REFERENCES

[1] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 1996, Tech. Rep. CRG-TR-96-1.

[2] G. J. McLachlan, D. Peel, and R. W. Bean, "Modelling high-dimensional data by mixtures of factor analyzers," *Comput. Statist. Data Anal.*, vol. 41, pp. 379–388, Jan. 2003.

[3] G. McLachlan, R. Bean, and L. B.-T. Jones, "Extension of the mixture of factor analyzers model to incorporate the multivariate t distribution," *Comput. Statist. Data Anal.*, vol. 51, pp. 5327–5338, 2007.

[4] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Comput.*, vol. 11, pp. 443–482, 1999.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data using the EM algorithm (with discussion)," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[6] X. L. Meng and D. A. van Dyk, "The EM algorithm—An old folk-song sung to a fast new tune," *J. Roy. Statist. Soc. B*, vol. 59, no. 3, pp. 511–567, 1997.

[7] J. Zhao, P. L. H. Yu, and Q. Jiang, "ML Estimation for factor analysis: EM or non-EM?," *Statist. Comput.*, vol. 18, no. 2, pp. 109–123, 2008.

[8] D. B. Rubin and T. T. Thayer, "EM Algorithms for factor ML analysis," *Psychometrika*, vol. 47, pp. 69–76, 1982.

[9] K. B. Petersen, O. Winther, and L. K. Hansen, "On the slow convergence of EM and VBEM in low-noise linear models," *Neural Comput.*, vol. 17, no. 9, pp. 1921–1926, 2005.

[10] K. G. Jöreskog, "Some contributions to maximum likelihood factor analysis," *Psychometrika*, vol. 32, no. 4, pp. 433–482, 1967.

[11] C. Liu, "The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence," *Biometrika*, vol. 81, pp. 633–648, 1994.

[12] X.-L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[13] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton, "SMEM algorithm for mixture models," *Neural Comput.*, vol. 12, no. 9, pp. 2109–2128, Sep. 2000.

[14] K. Lange, *Numerical Analysis for Statisticians*. New York: Springer-Verlag, 1999.

A-Optimality Orthogonal Forward Regression Algorithm Using Branch and Bound

Xia Hong, Sheng Chen, and Chris J. Harris

***Abstract*—In this brief, we propose an orthogonal forward regression (OFR) algorithm based on the principles of the branch and bound (BB) and A-optimality experimental design. At each forward regression step, each candidate from a pool of candidate regressors, referred to as $\mathcal{S}$, is evaluated in turn with three possible decisions: 1) one of these is selected and included into the model; 2) some of these remain in $\mathcal{S}$ for evaluation in the next forward regression step; and 3) the rest are permanently eliminated from $\mathcal{S}$. Based on the BB principle in combination with an A-optimality composite cost function for model structure determination, a simple adaptive diagnostics test is proposed to determine the decision boundary between 2) and 3). As such the proposed algorithm can significantly reduce the computational cost in the A-optimality OFR algorithm. Numerical examples are used to demonstrate the effectiveness of the proposed algorithm.**

***Index Terms*—Branch and bound (BB), experimental design, forward regression, structure identification.**

I. INTRODUCTION

A large class of nonlinear models and neural networks can be classified as a linear-in-the-parameters model [1], [2]. The linear-in-the-parameters models are well structured for adaptive learning, have provable learning and convergence conditions, have the capability of parallel processing, and have clear applications in many engineering applications [3]–[5]. A basic principle in practical nonlinear data modeling is the parsimonious principle that ensures the smallest possible model for the explanation of the observational data. For the linear-in-the-parameters models, the forward orthogonal least squares (OLS) algorithm efficiently constructs parsimonious models [6], [7], and has been a popular tool in associative neural networks such as fuzzy/neurofuzzy systems [8], [9] and wavelet neural networks [10], [11]. The algorithm has also been utilized in a wide range of engineering applications, e.g., aircraft gas turbine modeling [12], fuzzy control of multiple-input–multiple-output (MIMO) nonlinear systems [13], power system control [14], and fault detection [15].

In optimum experimental design [16], the model adequacy is evaluated by design criteria that are statistical measures of goodness of experimental designs by virtue of design efficiency and experimental effort. Quantitatively, model adequacy is measured as function of the eigenvalues of the design matrix. In order to produce a model with good generalization capabilities, the A-optimality composite cost function has been used as the model selection criterion in the A-optimality-based orthogonal forward regression (OFR) algorithms [17].

Note that the nonlinear system identification is an intractable optimization problem of mixed integer programming that involves both continuous variables, e.g., model parameters and discrete variables, e.g., enumeration of possible model terms. The principle of branch-

Manuscript received March 31, 2008; revised June 23, 2008; accepted July 25, 2008. First published September 30, 2008; current version published November 5, 2008.

X. Hong is with the School of Systems Engineering, University of Reading, Reading RG6 61Y, U.K. (e-mail: x.hong@reading.ac.uk).

S. Chen and C. J. Harris are with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: sqc@ecs.soton.ac.uk; cjh@ecs.soton.ac.uk).

Color versions of one or more of the figures in this brief are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2008.2003251

and-bound (BB) approach [18] is well understood in the operational research community and is by far the most widely used approach for optimization with mixed integer programming. The basic idea is that the search spaces are kept being divided into a feasible subset and an infeasible subset. The infeasible subset is initially determined by the current optimal solution, and then eliminated from further search efforts. A common difficulty with the BB is that this is only a modeling paradigm since there is always a gap between the idea and any specific problem in terms of the BB strategy design. For any application, it is necessary to integrate the BB procedure into the problem domain and to have provable results such that the infeasible subset definitely does not contain solutions superior to the current solutions.

We point out that in spite of the fact that the OFR algorithms are regarded as efficient model subset selection approaches, there is not only a practical need, but also the opportunities to further reduce significantly the computation cost of the OFR algorithms. In this brief, a new A-optimality OFR algorithm is introduced to reduce the search space/computation cost based on a new simple adaptive decision rule/boundary, which is shown to be a provable application of the BB technique.

