n,v}$) or trace ($\lambda^*_{n,t}$) weight.

3. Item 2 is true for any volume or trace minimizing UOBE algorithm.

Sketch of Proof: Item 1 is proven in [3],[17]. (One key feature of $F_\kappa^{SM-WRLS}$ which facilitates this result is that its second order coefficient is always positive. This is not true of F_κ^{D-H}.) Item 3 follows the fact that, if (49) holds, then a_0 of F_v, and b_0 of F_t (see Theorem 1) are both negative. Now see the remarks under Theorem 1. Item 2 is a special case of 3. □

The point of including Theorem 5 is to illustrate one case (SM-WRLS) in which κ minimization has many implications for intepretable performance. Indeed, (49) is a very powerful test. It is an indicator that not only $\kappa(n)$, but also either of the other two (interpretable) measures can be minimized at time n for SM-WRLS. Further note that, due to Theorem 3, κ *minimization implies*

[6] Their work is carried out for the one-dimensional case.

[7] Actually, a simpler rule is available. Because of the weighting strategy, $\lambda^*_{n,\kappa}$ must be in the interval $(0,1)$. A little thought will indicate that, once (48) is met, $F_\kappa^{D-H}(s) = 0$ can only have a root on $(0,1)$ if $F_\kappa^{D-H}(1) > 0$. If this is the case, then the quadratic equation can be used to find the root. Otherwise $\lambda^*_{n,\kappa}$ is taken to be zero or some predetermined number on $(0,1)$, depending on the relative magnitudes of $F_\kappa^{D-H}(0)$ and $F_\kappa^{D-H}(1)$

likely volume decrease (though not optimally) as well. Coincidently, (49) can also be (suboptimally, since it is only a sufficient condition) used to test whether volume or trace can be minimized at time n for any UOBE algorithm[8]. We must not lose site of the very important fact that convergence has only been proven for the D-H/OBE, and none of these findings does anything to change that fact.

Obviously, the next step is to inquire whether the D-H test (48) has similar implications for interpretable measures. Unfortunately, the answer appears to be no. Whereas (49) is equivalent to testing whether $a_0 + K_v < 0$ or $b_0 + K_t < 0$ with both K_v and K_t positive, (48) is only equivalent to testing whether $a_0 + \bar{K}_v < 0$ or $b_0 + \bar{K}_t < 0$, where neither $\bar{K}_v$ nor $\bar{K}_t$ is necessarily positive. Unfortunately, the truth of (48) is therefore not sufficient to assure that a_0 and b_0 are negative. If it were additionally known that

$$G(n) > mk, \tag{50}$$

then (48) would be a sufficient condition for the existence of $\lambda^*_{n,v}$ and $\lambda^*_{n,t}$ in the D-H case, and an indicator that any weight, even $\lambda^*_{n,\kappa}$, would likely diminish the volume. Because of the weighting strategy used in D-H/OBE, however, (50) does not hold in general. While some heuristic arguments can be made indicating circumstances under which (50) might be true, support for the notion that the D-H test might be similar to a volume or trace test is very weak in these terms.

So, in the analysis above at least, the D-H test comes intriguingly close to being a check for the existence of $\lambda^*_{n,v}$ or $\lambda^*_{n,t}$ (hence for an indicator that D-H/OBE is minimizing both κ and volume), but falls somewhat short. Again, we have not been able to find that convergence and intepretability exist in a single UOBE algorithm. However, the connections that apparently exist between D-H/OBE and more interpretable algorithms offer some hope that a meaningful interpretation of the dynamics of D-H/OBE might ultimately be found.

5 Summary and Conclusions

We have shown that all existing OBE, and, in fact, a very broad class of OBE algorithms, can be unified into a single framework which we have called the UOBE algorithm. This framework is based on generalized WRLS in which very wide classes of "forgetting factors" and data weights may be employed. Different instances of UOBE are distiguished by their weighting policies and the criteria used to determine their optimal values.

With the UOBE as a framework for discussion, we then turned our attention to existing algorithms. The main advantage of those which minimize ellipsoid volume and trace is the ease with which the performance principles are interpreted. However, to date no volume or trace algorithm

[8] This idea has been employed in [17]–[19] as an efficient way to implement the testing for real-time applications.

has been formally shown to converge in set-theoretic terms. We have however, introduced a new algorithm, SM-SA, for which conditions may be stated for convergence of the *estimator*. Several results are presented which offer promise that a proof of set convergence for volume and trace algorithms will ultimately be found.

Interestingly, SM-SA uses an equivalent weighting strategy to D-H/OBE, the only published UOBE algorithm for which set convergence and cessation of updating has been proven. D-H/OBE, however, uses κ minimization which does not lend itself well to interpretation of algorithm performance. An inquiry into the interpretability of D-H/OBE yielded some interesting connections of this method to volume and trace algorithms, but fell short of showing that D-H/OBE in fact minimizes something meaningful at each step. It was discovered that κ minimization *can* imply volume or trace minimization, but this was not demonstrated for any *converging* (ellipsoid) algorithm.

Hence the pursuit of an interpretable, set converging UOBE algorithm remains an open issue. The UOBE framework developed in this paper should be an asset in the discovery of this desirable algorithm.

References

[1] E. Walter and H. Piet-Lahanier, "Estimation of parameter bounds from bounded-error data: A survey," *Mathematics and Computers in Simulation*, vol. 32, pp. 449-468, 1990.

