Copyright © 2002 IFAC
15th Triennial World Congress, Barcelona, Spain

A LEARNING THEORY APPROACH TO SYSTEM IDENTIFICATION

M. Vidyasagar[1] **Rajeeva L. Karandikar**[2]

[1] Advanced Technology Centre, Tata Consultancy Services
Khan Lateefkhan Estate, Fateh Maidan Road, Hyderabad 500 001, INDIA
sagar@atc.tcs.co.in

[2] Indian Statistical Institute, S. J. S. Sansawal Marg, New Delhi 110 016, INDIA
rlk@isid.ac.in

Abstract In this paper, we present a new approach to system identifiction and stochastic adaptive control, by viewing these as problems in statistical learning theory. This approach leads to *finite time* estimates for the distance between the system being identified and the unknown system. This approach permits one to combine system identification with robust control. learning theory. As an illustration of the approach, a result is derived showing that in the case of systems with fading memory, it is possible to combine standard results in statistical learning theory (suitably modified to the present situation) with some fading memory arguments to obtain finite time estimates of the desired kind. Though the actual results derived here are rather restricted in scope, it is hoped that future researchers will pursue the ideas presented here to extend the theory further.

Key words: Identification, learning algorithms, stochastic control

1 Introduction

The aim of system identification is to fit given data, usually supplied in the form of a time series, with models from within a given model class. One can divide the main challenges of system identificdation into three successively stronger questions, as follows: As more and more data is provided to the identification algorithm,

1. Does the estimation error between the outputs of the identified model and the actual time series approach the minimum possible estimation error achievable by any model within the given model class?
2. Does the identified model converge to the best possible model within the given model class?
3. Assuming that the data is generated by a 'true' model whose output is corrupted by measurement noise, does the identified model converge to the 'true' model?

From a technical standpoint, Questions 2 and 3 are easier to answer than Question 1. Following the notational conventions of system identifiction, let $\{h(\theta), \theta \in \Theta\}$ denote the family of models, where θ denotes a parameter that characterizes the model, and Θ is a topological space (usually a subset of $\mathbb{R}^\ell$ for some ℓ). Since identification is carried out recursively, the output of the identification algorithm is a sequence of estimates $\{\theta_t\}_{t\geq 1}$, or what is the same thing, a sequence of estimated models $\{h(\theta_t)\}_{t\geq 1}$. Traditionally a positive answer to Question 2 is assured by assuming that Θ is a *compact* set, which in turn ensures that the sequence $\{\theta_t)\}$ contains a convergent subsequence. If the answer to Question 1 is 'yes,' and if θ^* is a limit point of the sequence, it is usually not difficult to establish that the model $h(\theta^*)$ is an 'optimal' fit to the data among the family $\{h(\theta), \theta \in \Theta\}$. Coming now to Question 3, suppose θ_{true} is the parameter of the 'true' model, and let f_{true} denote the 'true' system. Suppose θ^* is a limit point of the sequence $\{\theta_t\}$. The traditional way to ensure that $\theta_{\text{true}} = \theta^*$ is to assume that the input to the true system is 'persistingly exciting' or 'sufficiently rich,' so that the only way for $h(\theta^*)$ to match the performance of f_{true} is to have $\theta^* = \theta_{\text{true}}$.

With this background, the present paper concentrates on providing an affirmative answer to Question 1. In a seminal paper [10], Lennart Ljung has shown that indeed Question 1 can be answered in the affirmative provided empirical estimates of the performance of each model $h(\theta)$ converge *uniformly* to the corresponding true performance, where the uniformity is with respect to $\theta \in \Theta$. Very closely related results are proven by Caines [3]. Ljung also showed that this particular uniform convergence property does hold, provided two assumptions are satisfied, namely:

- The model class consists of uniformly exponentially stable systems, and
- The parameter θ enters the description of the model $h(\theta)$ in a 'differentiable' manner. Coupled with the assumption that Θ is a compact set, this assumption implies that various quantities have bounded gradients with respect to θ.

The uniform convergence property in question is referred to hereafter as UCEM (uniform convergence of empirical means). A precise definition of the UCEM property, as well as a rationale for its name, is given in subsequent sections.

Now it turns out that a study of the UCEM property in various forms lies at the heart of a branch of applied probability theory, variously known as empirical process theory or statistical learning theory. One of the distinguishing features of statistical learning theory is its emphasis on *finite time estimates*. This is in contrast to the *asymptotic results* provided by nearby branches of probability theory such as large deviation theory. Note that the main results of system identification theory of relevance to the present discussion, such as [10], Lemma 3.1, or [11], Theorem 2B.3, are also asymptotic. Actually, the proofs of these results can in fact provide finite time estimates. However, these estimates are not very tight, possibly because by tradition the emphasis in system identification theory has not been on deriving finite time estimates.

This brings us to the motivation of the present paper, which is to apply the techniques of statistical learning theory (if not exactly the actual results from that theory) to the problem of system identification. The results presented here are among the first attempts at applying statistical learning theory to the long-standing problem of system identification. See [14, 15, 4, 16] for related results. Undoubtedly it is possible to improve both the results themselves and also the proofs of the results. It is the hope of the authors that the paper will spur further research in the subject.