II. The A-Optimality-Based Orthogonal Forward Regression Algorithm

A linear-in-the-parameter model [radial basis function (RBF) neural network, B-spline neurofuzzy network] can be formulated as [1], [2]

$$y(t) = \sum_{k=1}^{M} p_k\left(\mathbf{x}(t)\right)\theta_k + \xi(t) \tag{1}$$

where $t = 1, 2, \cdots, N$, and N is the size of the estimation data set. $y(t)$ is a system output variable and $\mathbf{x}(t) = [y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u)]^T$ is a system input vector of observables with assumed known dimension of $(n_y + n_u)$. $u(t)$ is a system input variable. $p_k(.)$ is a known nonlinear basis function, such as RBF, or B-spline fuzzy membership functions. $\xi(t)$ is an uncorrelated model residual sequence with zero mean and variance of σ^2. Equation (1) can be written in the matrix form as

$$\mathbf{y} = \mathbf{P}\Theta + \Xi \tag{2}$$

where $\mathbf{y} = [y(1), \cdots, y(N)]^T$ is the output vector, $\Theta = [\theta_1, \cdots, \theta_M]^T$ is the parameter vector, $\Xi = [\xi(1), \cdots, \xi(N)]^T$ is the residual vector, and $\mathbf{P}$ is the regression matrix with $p_k(\mathbf{x}(t))$ as the element at tth row and kth column. Denote the column vectors of $\mathbf{P}$ as $\mathbf{p}_j, j = 1, \ldots M$.

An orthogonal decomposition of $\mathbf{P}$ is

$$\mathbf{P} = \mathbf{W}\mathbf{A} \tag{3}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & a_{1,2} & \cdots & a_{1,M} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{M-1,M} \\ 0 & \cdots & 0 & 1 \end{bmatrix} \tag{4}$$

and $\mathbf{W}$ is an $N \times M$ matrix with orthogonal columns that satisfy

$$\mathbf{W}^T\mathbf{W} = \mathrm{diag}\{\kappa_1, \cdots, \kappa_M\} \tag{5}$$

with

$$\kappa_k = \mathbf{w}_k^T\mathbf{w}_k, \qquad k = 1, \ldots, M. \tag{6}$$

Equation (2) can be expressed as

$$\mathbf{y} = (\mathbf{P}\mathbf{A}^{-1})(\mathbf{A}\Theta) + \Xi = \mathbf{W}\Gamma + \Xi \tag{7}$$

where $\Gamma = [\gamma_1, \cdots, \gamma_M]^T$ is an auxiliary vector. It may be shown that the least squares solution minimizing the cost function of $J_{\mathrm{MSE}} = (1/N)\sum_{t=1}^{N}(y(t) - \sum_{k=1}^{M} w_k(\mathbf{x}(t))\gamma_k)^2$ is given by [6]

$$\gamma_k = \frac{\mathbf{w}_k^T\mathbf{y}}{\mathbf{w}_k^T\mathbf{w}_k}, \qquad k = 1, \ldots, M \tag{8}$$

where $w_k(\mathbf{x}(t))$ denotes the element of $\mathbf{W}$ at the tth row and kth column.

One way of implementing the above orthogonal decomposition is to use the modified Gram–Schmidt orthogonalization procedure (see the Appendix).

Based on the above, the OFR algorithms may be derived [6], [19], which select model terms one at a time in order to construct a subset model consisting of n_θ regressors, $n_\theta \ll M$, from the full model with regression matrix $\mathbf{P}$. The resultant regression matrix is denoted $\mathbf{P}_{n_\theta} \in \Re^{N \times n_\theta}$. A subset model can be achieved via a model term selective criterion, e.g., the minimization of J_{MSE}.

While J_{MSE} represents the model's approximation capability, the experimental design criteria focus on the model's adequacy and robustness [16], hence it is natural to consider model subset selection in the framework of the optimal experiment design. In optimal experimental design for model given by (2), $\mathbf{P}^T\mathbf{P}$ is referred to as the design matrix.

Consider the application of experimental design criteria in the context of model subset selection. We initially introduce the concept of A-optimality based on using a fixed sized subset model with size n_θ. The resultant regression matrix is still denoted as $\mathbf{P}_{n_\theta}$, and hence, the resultant design matrix is $[\mathbf{P}_{n_\theta}]^T\mathbf{P}_{n_\theta}$. Let λ_k, $k = 1, \ldots, n_\theta$, be the eigenvalues of $[\mathbf{P}_{n_\theta}]^T\mathbf{P}_{n_\theta}$.

Definition 1: The A-optimality criterion minimizes the sum of the variance of a parameter estimate vector $\hat{\mathbf{\Theta}} = [\theta_1, \cdots, \theta_{n_\theta}]^T$

$$\min\left\{J_1 = \mathrm{tr}[\mathrm{cov}\hat{\mathbf{\Theta}}] = \sigma^2\sum_{k=1}^{n_\theta}\frac{1}{\lambda_k}\right\}. \tag{9}$$

Unfortunately, the experimental design criterion of (9) is inherently computational inefficient if applied to model subset selection, due to the derivation of eigenvalues, and exponential growth of possible subsets. In our previous work [17], we aim to overcome this problem by initially introducing an alternative A-optimality based on orthogonal basis $\mathbf{w}_k$ rather than original regressor $\mathbf{p}_k$, followed by integrating this into the OFR framework. This is advantageous in that the computation efficiency in the conventional OFR algorithms is maintained. The basic idea in [17] is briefly explained below.

Note that (2) and (7) are just two alternative model representations. Similarly, an alternative A-optimality design criterion may be based on model (7) with orthogonal basis $\mathbf{w}_k$, rather than model (2). The A-optimality cost function proposed in [17] is described below. Let the subset regression matrix based on model (7) be $\mathbf{W}_{n_\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_{n_\theta}]$. Clearly, the resultant design matrix is $[\mathbf{W}_{n_\theta}]^T\mathbf{W}_{n_\theta}$, with eigenvalues as κ_k, $k = 1, \ldots, n_\theta$.

Definition 2: The A-optimality criterion minimizes the sum of the variance of the parameter estimate vector $\hat{\mathbf{\Gamma}} = [\gamma_1, \cdots, \gamma_{n_\theta}]^T$

$$\min\left\{J_A = \mathrm{tr}[\mathrm{cov}\hat{\mathbf{\Gamma}}] = \sigma^2\sum_{k=1}^{n_\theta}\frac{1}{\kappa_k}\right\}. \tag{10}$$

Although (9) and (10) are not exactly equivalent, it can be assumed that penalizing the large variance of the auxiliary parameter vector $\mathbf{\Gamma}$ also leads to penalizing the large variance of parameter vector $\mathbf{\Theta}$ because $\mathbf{A}\mathbf{\Theta} = \mathbf{\Gamma}$.