[2] M. Milanese and A. Vicino, "Estimation theory for dynamic systems with unknown but bounded uncertainty: An overview," *Proceedings of the* 9^{th} *IFAC / IFORS Symp. on Identification and System Parameter Estimation*, vol. 2, pp. 859-867, Budapest, July 1991.

[3] J.R. Deller, Jr., M. Nayeri, and S.F. Odeh, "System identification using set-membership-based signal processing," *IEEE Proceedings* (in review).

[4] E. Fogel and Y.F. Huang, "On the value of information in system identification - Bounded noise case," *Automatica*, vol. 18, pp. 229-238, 1982.

[5] J.R. Deller, Jr. and T.C. Luk, "Linear prediction analysis of speech based on set-membership theory," *Computer Speech and Language*, vol. 3, pp. 301-327, Oct. 1989.

[6] J.P. Norton, *An Introduction to Identification*, London and Orlando, Florida: Academic Press, 1986.

[7] C.R. Johnson, *Lectures on Adaptive Parameter Estimation*, Englewood Cliffs, New Jersey: Prentice-Hall, 1988.

[8] D. Graupe, *Time Series Analysis, Identification, and Adaptive Systems*, Malabar, Florida: Krieger, Ch. 5, 1989.

[9] J.R. Deller, Jr. and D. Hsu, "An alternative adaptive sequential regression algorithm and its application to the recognition of cerebral palsy speech," IEEE Trans. Circuits and Systems, vol. CAS-34, pp. 782-787, July 1987.

[10] J.R. Deller, Jr., "A 'systolic array' formulation of the optimal bounding ellipsoid algorithm," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 37, pp. 1432-1436, Sept. 1989.

[11] J.R. Deller, Jr., "Set-membership identification in digital signal processing," *IEEE ASSP Magazine*, vol. 6. pp. 4-22, Oct. 1989.

[12] J.R. Deller, Jr. and S.F. Odeh, "Implementing the optimal bounding ellipsoid algorithm on a fast processor," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. '89*, Glasgow, vol. 2, pp. 1067-1070, May 1989.

[13] S.F. Odeh and J.R. Deller, Jr., "Systolic algorithms for adaptive set-membership identification," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. '90*, Albuquerque, vol. 5, pp. 2419-2422, Apr. 1990.

[14] S. Dasgupta and Y.F. Huang, "Asymptotically convergent modified recursive least squares with data dependent updating and forgetting factor for systems with bounded noise," *IEEE Trans. Information Theory*, vol. IT-33, pp. 383-392, 1987.

[15] M. Nayeri, J.R. Deller, Jr., and M.S. Liu, Paper in preparation.

[16] J.P. Norton and S.H. Mo, "Parameter bounding for time-varying systems," *Mathematics and Computers in Simulation*, vol. 32, pp. 527-534, 1990.

[17] J.R. Deller, Jr. and S.F. Odeh, "SM-WRLS algorithms with an efficient test for innovation," *Proceedings of the* 9^{th} *IFAC / IFORS Symp. on Identification and System Parameter Estimation*, vol. 2, pp. 1044-1049, Budapest, July 1991.

[18] S.F. Odeh and J.R. Deller, Jr., "An SM-WRLS algorithm with an efficient test for innovation: Simulation studies and complexity issues," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. '91*, Toronto, vol. 5, pp. 3413-3416, May 1991.

[19] J.R. Deller, Jr. and S.F. Odeh, "Adaptive set-membership identification in $\mathcal{O}(m)$ time for linear-in-parameters models," *IEEE Trans. Signal Process.* (in press).

[20] A.K. Rao, Y.-F. Huang, and S. Dasgupta, "ARMA parameter estimation using a novel recursive estimation algorithm with selective updating," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 38, pp. 447-457, March 1990.

[21] M. Nayeri, J.R. Deller, Jr., and M.M. Krunz, "Convergence and colored noise issues in bounding ellipsoid identification," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. '92*, San Francisco vol. 5, pp. 337-340, Mar. 1992.

[22] M. Krunz, M.S. dissertation in preparation, Michigan State University, East Lansing.

[23] G. Belforte and B. Bona, "An improved parameter identification algorithm for signals with unknown-but-bounded errors," *Proc.* 7^{th} *IFAC/IFORS Symp. on Identification and System Parameter Estimation*, York, pp. 1507-1512, 1985.

[24] G. Belforte, B. Bona, and V. Cerone, "Parameter estimation algorithms for a set-membership description of uncertainty," *Automatica*, vol. 26, pp. 887-898, 1990.

[25] T. Söderström and P. Stoica, *System Indentification*, Englewood Cliffs, NJ: Prentice-Hall, 1989.

[26] G.C. Goodwin and K.S. Sin, *Adaptive Filtering, Prediction, and Control*, Englewood Cliffs, NJ: Prentice-Hall, 1984.

[27] G. Belforte, B. Bona, and V. Cerone, "Identification, structure selection and validation of uncertain models with set-membership error description," *Mathematics and Computers in Simulation*, vol. 32, pp. 561-569, 1990.

[28] J.P. Norton, "Recursive computation of inner bounds for the parameters of linear models," *Int. J. Control*, vol. 50, pp. 2423-2430, 1989.

[29] A. Vincino and M. Milanese, "Optimal approximation of feasible parameter set in estimation with unknown but bounded noise," *Proceedings of the* 9^{th} *IFAC / IFORS Symp. on Identification and System Parameter Estimation*, vol. 2, pp. 1365-1369, Budapest, July 1991.

[30] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*, Cambridge, MA: MIT Press, 1983.

[31] H. Robbins and S. Monroe, "A stochastic approximation method," *Ann. of Math. Stat.*, vol. 22, pp. 400-407, 1951.

[32] J.R. Blum, "Multidimensional stochastic approximation methods," *Ann. of Math. Stat.*, vol. 25, pp. 737-744, 1954.

Figure and Table List

At time n,

1. In conjunction with the incoming data set $(\boldsymbol{y}(n), \boldsymbol{x}(n))$, find optimal values of α_n and/or β_n, say α_n^* and/or β_n^*. Optimality criteria are described in the text;
2. If optimal positive (and sometimes further constrained) values α_n^* and/or β_n^* do not exist, then discard the data set (set $\beta_n^* = 0$ and / or $\alpha_n^* = 1$);
3. Update $\boldsymbol{C}(n)$, $\boldsymbol{\Theta}(n)$, and $\kappa(n)$ using (6), (7), and a recursion for $\kappa(\cdot)$ described in Lemma 1.

Figure 1: General steps of the UOBE algorithm.

Table 1: Specification of Existing UOBE Algorithms

Algorithm	$\alpha_n(\lambda_n^*)$	$\beta_n(\lambda_n^*)$	Optimization
F-H/OBE	$1/\kappa(n-1)$	λ_n^*/γ_n	$\mu_v(n)$ or $\mu_t(n)$
SM-WRLS	1	λ_n^*	$\mu_v(n)$ or $\mu_t(n)$
Dual SM-WRLS	λ_n^*	1	$\mu_v(n)$ or $\mu_t(n)$
D-H/OBE	$1-\lambda_n^*$	λ_n^*	$\kappa(n)$
SM-SA	$\Lambda_{n-1}^*/(\Lambda_{n-1}^*+\lambda_n^*)$ $\Lambda_n^* \stackrel{\text{def}}{=} \sum_{\tau=1}^{n}\lambda_\tau^*$	$\lambda_n^*/(\Lambda_{n-1}^*+\lambda_n^*)$	$\mu_v(n)$ or $\mu_t(n)$

A MILDLY WEAKER SUFFICIENT CONDITION IN IIR ADAPTIVE FILTERING[1]

Majid Nayeri[2]

ABSTRACT

The cross-covariance matrix of two stable autoregressive (AR) sequences is considered. A mildly weaker condition is identified which ensures the nonsingularity of this matrix. As one consequence of this result, a weaker sufficient condition is obtained which would guarantee the unimodality of the mean-square output error surface of an IIR adaptive filter with white noise excitation.

PERMISSION TO PUBLISH THIS ABSTRACT SEPARATELY IS GRANTED.

EDICS: 4.3.2 - Analysis of Adaptive Filters.

---

[1] Manuscript received _______________; revised on _______________.

[2] The author is with the Department of Electrical Engineering, Michigan State University, East Lansing, MI, 48824-1226.

This work was supported in part by the Office of Naval Research (ONR) under grant no. N00014-91-J-1329, and in part by the National Science Foundation (NSF) under grant no. MIP-9016734.

Submitted to IEEE Transaction on Signal Processing

I. INTRODUCTION

The two most popular approaches to filtering, identification, prediction, estimation, etc., are the equation error method and the output error method. The algorithms which are based on these two methods are often designed to minimize the mean square error. Specifically, they minimize the mean square equation error (MSEE) or the mean square output error (MSOE). The asymptotic analysis of these algorithms involves: 1) finding the attracting solution(s), and 2) investigating the local or global convergence to the solution(s).

Goodwin and Sin [1], and Ljung and Soderstrom [2] have treated the equation error based algorithms thoroughly from the perspective of convergence and applications. An attractive feature of the equation error is its unique minimum MSEE solution regardless of the linear model and the properties of the input. [3] However, this property is not shared with the output error method in general when the model is an infinite impulse response (IIR) filter. But, there are sufficient conditions which guarantee the uniqueness of the minimum MSOE solution in the identification setting [3], where the model (adaptive filter) can characterize the plant (unknown) completely.

The goal of this paper is to present a weaker sufficient condition than what was presented in [3] when the input is white noise and the model order exactly matches the order of the plant. This is ultimately intended to take us a step closer to establishing the necessary and sufficient conditions for the unimodality of the MSOE surface. In this paper, first a cross-correlation matrix is introduced in section II where some of its properties are outlined. In section III, these properties are used to extract the weaker sufficient condition for the uniqueness of the IIR identifier which would minimize the MSOE.