Taking into account both the $J_{\rm MSE}$ and the A-optimality objective as in Definition 2, a composite cost function can be defined as [17]

$$J = J_{\rm MSE} + \beta_1 J_A = \frac{1}{N}\left(\mathbf{y}^T\mathbf{y} - \sum_{k=1}^{n_\theta}\gamma_k^2\kappa_k\right) + \beta\sum_{k=1}^{n_\theta}\frac{1}{\kappa_k} \tag{11}$$

where β_1 is a predetermined small positive number, and $\beta = \sigma^2\beta_1$. Alternatively, (11) can be written as

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}\gamma_k^2\kappa_k + \frac{\beta}{\kappa_k} \tag{12}$$

with $J^{(0)} = \mathbf{y}^T\mathbf{y}/N$.

Based on (12), the A-optimality-based OFR selects the most relevant kth regressor at the kth forward regression stage [17]. At the kth forward regression stage, a candidate regressor is selected if it produces the smallest $J^{(k)}$ and provides further reduction on $J^{(k-1)}$.

The OFR algorithms are regarded as efficient model subset selection approaches. Considering the subset selection of choosing n_θ from M candidate terms and taking $M = 500$ and $n_\theta = 40$, there are $M!/n_\theta!(M-n_\theta)! = 2.2443\times 10^{59}$ possible model structures to select from. For the same $M = 500$ and $n_\theta = 40$ by OFR, the number of candidate model evaluation is reduced to $\sum_{k=1}^{n_\theta}(M-k+1) < n_\theta M = 2\times 10^4$. Despite this, it is still desirable to further reduce the computational cost, e.g., when M is very high.

III. New A-Optimality Orthogonal Forward Regression Algorithm Using Branch and Bound

A. The BB Based on the A-Optimality Composite Cost Function

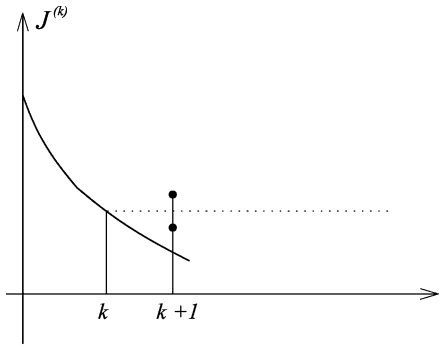
The BB technique consists of a systematic evaluation procedure for all candidate solutions by using the upper and lower estimated bounds of the quantity being optimized, such that large subsets of fruitless candidates are discarded. The branching refers to the procedure of successively dividing a candidate solution set into the subsets. The bounding refers to computing the upper and lower bounds for optimum value within a given subset. Suppose that the problem is to find the minimum of all candidate solutions, and the candidate set can be divided into two disjoint subsets $\mathcal{A}$ and $\mathcal{B}$. If the lower bound for the subset $\mathcal{A}$ is greater than the upper bound for $\mathcal{B}$, then $\mathcal{A}$ can be discarded. Alternatively, a bounding function could be based on the current best solution. If the lower bound for $\mathcal{A}$ is greater than the current best solution, it is discarded and the search space is reduced to $\mathcal{B}$.

Based on the BB principle, we propose an adaptive diagnostics test based upon the fact that the evolution of $J^{(k)}$, as a function of the forward regression step k, should be monotonically decreasing, as illustrated in Fig. 1. Specifically, at regression step k, the proposed test predicts whether a candidate regressor $\mathbf{p}_j$ would certainly increase $J^{(k)}$ if being included in the model for all subsequent regression steps (including and after the $(k+1)$th step). If this is true, then this regressor may be safely removed from $\mathcal{S}$.

Before proceeding to Theorem 1, we initially present some mathematical results so that these are readily usable for its proof.

Supposing at the kth forward regression step, a candidate regressor produces the smallest $J^{(k)}$, provides further reduction on $J^{(k-1)}$, and is selected with the resultant mean squares error of $J_{\rm MSE}^{(k)}$. Consider any other candidate regressor $\mathbf{p}_j$. The Gram–Schmidt orthogonalization procedure enables $\mathbf{p}_j$ to be orthogonal to $(k-1)$ orthogonal bases $\mathbf{w}_i, i = 1,\ldots,(k-1)$, in the current model of kth regression step, as

$$\mathbf{w}_{(-)} = \mathbf{p}_j - \sum_{i=1}^{k-1} a_{i,j}\mathbf{w}_i \tag{13}$$



Forward regression step

Fig. 1. A-optimality composite cost-function-based BB. The dotted horizontal line illustrates the decision boundary for regressors removal at the kth regression step. The circle dots illustrate the lower bound of the resultant A-optimality composite cost function from a candidate regressor, which can be removed if higher than the decision boundary.

with $a_{i,j} = (\mathbf{p}_j^T\mathbf{w}_i)/(\mathbf{w}_i^T\mathbf{w}_i)$. Furthermore, considering if this candidate regressor stays in the pool for the $(k+1)$th step forward regression, we have

$$\mathbf{w}_{(+)} = \mathbf{p}_j - \sum_{i=1}^{k} a_{i,j}\mathbf{w}_i \tag{14}$$

to enable it to be orthogonal to k orthogonal bases $\mathbf{w}_i$, $i = 1,\ldots,k$, in the model of $(k+1)$th regression step.

Substitute (13) into (14) to yield

$$\mathbf{w}_{(+)} = \mathbf{w}_{(-)} - a_{k,j}\mathbf{w}_k. \tag{15}$$

In summary, (13)–(15) describe the relationship between $\mathbf{w}_{(+)}$ and $\mathbf{w}_{(-)}$, which are based upon the same $\mathbf{p}_j$, but as a result of its being made orthogonal at two consecutive forward regression steps k and $(k+1)$. Furthermore, noting that the orthogonality condition $[\mathbf{w}_{(+)}]^T\mathbf{w}_k = 0$ holds, we have

$$a_{k,j} = \left(\mathbf{w}_k^T\mathbf{w}_{(-)}\right) / \left(\mathbf{w}_k^T\mathbf{w}_k\right). \tag{16}$$

To elaborate the motivation for the establishment of (15) and (16) (also for Theorem 1), we point out that the significance of a regressor $\mathbf{p}_j$ towards the model changes as the forward regression step k increases. For the BB principle to be applicable in the proposed algorithm, it is necessary to quantify the contribution of $\mathbf{p}_j$ towards the model as a function of k, as described in Theorem 1 and its proof. The use of (15) and (16) will become evident later on.

Theorem 1: The following diagnostic test is a feasible application of BB technique. If $\kappa_{(-)} = [\mathbf{w}_{(-)}]^T\mathbf{w}_{(-)} < \beta/J_{\rm MSE}^{(k)}$, then $\mathbf{p}_j$ is eliminated from the pool before the $(k+1)$th forward regression step.

Proof: Because our objective is to minimize the A-optimality composite cost function as given by (12), it is possible to determine a subset of infeasible candidate regressors, which would produce solutions worse off than the current solution using the BB technique. These candidate regressors can be eliminated from $\mathcal{S}$. Assuming $\mathbf{p}_j$ is included into the model at the $(k+1)$th step, let the resultant A-optimality composite cost function be $J^{(k+1,+)}$, consisting of the mean squares error $J_{\rm MSE}^{(k+1,+)}$ and the A-optimality objective $J_A^{(k+1,+)}$. Similar to (11) and (12), we have

$$\begin{aligned} J^{(k+1,+)} &= J_{\rm MSE}^{(k+1,+)} + \beta_1 J_A^{(k+1,+)} \\ &= J_{\rm MSE}^{(k)} - \frac{1}{N}\gamma_{(+)}^2\kappa_{(+)} + \beta_1 J_A^{(k)} + \frac{\beta}{\kappa_{(+)}} \\ &= J^{(k)} - \frac{1}{N}\gamma_{(+)}^2\kappa_{(+)} + \frac{\beta}{\kappa_{(+)}} \end{aligned} \tag{17}$$

where $\kappa_{(+)} = [\mathbf{w}_{(+)}]^T \mathbf{w}_{(+)}$ and $\gamma_{(+)} = \mathbf{w}_{(+)}^T \mathbf{y} / \mathbf{w}_{(+)}^T \mathbf{w}_{(+)}$. The reduction of the A-optimality composite cost function due to adding $\mathbf{p}_j$ to the model is then given by $(J^{(k)} - J^{(k+1,+)})$.

In the following, an upper bound of $(J^{(k)} - J^{(k+1,+)})$ is derived, which is the difference between the upper bound of the term $(1/N)\gamma_{(+)}^2 \kappa_{(+)}$ and the lower bound of term $\beta/\kappa_{(+)}$ in (17). Specifically, we suggest a choice for these two bounds to be used in the proposed algorithm in the points i) and ii).

We note the following.

i) Clearly, $J_{\mathrm{MSE}}^{(k+1,+)} > 0$ such that we have $(1/N)\gamma_{(+)}^2 \kappa_{(+)} < J_{\mathrm{MSE}}^{(k)}$.

ii) Making use of (15) and (16), we have

$$\begin{aligned} \kappa_{(+)} &= \left[\mathbf{w}_{(+)}\right]^T \mathbf{w}_{(+)} \\ &= \left[\mathbf{w}_{(-)}\right]^T \mathbf{w}_{(-)} - 2a_{k,j}\mathbf{w}_k^T \mathbf{w}_{(-)} + a_{k,j}^2 \mathbf{w}_k^T \mathbf{w}_k \\ &= \left[\mathbf{w}_{(-)}\right]^T \mathbf{w}_{(-)} - a_{k,j}^2 \mathbf{w}_k^T \mathbf{w}_k \\ &< \left[\mathbf{w}_{(-)}\right]^T \mathbf{w}_{(-)} = \kappa_{(-)}. \end{aligned} \tag{18}$$

Thus

$$\frac{\beta}{\kappa_{(+)}} > \frac{\beta}{\kappa_{(-)}}. \tag{19}$$

From i) and ii), we see that the reduction of the A-optimality composite cost function due to adding $\mathbf{p}_j$ to the model is upper bounded by $[J_{\mathrm{MSE}}^{(k)} - (\beta/\kappa_{(-)})]$.

To find the subset of infeasible candidate regressors, we set the upper bound of $(J^{(k)} - J^{(k+1,+)})$ to be less than zero, i.e., a negative reduction of the A-optimality composite cost function, yielding $\kappa_{(-)} < \beta/J_{\mathrm{MSE}}^{(k)}$.

Finally, it is straightforward to verify by induction that if any regressor is eligible to be eliminated from the pool at the kth regression step, but is kept in the pool, then this is also eligible to be eliminated from the pool at any of all future regression steps. This concludes the proof of Theorem 1.

We point out that (13)–(19) are not for the real implementation, but for analysis only. Particularly, note that any regressors satisfying $\kappa_{(-)} < \beta/J_{\mathrm{MSE}}^{(k)}$ are eliminated immediately after the kth regression step for $k \geq 2$, such that for these regressors, no real $(k+1)$th step orthogonalization are implemented, resulting in significant reduction in computational cost.

B. The Algorithm

Combining the BB technique based on the A-optimality composite cost function with the modified Gram–Schmidt orthogonalization procedure (see the Appendix), an efficient algorithm for selecting a subset model is derived as below. Let $M_k \leq (M - k + 1)$ denote the number of the candidate regressors in the pool $\mathcal{S}$ at the kth regression stage. Define

$$\mathbf{P}^{(k-1)} = \left[\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}, \underbrace{\mathbf{p}_k^{(k-1)}, \ldots, \mathbf{p}_{M_k+k-1}^{(k-1)}}_{\mathcal{S}}, \underbrace{\cdots, \cdots, \cdots}_{\text{removed regressors}}\right] \in \Re^{N \times M}. \tag{20}$$

If some of the columns in $\mathbf{P}^{(k-1)}$ have been interchanged, this will still be referred to as $\mathbf{P}^{(k-1)}$ for notational simplicity. The kth stage of selection procedure is given as follows.