II. A CROSS-CORRELATION MATRIX

Consider the $m \times m$ matrix $\mathbf{P}$ defined by

$$\mathbf{P}(A, C, x, m) = E\left[\boldsymbol{\phi}_m(n)\boldsymbol{\psi}_m^T(n)\right] \tag{1}$$

where

[3] The input is assumed to be persistently exciting.

$$\phi_m(n) \triangleq \begin{bmatrix} \frac{1}{A(q^{-1})}x(n) \\ \cdot \\ \cdot \\ \frac{1}{A(q^{-1})}x(n-m+1) \end{bmatrix}, \quad \psi_m(n) \triangleq \begin{bmatrix} \frac{1}{C(q^{-1})}x(n) \\ \cdot \\ \cdot \\ \frac{1}{C(q^{-1})}x(n-m+1) \end{bmatrix}$$

Here, it is assumed that $A(q^{-1})$ and $C(q^{-1})$ are both N^{th} order stable polynomials of the form

$$\begin{aligned} A(q^{-1}) &= 1 + \sum_{i=1}^{N} a_i \, q^{-i} \equiv \prod_{i=1}^{N} (1-p_i q^{-1}) \\ C(q^{-1}) &= 1 + \sum_{i=1}^{N} c_i \, q^{-i} \equiv \prod_{i=1}^{N} (1-r_i q^{-1}) \end{aligned} \tag{2}$$

with $|p_i| < 1$ and $|r_i| < 1$, for $i = 1, \cdots, N$.

Ljung and Soderstrom [2] encountered the matrix **P** during the convergence analysis of the nonsymmetric instrumental variable method (IVM). They argued that **P** is singular only on a measure zero set which is determined by $det(\mathbf{P})=0$. As a result, it was concluded that the nonsymmetric IVM converges almost everywhere and that **P** is *generically* nonsingular. They also provided the sufficient condition for nonsingularity of **P** in [2, Lemma 4.7]. But, since we are only interested in the case where x(n) is white, let us restate this Lemma.

Lemma 1: [2]

Assume that $x(n)$ is a white sequence. Then, the matrix **P** *is nonsingular if either*

(i) $\dfrac{A(q^{-1})}{C(q^{-1})}$ *is strictly positive real,*

or

(ii) $m \geq N$

The positive realness in (i) is such a strong condition that it also guarantees **P** to be positive definite. An obvious special case is when $C(q^{-1}) = A(q^{-1})$. The sufficient condition stated in (ii) is much less restrictive than (i) which we will further explore next. In particular, it is of considerable interest to know how tight this sufficient condition is. That is, can $m \geq N-1$, or $m \geq N-2$, etc., replace (ii)? If so, a weaker sufficient condition has been identified.

The Toeplitz matrix **P** for which no symmetry assumption is assumed is fully determined by g_k , $1-m \leq k \leq m-1$, where

$$g_{j-i}(A,C) \triangleq P_{ij}(A,C,x,m) \qquad , \qquad i,j = 1,2,\cdots,m$$
$$= \frac{1}{2\pi j}\oint_{|z|=1} \frac{1}{z^N A(z^{-1})}\frac{1}{C(z)}\Phi_x(z)\, z^{(N-1)+j-i}\, dz \tag{3}$$

or, equivalently,

$$\mathbf{P}(A,C,x,m)^T = \mathbf{J}\,\mathbf{P}(A,C,x,m)\,\mathbf{J}$$
$$= \mathbf{P}(C,A,x,m) \tag{4}$$

where $\Phi_x(z)$ is the spectral density of $x(n)$ and

$$\mathbf{J} = \begin{bmatrix} 0 & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & 0 \end{bmatrix}$$

The following Lemma presents an identity involving the $N\times N$ matrix $\mathbf{P}$. In the sequel, $\mathbf{P}_m$ will be occasionally used to denote $\mathbf{P}(A,C,x,m)$ for brevity.

Lemma 2:

If $x(n)$ is a zero-mean white noise sequence with variance σ^2, then

$$\mathbf{P}_N - \left[\begin{array}{c|c} -\mathbf{a} & \begin{array}{c}\mathbf{I}\\ \hline \mathbf{0}^T\end{array}\end{array}\right]^T \mathbf{P}_N \left[\begin{array}{c|c} -\mathbf{c} & \begin{array}{c}\mathbf{I}\\ \hline \mathbf{0}^T\end{array}\end{array}\right] = \left[\begin{array}{c|c} \sigma^2 & \mathbf{0}^T \\ \hline \mathbf{0} & \mathbf{O}\end{array}\right] \tag{5}$$

where $\mathbf{a}=[a_1\, a_2\, \cdots\, a_N]^T$, $\mathbf{c}=[c_1\, c_2\, \cdots\, c_N]^T$. Also, $\mathbf{0}$ is the $N-1$ zero vector, and $\mathbf{O}$ and $\mathbf{I}$ are the $N-1\times N-1$ zero and identity matrices, respectively.

proof: See the Appendix. □

One immediate consequence of this result is stated below in Lemma 3 and Theorem 1 which establish the non-singularity of $\mathbf{P}_{N-1}$ by identifying a direct relationship between $det(\mathbf{P}_N)$ and $det(\mathbf{P}_{N-1})$.

Lemma 3:

Let us define

$$\Delta_m = \det\left(\mathbf{P}(A,C,x,m)\right)$$

Then, we have

(i) $\displaystyle \Delta_N = \frac{\sigma^{2N}}{\prod_{i,j=1}^{N}(1-p_i r_j)}$

$$\text{(ii)}\ \Delta_{N-1} = \frac{1}{\sigma^2}(1 - a_N c_N)\,\Delta_N = \frac{(1 - a_N c_N)}{\prod_{i,j=1}^{N}(1 - p_i r_j)}\,\sigma^{2(N-1)} \tag{6}$$

proof: See the Appendix. □

Theorem 1:

The matrix **P** *is nonsingular for* $m \geq N-1$.