Step 1) For $k \leq j \leq M_k + k - 1$, compute

$$\left.\begin{aligned} \gamma_k^{(j)} &= \left[\mathbf{p}_j^{(k-1)}\right]^T \mathbf{y}^{(k-1)} / \left(\left[\mathbf{p}_j^{(k-1)}\right]^T \mathbf{p}_j^{(k-1)}\right) \\ J^{(k,j)} &= J^{(k-1)} - \left(\gamma_k^{(j)}\right)^2 \left(\left[\mathbf{p}_j^{(k-1)}\right]^T \mathbf{p}_j^{(k-1)}\right)/N \\ &\quad + \beta / \left(\left[\mathbf{p}_j^{(k-1)}\right]^T \mathbf{p}_j^{(k-1)}\right) \end{aligned}\right\}.$$

Step 2) Find

$$J^{(k)} = J^{(k,j_k)} = \min\left\{J^{(k,j)}, k \leq j \leq M_k + k - 1\right\}. \tag{21}$$

Then, the j_kth and the kth column of $\mathbf{P}^{(k-1)}$ are interchanged. The j_kth column and the kth column of $\mathbf{A}$ are interchanged up to the $(k-1)$th row. This effectively selects the kth regressor in the subset model.

Step 3) Set $\mathbf{w}_k = \mathbf{p}_k^{(k-1)}$. Calculate γ_k and update $\mathbf{y}^{(k-1)}$ into $\mathbf{y}^{(k)}$ according to (25) of the modified Gram–Schmidt orthogonalization procedure shown in the Appendix. Update

$$\begin{aligned} J_{\mathrm{MSE}}^{(k)} &= J_{\mathrm{MSE}}^{(k-1)} - \gamma_k^2 \mathbf{w}_k^T \mathbf{w}_k / N \\ \alpha(k) &= \beta / J_{\mathrm{MSE}}^{(k)}. \end{aligned} \tag{22}$$

Update

$$\mathcal{S} = \arg\left\{\left[\mathbf{p}_j^{(k-1)}\right]^T \mathbf{p}_j^{(k-1)} > \alpha(k), k+1 \leq j \leq M_k + k - 1\right\} \tag{23}$$

and M_{k+1} (the reduced size of $\mathcal{S}$). According to $\mathcal{S}$, update $\mathbf{P}^{(k)}$ and $\mathbf{A}$ in the same column order such that the regressors in $\mathcal{S}$ are placed from $(k+1)$th to $(M_{k+1} + k)$th row.

Step 4) Perform (24) of the modified Gram–Schmidt orthogonalization procedure shown in the Appendix, but *only up to* $(M_{k+1} + k)$th *column*. That is, for the set $\mathcal{S}$, to derive the kth row of $\mathbf{A}$, transform $\mathbf{P}^{(k-1)}$ into $\mathbf{P}^{(k)}$. This procedure is terminated at the $(n_\theta + 1)$th stage when $J^{(n_\theta+1)} > J^{(n_\theta)}$ is detected and this produces a subset model with n_θ significant regressors.

For both A-optimality-based OFR [17] and the proposed algorithm above, the computational complexity for each evaluation of a candidate regressor is in the order of $O(N)$. Therefore, the computational saving offered by the proposed algorithm can be indicated by the total number of regressors evaluation $(\sum_{k=1}^{n_\theta} m_k)$ in comparison with the conventional A-optimality-based OFR $(\sum_{k=1}^{n_\theta}(M - k + 1))$, with $m_k - m_{k+1} > 1$. Although the rate of m_k with k depends on the data itself, it is clear that the proposed algorithm offers opportunities to significantly reduce the computation cost of the OFR algorithms. For illustration, assuming $M \gg 1$, m_k is reduced at a constant rate to 1, and the final model size is $n_\theta = 5\%M$, the computational cost of the proposed algorithm is only 52% of the A-optimality-based OFR [17]. In practice, as found in the simulations, the reduction rate of m_k is small at small k, but increases with k, and there is likely about 20%∼40% saving of the computational cost.

IV. Modeling Examples

Example 1: The relationship between the fuel rack position [input $u(t)$] and the engine speed [input $y(t)$] is modeled for a Leyland TL11 turbocharged, direct injection diesel engine that is operated at a low engine speed. Detailed system description and experimental setup can be found in [20]. The data set, depicted in Fig. 2, contains 410 samples. The first 210 data samples were used in training and the last 200 data samples for model validation. The previous study has shown that the data set can be modeled adequately using the system input vector $\mathbf{x}(t) = [y(t-1), u(t-1), u(t-2)]^T$. The best Gaussian kernel model was provided by the locally regularized orthogonal least squares (LROLS) algorithm with the leave-one-out (LOO) test score, consisting of 22 terms [21] and with the mean square error (MSE) values over the training and validation data sets of 0.000453 and 0.000490, respectively.

We use the Gaussian RBF $p_k(\mathbf{x}(t)) = \exp\{-(\|\mathbf{x}(t) - \mathbf{c}_k\|^2/2\tau^2)\}$ to construct our model using the proposed algorithm, where $\tau = 2.5$ was set empirically. $\mathbf{c}_k$ was formed using all the training data samples.

TABLE I
COMPARISON OF MODELING PERFORMANCE FOR ENGINE DATA SET

A-optimality Weighting β	MSE over training data A-OFR	MSE over training data A-OFR +BB	MSE over testing data A-OFR	MSE over testing data A-OFR + BB	Model size A-OFR	Model size A-OFR + BB	A-OFR +BB computing cost reduction
10^{-11}	0.000479	0.000496	0.000499	0.000518	16	16	23.15%
10^{-12}	0.000489	0.000489	0.000485	0.000484	18	17	27.3%
10^{-13}	0.000479	0.000482	0.000492	0.000492	20	18	32.4%
10^{-14}	0.000467	0.000484	0.000495	0.000487	23	19	39.2%