Proof:

Since $|a_N| < 1$ and $|c_N| < 1$ for the stable polynomials $A(z^{-1})$ and $C(z^{-1})$, according to Lemma 3(ii) Δ_{N-1} would be nonzero. Since **P** is nonsingular for $m \geq N$ according to Lemma 1, part (ii), the result immediately follows. □

Theorem 1 presents a weaker sufficient condition for nonsingularity of **P**, namely, $m \geq N-1$. But, is this the *weakest* sufficient condition? To answer this question, consider the following example.

Example 1:

Let $N=3$ and $m=1$ which signifies the case $m = N-2$. Then, **P** is a scalar which is given by

$$\mathbf{P} = g_0(A, C) = \frac{(1-a_3c_3)^2 - (a_1c_3-c_2)(c_1a_3-a_2)}{\prod_{i,j=1}^{3}(1-p_i r_j)}$$

However, if c_1=−2.4, c_2=1.91, c_3=−0.504, a_1=a_2=0, and a_3=0.2854 , which correspond to two stable polynomials $A(q^{-1})$ and $C(q^{-1})$, result in $\mathbf{P}=0$. □

Similar examples can be found for the case where $m < N-2$. Therefore, it is concluded that $m \geq N-1$ represents the weakest sufficient condition.

One significant implication of the above result is presented next where the connection between **P** and the stationary points of MSOE is established and then a weaker condition for the uniqueness is stated.

III. STATIONARY POINTS OF MSOE

Consider the system identification model where it is assumed that

$$y(n) = \frac{D(q^{-1})}{C(q^{-1})} x(n) + v(n) \tag{7}$$

where

$$D(q^{-1}) = \sum_{i=0}^{n_d} d_i \; q^{-i} \qquad , \qquad C(q^{-1}) = 1 + \sum_{i=1}^{n_c} c_i \; q^{-i}$$

are coprime polynomials in q^{-1} and $v(n)$ is additive noise. We further assume that the zeros of $C(z^{-1})$ are inside the unit circle and that the additive noise $v(n)$ is a zero mean stochastic process which is independent of $x(n)$. Let the adaptive system be an IIR filter whose input-output relation is governed by

$$\hat{y}(n) = \frac{B(q^{-1})}{A(q^{-1})} x(n), \tag{8}$$

where

$$B(q^{-1}) = \sum_{i=0}^{n_b} b_i \; q^{-i} \qquad , \qquad A(q^{-1}) = 1 + \sum_{i=1}^{n_a} a_i \; q^{-i}$$

If we define the output error by $e(n) = y(n) - \hat{y}(n)$, the MSOE is given by

$$E[e^2(n)] = E\left[\left[\frac{D(q^{-1})}{C(q^{-1})} - \frac{B(q^{-1})}{A(q^{-1})}\right] x(n)\right]^2 + E[v^2(n)], \tag{9}$$

The stationary points of (9) are the solutions of

$$E\left[\left[\frac{D(q^{-1})}{C(q^{-1})} - \frac{B(q^{-1})}{A(q^{-1})}\right] x(n) \cdot \frac{B(q^{-1})}{A^2(q^{-1})} \; x(n-i)\right] = 0, \tag{10}$$

$$E\left[\left[\frac{D(q^{-1})}{C(q^{-1})} - \frac{B(q^{-1})}{A(q^{-1})}\right] x(n) \cdot \frac{1}{A(q^{-1})} x(n-j)\right] = 0, \tag{11}$$

$$1 \le i \le n_a \;\; , \;\; 0 \le j \le n_b$$

It is shown in [3] that (10) and (11) accept a unique minimally realizable solution if for white input $x(n)$

$$n^* = min(n_a - n_c \; , \; n_b - n_d) \ge 0 \tag{12}$$

$$n_b + 1 - n_c \ge 0 \tag{13}$$

The expression (12) merely states that the adaptive filter should be of sufficient order. Here, we consider the case where $n_a = n_c$ and $n_b = n_d$ which is referred to as exactly matching (EM) adaptive IIR filter. Violating (13) may lead to the existence of local minima on MSOE surface as shown in [6], [7]. Weakening (13) may naturally seem to be in contradiction with this result. However, the case considered in [6] and [7] is for $n_b = 0$, $n_c = 3$ and, therefore, $n_b + 1 - n_c = -2$. Here, we show that $n_b + 1 - n_c = -1$, for EM case, is the underlying weakest sufficient condition for uniqueness of the global minimum.

***Degenerated Solutions*:** The existence of stable degenerated solutions is a sufficient condition for the existence of local minima [5,6,7]. For the EM case, the degenerated solutions corresponding to $n_b + 2 = n_c$ are found by setting $B(q^{-1}) \equiv 0$ in (10) and (11). As a result, (10) vanishes and (11) is reduced to

$$\mathbf{P}(A, C, x, n_b+1) \cdot \begin{bmatrix} d_0 \\ \cdot \\ \cdot \\ \cdot \\ d_{n_b} \end{bmatrix} = 0 \tag{14}$$

Corollary 2:

For a given stable system (7) with white input, no stable polynomial $A(q^{-1})$ satisfies (14), and hence no degenerated solution exists.

Proof:

If such a stable solution exists, say A^*, then $d_0, \ldots, d_{n_b}$ can be solved for. The matrix $\mathbf{P}$ has to be singular at A^* since not all d_i are zero. But, according to Theorem 1, $\mathbf{P}(A, C, x, n_b+1)$ is nonsingular for any stable A^* and C since $n_b + 1 = n_c - 1 = n_a - 1$. □

***Other Solutions*:** For the EM case, all the stationary, nondegenerated points of MSOE corresponding to $n_b + 2 = n_c$ which solve (10) and (11) fulfill (see [5])

$$\mathbf{P}(A\bar{A}, C\bar{A}, x, n_a+n_b-n_L+1) \cdot \begin{bmatrix} h_0 \\ \cdot \\ \cdot \\ \cdot \\ h_{n_a+n_b-n_L} \end{bmatrix} = 0 \tag{15}$$

where for $L(q^{-1}) = 1 + \sum_{i=1}^{n_L} l_i q^{-i}$ (, $n_L \geq 0$)

$$A(q^{-1}) \equiv \bar{A}(q^{-1})L(q^{-1}) \quad , \quad B(q^{-1}) \equiv \bar{B}(q^{-1})L(q^{-1})$$

in which $\bar{A}(q^{-1})$ and $\bar{B}(q^{-1})$ are coprime. Also, h_i are such that

$$\sum_{i=0}^{n_a+n_b-n_L} h_i \, q^{-i} = \bar{A}(q^{-1})D(q^{-1}) - \bar{B}(q^{-1})C(q^{-1}) \tag{16}$$

Theorem 2:

Consider an exactly matching adaptive IIR filter in which $n_b + 2 \geq n_c$. The MSOE surface of this filter is unimodal, with a unique global minimum, when the input is white.

Proof:

We note that

$$N = deg(A\bar{A}) = deg(C\bar{A}) = 2n_a - n_L \quad , \quad m = n_a + n_b - n_L + 1$$

and since

$$\begin{aligned} N - m &= n_a - n_b - 1 \\ &= n_c - n_b - 1 \leq 1 \end{aligned}$$

then, Theorem 1 implies the nonsingularity of $P(A\bar{A}, C\bar{A}, x, n_a+n_b-n_L+1)$ for any value of n_L . Therefore, (15) yields

$$h_i = 0 \ , \quad i = 0, \ldots, n_a+n_b-n_L \tag{17}$$

Using (17) in equation (16) reveals that

$$\frac{\bar{B}(q^{-1})}{\bar{A}(q^{-1})} = \frac{D(q^{-1})}{C(q^{-1})} \tag{18}$$

Since $C(q^{-1})$ and $D(q^{-1})$ are coprime polynomials, (19) implies that

$$\bar{A}(q^{-1}) = C(q^{-1}) \quad , \quad \bar{B}(q^{-1}) = D(q^{-1})$$

But this can happen only when $n_L = 0$. Therefore, there exists a unique stationary point which is a unique global minimum of MSOE and is givven by

$$A(q^{-1}) = C(q^{-1}) \quad , \quad B(q^{-1}) = D(q^{-1}) \tag{19}$$

This is a weaker sufficient condition than (13) for the EM filters. □

Example 2:

Consider the case where $n_b = n_d = 0$ and $n_a = n_c = 2$ which was considered by Stearns [9]. The MSOE surfaces of filters in this class were observed to be unimodal by examining different pole locations of the unknown system. Theorem 2 provides a proof in support of this observation. □

IV. CONCLUSION

A nonsymmetric Toeplitz matrix was introduced and a weaker sufficient condition for its invertibility was presented. This obtained weaker condition was used to conclude that if the adaptive IIR filter is exactly matching, the MSOE surface is unimodal if $(n_b+2) - n_c \geq 0$. This is a weaker sufficient condition than what is reported in the literature [3]. In fact, this can be regarded as the weakest sufficient condition in general.

REFERENCES

[1] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction, and Control.* Englewood Cliffs, NJ: Prentice Hall, Inc., 1984.

[2] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification.* Cambridge, MA: The MIT Press, 1983.

[3] T. Soderstrom and P. Stoica, "Some properties of the output error method," *Automatica*, vol. 18, no. 12, pp. 1692-1716, Dec. 1982.

[4] M. Nayeri, H. Fan and W. K. Jenkins, "Some Characteristics of Error Surfaces For Insufficient Order Adaptive IIR Filters," *IEEE Trans. on Acoust., Speech, and Signal Processing* , Vol. 38, No. 7, pp. 1222-1227, July 1990.

[5] T. Soderstrom, "On the uniqueness of maximum likelihood identification," *Automatica*, vol. 11, no. 2, pp. 193-197, 1975.

[6] H. Fan and M. Nayeri, "On Error Surfaces of Sufficient Order Adaptive IIR Filters," *IEEE Trans. on Acoust., Speech, and Signal Processing* , Vol. ASSP-37, No. 9, pp. 1436-1442, Sept. 1989.

[7] H. Fan and M. Nayeri, "Counter examples to Stearns' conjecture on error surfaces of adaptive IIR filters," *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, New York, NY, Apr. 11-14, 1988.

[8] M. Nayeri, "Uniqueness of MSOE Estimates in IIR Adaptive Filtering; A Search For Necessary Conditions," *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing* , pp. 1047-1051, Glasgow, Scotland, U. K., May 23-26, 1989.

[9] S. D. Stearns,"Error Surfaces of Recursive Adaptive Filters," *IEEE Trans. on Acoust., Speech, and Signal Processing*, Vol. ASSP-29, No. 3, pp. 763-766, June 1981.