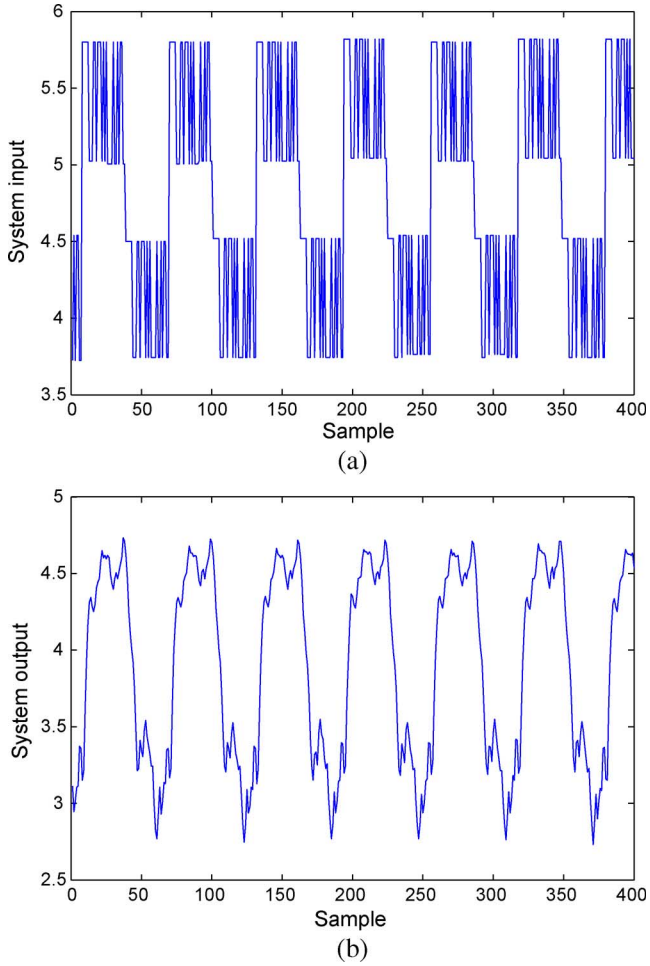


Fig. 2. Engine data set. (a) System input $u(t)$. (b) System output $y(t)$.

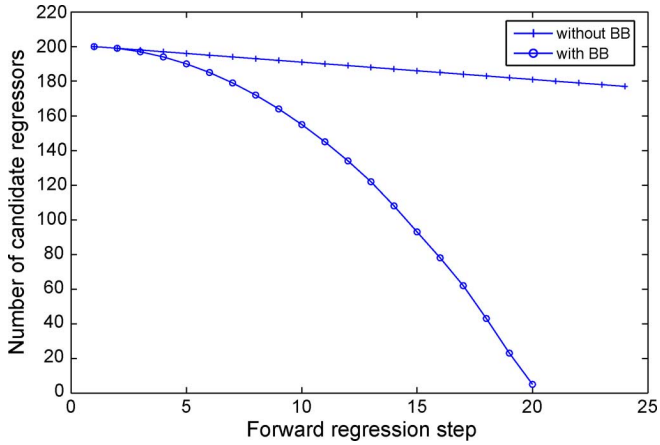


Fig. 3. Size of $\mathcal{S}$ as a function of the forward regression step.

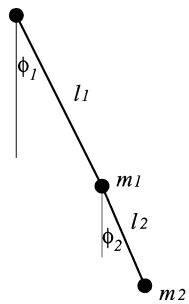


Fig. 4. Double pendulum.

The modeling performance of the proposed algorithm (A-OFR+BB) is shown in Table I in comparison with the A-optimality-based OFR without BB applied (A-OFR). Clearly the modeling accuracy of the models are comparable to that of [21]. The main computational cost reduction is indicated by the total number of regressors evaluation ($\sum_{k=1}^{n_\theta} m_k$) in comparison with the conventional A-optimality-based OFR ($\sum_{k=1}^{n_\theta}(M-k+1)$). The evolution of m_k in the case of $\beta = 10^{-14}$ is shown in Fig. 3 in order to demonstrate the faster reduction of the search space due to the proposed application of the BB technique. Finally, we note that the proposed A-OFR+BB does not yield to the exact model as that of A-OFR. We found in simulations that the cause is due to fact that a tie may happen in selecting the regressor producing the minimal $J^{(k)}$ at some k, such that different regressors are selected. Note that it is also possible to modify the A-OFR+BB so that it produces the exact model as of A-OFR, e.g., via carefully preserving the order of regressors in $\mathcal{S}$. However, this procedure is generally unnecessary because: 1) this would incur additional computational cost, and more importantly, 2) either algorithm does not necessarily produce model superior to the other. Nevertheless, we point out that for any forward regression algorithm, how to deal with a tie and its implications in model selection pose an interesting open problems, especially, if there are multiple objectives or other requirements involved, e.g., possibly those from the application domain.

Example 2 (Nonlinear Time Series): The motion equations of a double pendulum system, as shown in Fig. 4, are given by

$$\begin{aligned}
\dot{\phi}_1 &= \omega_1 \\
\dot{\omega}_1 &= \left\{m_2 l_1 \omega_1^2 \sin(\varphi)\cos(\varphi) + m_2 g \sin(\phi_2)\cos(\varphi) \right. \\
&\qquad \left. + m_2 l_2 \omega_2^2 \sin(\varphi) - (m_1+m_2) g \sin(\phi_1)\right\} \\
&\quad / \left\{(m_1+m_2) l_1 - m_2 l_1 \cos^2(\varphi)\right\} \\
\dot{\phi}_2 &= \omega_2 \\
\dot{\omega}_2 &= \left\{-m_2 l_2 \omega_2^2 \sin(\varphi)\cos(\varphi) + (m_1+m_2) \right. \\
&\qquad \left. \times \left[g \sin(\phi_1)\cos(\varphi) - l_1 \omega_1^2 \sin(\varphi) - g \sin(\phi_2)\right]\right\} \\
&\quad / \left\{(m_1+m_2) l_2 - m_2 l_2 \cos^2(\varphi)\right\}
\end{aligned}$$

TABLE II
COMPARISON OF MODELING PERFORMANCE FOR THE LOWER PENDULUM ANGLE ϕ_2

A-optimality Weighting β	MSE over training data A-OFR	MSE over training data A-OFR +BB	MSE over testing data A-OFR	MSE over testing data A-OFR + BB	Model size A-OFR	Model size A-OFR + BB	A-OFR +BB computing cost reduction
10^{-11}	0.000127	0.000176	0.000316	0.000515	31	29	23.02%
10^{-12}	0.000081	0.000088	0.000196	0.000174	33	35	20.0%
10^{-13}	0.000062	0.000078	0.000163	0.000262	42	38	35.1%
10^{-14}	0.000046	0.000061	0.000176	0.000162	48	39	42.8%