APPENDIX

Proof of Lemma 2:
First, note that

$$\mathbf{a}^T \boldsymbol{\phi}_N(n) = x(n+1) - \frac{x(n+1)}{A(q^{-1})} \quad , \quad \mathbf{c}^T \boldsymbol{\psi}_N(n) = x(n+1) - \frac{x(n+1)}{C(q^{-1})} \tag{A1}$$

Then,

$$\begin{aligned}\mathbf{a}^T \mathbf{P}_N \mathbf{c} &= E\left[\left[x(n+1) - \frac{x(n+1)}{A(q^{-1})}\right]\left[x(n+1) - \frac{x(n+1)}{C(q^{-1})}\right]\right] \\ &= E\left[x^2(n+1)\right] + E\left[\frac{x(n+1)}{A(q^{-1})}\frac{x(n+1)}{C(q^{-1})}\right] - E\left[x(n+1)\frac{x(n+1)}{A(q^{-1})}\right] - E\left[x(n+1)\frac{x(n+1)}{C(q^{-1})}\right] \\ &= \sigma^2 + g_0 - \sigma^2 - \sigma^2 = g_0 - \sigma^2\end{aligned} \tag{A2}$$

since $x(n)$ is white. Also.

$$\begin{aligned}-\mathbf{a}^T \mathbf{P}_N \begin{bmatrix} \mathbf{I} \\ - - - \\ \mathbf{0}^T \end{bmatrix} &= -\mathbf{a}^T E\left[\boldsymbol{\phi}_N(n)\boldsymbol{\psi}_{N-1}^T(n)\right] \\ &= -E\left[\left[x(n+1) - \frac{x(n+1)}{A(q^{-1})}\right]\boldsymbol{\psi}_{N-1}^T(n)\right] \\ &= \left[g_1 \cdots g_{N-1}\right]\end{aligned} \tag{A3}$$

Similarly,

$$-\begin{bmatrix} \mathbf{I} \\ - - - \\ \mathbf{0}^T \end{bmatrix}^T \mathbf{P}_N \mathbf{c} = \begin{bmatrix} g_{-1} \\ \cdot \\ \cdot \\ \cdot \\ g_{1-N} \end{bmatrix} \tag{A4}$$

Finally, straightforward calculations suggest that

$$\begin{aligned}\begin{bmatrix} \mathbf{I} \\ - - - \\ \mathbf{0}^T \end{bmatrix}^T \mathbf{P}_N \begin{bmatrix} \mathbf{I} \\ - - - \\ \mathbf{0}^T \end{bmatrix} &= E\left[\boldsymbol{\phi}_{N-1}(n)\boldsymbol{\psi}_{N-1}(n)\right] \\ &= \mathbf{P}_{N-1}\end{aligned} \tag{A5}$$

Therefore,

$$\begin{bmatrix} -\mathbf{a} & \begin{array}{c}\mathbf{I} \\ \hline \mathbf{0}^T\end{array} \end{bmatrix}^T \mathbf{P}_N \begin{bmatrix} -\mathbf{c} & \begin{array}{c}\mathbf{I} \\ \hline \mathbf{0}^T\end{array} \end{bmatrix} = \left[\begin{array}{c|cccc} g_0-\sigma^2 & g_1 & \cdot & \cdot & g_{N-1} \\ \hline g_{-1} & & & & \\ \cdot & & & & \\ \cdot & & \mathbf{P}_{N-1} & & \\ \cdot & & & & \\ g_{1-N} & & & & \end{array}\right] \tag{A6}$$

where the right side of the equality in (A6) follows immediately using (A2), (A3), (A4), and (A5). But this an alternative representation of (5) and the proof is complete. □

Proof of Lemma 3:

(i) First, let p_i's be *distinct*. Now let

$$\begin{aligned} \lambda_i &= Residue\ of\ \frac{\sigma^2}{z^N\ A(z^{-1})\ C(z)}\ at\ z = p_i \\ &= \frac{\sigma^2}{C(p_i)\ \prod_{\substack{j=1 \\ j\neq i}}^{N} (p_i - p_j)} \end{aligned} \tag{A7}$$

Then, Equation (3) is reduced to

$$g_{j-i} = \sum_{k=1}^{N} p_k^{N-1-i}\ \lambda_k\ p_k^j \tag{A8}$$

For the special case when $m = N$, using (A8) in (1) gives

$$\mathbf{P}(A, C, x, N) = V_1\ \Lambda\ V_2 \tag{A9}$$

where

$$V_1 = \begin{bmatrix} p_1^{N-1} & p_2^{N-1} & \cdot\ \cdot\ \cdot & p_N^{N-1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ p_1 & p_2 & & p_N \\ 1 & 1 & & 1 \end{bmatrix}$$

and

$$V_2 = \begin{bmatrix} 1 & p_1 & \cdot\ \cdot\ \cdot & p_1^{N-1} \\ 1 & p_2 & & p_2^{N-1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & p_N & & p_N^{N-1} \end{bmatrix}$$

and

$$\Lambda = Diag\left[\ \lambda_1\ ,\ \lambda_2\ ,\ .\ .\ .\ .\ ,\ \lambda_N\ \right]$$

Therefore, since the involved matrices in (A9) are $N \times N$, it follows that

$$\begin{aligned}\Delta_N &= det(V_1)\, det(V_2)\, det(\Lambda) \\ &= \prod_{\substack{j=1\\ j<i}}^{N} (p_i - p_j) \prod_{\substack{j=1\\ j>i}}^{N} (p_i - p_j) \prod_{i=1}^{N} \lambda_i \\ &= \frac{\sigma^{2N}}{\prod_{i=1}^{N} C(p_i)} = \frac{\sigma^{2N}}{\prod_{i,j=1}^{N} (1 - p_i r_j)}\end{aligned} \tag{A10}$$

This result was derived under the assumption that p_i's are distinct. But, since the determinant is a continuous (actually analytic) function of the elements of a matrix, and since each element of the matrix $\mathbf{P}$ is analytic for stable polynomials $A(z^{-1})$ and $C(z^{-1})$ (see [2]) then Δ_m is analytic and therefore continuous. It then follows that if $A(z^{-1})$ has multiple poles, Δ_N is given by (A10).