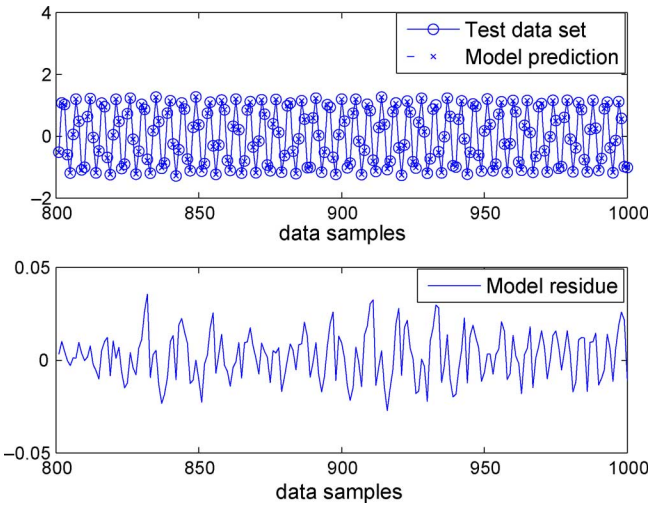


Fig. 5. Test data set of the lower pendulum angle ϕ_2.

where m_i, l_i, and ϕ_i denote the masses, lengths, and angles from the vertical of the upper ($i = 1$) and lower ($i = 2$) pendulum. $\varphi = \phi_2 - \phi_1$. The solutions can be obtained using numerical integration. The parameters were set as $m_1 = 700g$, $m_2 = 200g$, $l_1 = 1.2m$, and $l_2 = 0.2m$. With the initial condition of $[\phi_1(0), \omega_1(0), \phi_2(0), \omega_2(0)]^T$ as $[0, 0, \pi/2, 0]^T$, an integration time span of 200 s at a sampling rate of 0.2 s, we generated 1000 data points of four sequences of nonlinear time-series data set for ϕ_1, ω_1, ϕ_2, and ω_2. Consider the modeling of $\phi_2(t)$ as a nonlinear time series, with the system input vector $\mathbf{x}(t) = [\phi_2(t-1), \phi_2(t-1), \ldots, \phi_2(t-n_y)]^T$. $n_y = 6$. The first 800 data samples were used in training and the last 200 data samples for model validation.

The Gaussian RBF $p_k(\mathbf{x}(t)) = \exp\{-(\|\mathbf{x}(t) - \mathbf{c}_k\|^2/2\tau^2)\}$ to construct our model using the proposed algorithm, where $\tau = 3$ was set empirically. $\mathbf{c}_k$ was formed using all 800 training data samples. The modeling performance of the proposed algorithm (A-OFR+BB) is shown in Table II in comparison with the A-optimality-based OFR without BB applied (A-OFR). Similar to Example 1, we note that the proposed A-OFR+BB does not yield to the exact model as that of A-OFR, and the cause is due to fact that a tie may happen in selecting the regressors. In terms of the modeling errors, both methods yield comparable results. The modeling results of a 39 centers RBF model obtained using the proposed algorithm with $\beta = 10^{-14}$ is shown in Fig. 5. The main computational cost reduction is indicated by the total number of regressors evaluation in comparison with the conventional A-optimality-based OFR, which shows a significant amount of saving. Finally, we clarify that the amount of saving indicated in Tables I and II is a comparison based on the hidden nodes selection stage only, without taking into account the calculation load of earlier stages, e.g., input selection for high input dimension data set, or the formation of regression matrix $\mathbf{P}$. It is reasonable to assume that the same procedure is applied for both A-OFR and A-OFR+BB such that the same amount of extra computational cost should be added to obtain the computation cost of the complete identification algorithm. Consequently, the A-OFR+BB still provides a certain amount of computational saving, if not significant in the rare case that the computational cost in earlier stage is dominant.

V. CONCLUSION

In this brief, we have introduced a new A-optimality-based OFR algorithm by integrating the BB technique. The proposed algorithm can reduce the search efforts in the A-optimality-based OFR algorithm significantly. A new diagnostics test is proposed to reduce the size of the pool of candidate regressors at each regression step, and this is proven to be an application of the BB technique. Numerical examples are used to demonstrate the effectiveness of the proposed algorithm.

APPENDIX
THE LEAST SQUARES ALGORITHM USING THE MODIFIED GRAM–SCHMIDT ORTHOGONALIZATION PROCEDURE

The modified Gram–Schmidt orthogonalization procedure calculates $\mathbf{A}$ matrix row by row and orthogonalizes $\mathbf{P}$ as follows: at the kth stage, the columns $\mathbf{p}_j$ are made orthogonal to the kth column. The procedure is repeated for $1 \le k \le M-1$. Specifically, denoting $\mathbf{p}_j^{(0)} = \mathbf{p}_j$, $j = 1, \ldots, M$, then

$$\left.\begin{aligned} \mathbf{w}_k &= \mathbf{p}_k^{(k-1)} \\ a_{k,j} &= \mathbf{w}_k^T \mathbf{p}_j^{(k-1)} / \mathbf{w}_k^T \mathbf{w}_k, \qquad k+1 \le j \le M \\ \mathbf{p}_j^{(k)} &= \mathbf{p}_j^{(k-1)} - a_{k,j}\mathbf{w}_k, \qquad k+1 \le j \le M \end{aligned}\right\}, \quad 1 \le k \le M-1. \tag{24}$$

The last stage of the procedure is simply $\mathbf{w}_M = \mathbf{p}_M^{(M-1)}$. The elements of Γ are computed by transforming $\mathbf{y}^{(0)} = \mathbf{y}$ in a similar way

$$\left.\begin{aligned} \gamma_k &= \mathbf{w}_k^T \mathbf{y}^{(k-1)} / \mathbf{w}_k^T \mathbf{w}_k \\ \mathbf{y}^{(k)} &= \mathbf{y}^{(k-1)} - \gamma_k \mathbf{w}_k, \end{aligned}\right\}, \qquad 1 \le k \le M. \tag{25}$$

REFERENCES

[1] C. J. Harris, X. Hong, and Q. Gan, *Adaptive Modeling, Estimation and Fusion from Data: A Neurofuzzy Approach*. New York: Springer-Verlag, 2002.
[2] M. Brown and C. J. Harris, *Neurofuzzy Adaptive Modeling and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1994.
[3] A. E. Ruano, *Intelligent Control Systems using Computational Intelligence Techniques*. London, U.K.: IEE, 2005.
[4] R. Murray-Smith and T. A. Johansen, *Multiple Model Approaches to Modeling and Control*. New York: Taylor & Francis, 1997.
[5] S. G. Fabri and V. Kadirkamanathan, *Functional Adaptive Control: An Intelligent Systems Approach*. New York: Springer-Verlag, 2001.
[6] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their applications to non-linear system identification," *Int. J. Control*, vol. 50, pp. 1873–1896, 1989.