(ii) Lemma 2 implies that

$$\det\left\{ \left[\begin{array}{c|c} -\mathbf{a} & \mathbf{I} \\ \hline & 0^T \end{array} \right]^T \mathbf{P}_N \left[\begin{array}{c|c} -\mathbf{c} & \mathbf{I} \\ \hline & 0^T \end{array} \right] \right\} = \det\left\{ \mathbf{P}_N - \left[\begin{array}{c|c} \sigma^2 & 0^T \\ \hline 0 & \mathbf{O} \end{array} \right] \right\} \tag{A11}$$

The determinants of the Companion form matrices in the left side of (A11) are equal to $(-1)^{N-1} a_N$ and $(-1)^{N-1} c_N$, respectively. As a result, (A11) can be written as

$$a_N c_N \Delta_N = \begin{vmatrix} g_0 - \sigma^2 & g_1 & \cdot & \cdot & \cdot & g_{N-1} \\ g_{-1} & g_0 & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & g_1 \\ g_{1-N} & \cdot & \cdot & \cdot & g_{-1} & g_0 \end{vmatrix} \tag{A12}$$

$$= \Delta_N - \sigma^2 \Delta_{N-1}$$

where the last equality follows by evaluating the determinant with respect to the first row or the first column. Therefore,

$$\Delta_{N-1} = \frac{1}{\sigma^2}(1 - a_N c_N)\, \Delta_N \tag{A13}$$

and the proof is complete. □

This work was supported by the National Science Foundation of the United States under Grant No. MIP-9016734 and by the United States Office of Naval Research under Grant No. N00014-91-J-1329.

LAYER-WISE TRAINING OF FEEDFORWARD NEURAL NETWORKS BASED ON LINEARIZATION AND SELECTIVE DATA PROCESSING

By

Shawn David Hunt

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical Engineering

1992

ABSTRACT

LAYER-WISE TRAINING OF FEEDFORWARD NEURAL NETWORKS BASED ON LINEARIZATION AND SELECTIVE DATA PROCESSING

By

Shawn David Hunt

A class of algorithms is presented for training nonlinear feedforward neural networks using purely "linear" techniques. The algorithms are based upon linearizations of the network using error surface analysis, followed by a contemporary recursive least squares identification procedure which can be implemented using parallel processing. Specific algorithms are presented to estimate weights node-wise, layer-wise, and for estimating the entire set of network weights simultaneously. A procedure for modifying the algorithms to selectively use the training data and increase speed is also presented. A computationally inexpensive measure is developed with which to assess the effect of a particular training pattern on the weight estimates prior to its inclusion in any iteration. Data which do not significantly change the weights are not used in that iteration, obviating the computational expense of updating. Several experimental studies are presented showing the advantages of this class of algorithms. Specifically, the layer-wise algorithm is shown to be vastly superior to back-propagation in terms of the number of convergences and convergence rate. Additionally this algorithm is shown to be insensitive to the choice of initial weights and forgetting factor, eliminating two of the greatest problems in the implementation of existing training algorithms.

This work was supported by the National Science Foundation of the United States under Grant No. MIP-9016734 and by the United States Office of Naval Research under Grant No. N00014-91-J-1329.

NUMERICAL STABILITY AND CONVERGENCE ISSUES IN THE SM-WRLS ALGORITHM

By

Marwan Mahdi Krunz

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Department of Electrical Engineering

1992

ABSTRACT

NUMERICAL STABILITY AND CONVERGENCE ISSUES IN THE SM-WRLS ALGORITHM

By

Marwan Mahdi Krunz

This research is concerned with a particular class of *optimal bounding ellipsoid (OBE)* algorithms which implements an optimization criterion based on the volume of the optimal ellipsoid. The OBE algorithms belong to *set membership (SM)* identification techniques and are used to identify the parameters of linear system or signal models based on *a priori* information about the pointwise "energy bounds" on the error sequence. OBE algorithms define a set of solutions that takes the form of a "hyperellipsoid" in the parameter space. This ellipsoid is centered around the familiar WRLS estimate.

In this work, the convergence behavior of the ellipsoid for the class of OBE algorithms that utilizes the volume ratio measure is studied under different types of disturbances. The non-persistency in the excitation of the disturbances may result in the degeneration of the ellipsoid. The convergence of the ellipsoid under *both* persistently and non-persistently exciting colored noise is particularly investigated.

The conventional OBE algorithms with volume ratio measure employ a data selection strategy which is based on minimizing the volume of the ellipsoid and finding an optimal error minimization weight to be associated with the present datum. In this work, a new OBE algorithm, the *set membership past weight optimization (SM-PWO)*, is developed. The data selection technique in this algorithm is based on minimizing