[7] M. J. Korenberg, "Identifying nonlinear difference equation and functional expansion representations: the fast orthogonal algorithm," *Ann. Biomed. Eng.*, vol. 16, pp. 123–142, 1988.
[8] L. Wang and J. M. Mendel, "Fuzzy basis functions, universal approximation, and orthogonal least-squares learning," *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 807–814, Sep. 1992.
[9] X. Hong and C. J. Harris, "Neurofuzzy design and model construction of nonlinear dynamical processes from data," *Inst. Electr. Eng. Proc. —Control Theory Appl.*, vol. 148, no. 6, pp. 530–538, 2001.
[10] Q. Zhang, "Using wavelets network in nonparametric estimation," *IEEE Trans. Neural Netw.*, vol. 8, no. 2, pp. 227–236, Mar. 1997.
[11] S. A. Billings and H. L. Wei, "The wavelet-narmax representation: A hybrid model structure combining polynomial models with multiresolution wavelet decompositions," *Int. J. Syst. Sci.*, vol. 36, no. 3, pp. 137–152, 2005.
[12] N. Chiras, C. Evans, and D. Rees, "Nonlinear gas turbine modeling using narmax structures," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 4, pp. 893–898, Aug. 2001.
[13] Y. Gao and M. J. Er, "Online adaptive fuzzy neural identification and control of a class of MIMO nonliear systems," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 4, pp. 462–477, Aug. 2003.
[14] K. M. Tsang and W. L. Chan, "Adaptive control of power factor correction converter using nonlinear system identification," *Inst. Electr. Eng. Proc.—Electric Power Appl.*, vol. 152, no. 3, pp. 627–633, 2005.
[15] G. C. Luh and W. C. Cheng, "Identification of immune models for fault detection," *Proc. Inst. Mech. Eng. Part I: J. Syst. Control Eng.*, vol. 218, pp. 353–367, 2004.
[16] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*. Oxford, U.K.: Clarendon, 1992.
[17] X. Hong and C. J Harris, "Nonlinear model structure detection using optimum experimental design and orthogonal least squares," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 435–439, Mar. 2001.
[18] M. Brusco and S. Stahl, *Branch-and-Bound Applications in Combinatorial Data Analysis*. New York: Springer-Verlag, 2005.
[19] S. Chen, C. F. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Netw.*, vol. 2, no. 2, pp. 302–309, Mar. 1991.
[20] S. A. Billings, S. Chen, and R. J. Backhouse, "The identification of linear and nonlinear models of a turbocharged automotive diesel engine," *Mech. Syst. Signal Process.*, vol. 3, no. 2, pp. 123–142, 1989.
[21] S. Chen, X. Hong, C. J. Harris, and P. M. Sharkey, "Sparse modeling using orthogonal forward regression with PRESS statistic and regularization," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 34, no. 2, pp. 898–911, Apr. 2004.

Consensus in Networks of Multiagents With Cooperation and Competition Via Stochastically Switching Topologies

Bo Liu and Tianping Chen

***Abstract*—In this brief, we provide some theoretical analysis of the consensus for networks of agents via stochastically switching topologies. We consider both discrete-time case and continuous-time case. The main contribution of this brief is that the underlying graph topology is more general in both cases than those appeared in previous papers. The weight matrix of the coupling graph is not assumed to be nonnegative or Metzler. That is, in the model discussed here, the off-diagonal entries of the weight matrix of the coupling graph may be negative. This means that sometimes, the coupling may not benefit, but may prevent the consensus of the coupled agents. In the continuous-time case, the switching time intervals also take a more general form of random variables than those appeared in previous works. We focus our study on such networks and give sufficient conditions that ensure almost sure consensus in both discrete-time case and continuous-time case. As applications, we give several corollaries under more specific assumptions, i.e., the switching can be some independent and identically distributed (i.i.d.) random variable series or a Markov chain. Numerical examples are also provided in both discrete-time and continuous-time cases to demonstrate the validity of our theoretical results.**

***Index Terms*—Almost sure, consensus, stochastic, switching topology.**

I. Introduction

In a network of dynamical agents, groups of agents need to agree upon certain quantity of interest in order to realize coordination among them, which is the so-called "*consensus problem*." Consensus problems often arise in the applications of multiagent systems [1]–[4] and have received much attention in recent years. There is a large amount of papers concerning such problems (see [5]–[15], [17], [19] and references therein).

To achieve consensus, there should be some information flow from agent to agent, which may be directed or undirected. The agents with information flow can be described by a graph topology. The topology may be static, which means that it dose not change along with time. However, in many cases, it may dynamically change, which is often resulted from unreliable transmission or limited communication/sensing range. One of the important classes of dynamically changing network topologies is the so-called "*switching topoldogy*," where the network topology switches at a sequence of time points, randomly or controlled by a given rule. Consensus problems with switching topologies have been addressed in several papers such as [7], [9], [15], [17], [19], and others.

The weighted directed graph is an important class in modeling the network topology, where a directed information flow is modeled as a directed edge. When the information flow plays positive role to consensus between the agents, the corresponding edge is assigned a positive weight. Otherwise, it is assigned a negative weight. In real world, it is possible that there exists a positive or a negative role among agents to achieve consensus. This results in a graph topology with arbitrary weighted edges. Therefore, it is meaningful to investigate consensus problems for such network topologies in both theories and applications.

Manuscript received April 23, 2008; revised July 15, 2008; accepted August 5, 2008. First published September 30, 2008; current version published November 5, 2008. This work was supported by the National Science Foundation (NSF) of China under Grants 60574044 and 60774074.

The authors are with Key Laboratory of Nonlinear Science of Chinese Ministry of Education, Institute of Mathematics, Fudan University, Shanghai 200433, P. R. China (e-mail: tchen@fudan.edu.cn; 071018024@fudan.edu.cn).

Color versions of one or more of the figures in this brief are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2008.2004404