2 Problem Formulation

2.1 Preliminaries

For the class of systems under study, the output set is some $Y \subseteq \mathbb{R}^k$, while the input set is some $U \subseteq \mathbb{R}^\ell$ for some and ℓ. To avoid technicalities, let us suppose that the inputs are restricted to belong to a *bounded* set U; this assumption ensures that any random variable assuming values in U has bounded moments of all orders. There is also a "loss function" $\ell : Y \times Y \to [0, 1]$.

To set up the time series that forms the input to identification or stochastic adaptive control, let us first define $\mathcal{U} := \prod_{-\infty}^{\infty} U$, and define $\mathcal{Y}$ analogously. Equip the doubly infinite cartesian product $\mathcal{Y} \times \mathcal{U} := \prod_{-\infty}^{\infty} (Y \times U)$ with the product Borel -algebra, and call it $\mathcal{S}^\infty$. Next, introduce a probability measure $\tilde{P}_{\mathbf{y},\mathbf{u}}$ on the measurable space $(\mathcal{Y} \times \mathcal{U}, \mathcal{S}^\infty)$. Now let us define a 'stochastic process' as a measurable map from $(\mathcal{Y} \times \mathcal{U}, \mathcal{S}^\infty, \tilde{P}_{\mathbf{y},\mathbf{u}})$ into $\mathcal{Y} \times \mathcal{U}$. Let the coordinate random variables (y_t, u_t) be thought of as the components of the time series at time t, and let us assume that the time series is stationary (which means that the probability measure $\tilde{P}_{\mathbf{y},\mathbf{u}}$ is shift-invariant). Let $\tilde{P}_{y,\mathbf{u}}$ denote the one-dimensional marginal probability associated with $\tilde{P}_{\mathbf{y},\mathbf{u}}$ on Y, and note that $\tilde{P}_{y,\mathbf{u}}$ is a probability measure on the set $Y \times \mathcal{U}$. Let $U_{-\infty}^0$ denote the one-sided infinite cartesian product $U_{-\infty}^0 := \prod_{-\infty}^0 U$, and for a given two-sided infinite sequence $\mathbf{u} \in \mathcal{U}$, define

$$\mathbf{u}_t := (u_{t-1}, u_{t-2}, u_{t-3}, \ldots) \in U_{-\infty}^0.$$

With this preliminary notation, we can set up the problem under study.

2.2 System Identification

Let us begin with the problem of system identification, as the stochastic adaptive control problem is a ready modification thereof. The input to the identification process is a time series $\{(y_t, u_t)\}_{t \geq 1}$ generated through a stochastic process, as described above. To fit this time series, we use a family of models $\{h(\theta), \theta \in \Theta\}$, where each $h(\theta)$ denotes an input-output mapping from $U_{-\infty}^{0}$ to Y, and the parameter θ captures the variations in the model family. Thus the output at time t of the system parametrized by θ to the input sequence $\mathbf{u} \in \mathcal{U}$ is given by $h(\theta) \cdot \mathbf{u}_t$. Note that this definition automatically guarantees that each system is time-invariant.

For each parameter $\theta \in \Theta$, define the objective function

$$J(\theta) := E[\ell(y_t, h(\theta) \cdot \mathbf{u}_t), \tilde{P}_{\mathbf{y},\mathbf{u}}].$$

Thus $J(\theta)$ is the expected value of the loss we incur by using the model output $h(\theta) \cdot \mathbf{u}_t$ to predict the actual output y_t. Note that, since the only value of $\mathbf{y}$ that appears within the expected value is y_t, we can actually replace the measure $\tilde{P}_{\mathbf{y},\mathbf{u}}$ by $\tilde{P}_{y,\mathbf{u}}$. In other words, we can also write

$$J(\theta) := E[\ell(y_t, h(\theta) \cdot \mathbf{u}_t), \tilde{P}_{y,\mathbf{u}}]. \qquad (2.1)$$

Thus the expectation is taken with respect to the 'one-dimensional' marginal measure $\tilde{P}_{y,\mathbf{u}}$ on $Y \times \mathcal{U}$. One of the most commonly used loss functions is the squared error; thus

$$\ell(y, z) :=\| y - z \|^2,$$

where $|| \cdot ||$ is the usual Euclidean or ℓ_2-norm. In this case $J(\theta)$ is the expected value of the mean squared prediction error when the map $h(\theta)$ is used to predict y_t. Note that, by the assumption of stationarity, the quantity on the right side of (2.1) is independent of t. The objective of identification is to determine a $\theta \in \Theta$ that minimizes the error measure $J(\theta)$.

Suppose the measured output y_t corresponds to a noise-corrupted output of a 'true' system f_{true}, and that ℓ is the squared error, as above. In such a case, the problem formulation becomes the following: Suppose the input sequence $\{u_t\}_{-\infty}^{\infty}$ is i.i.d. according to some law P, and that $\{\eta_t\}_{-\infty}^{\infty}$ is a measurement noise sequence that is zero mean and i.i.d. with law Q. Suppose in addition that u_i, η_j are independent for each i, j. Now suppose that

$$y_t = f_{\text{true}} \cdot \mathbf{u}_t + \eta_t, \ \forall t. \qquad (2.2)$$

In such a case, the expected value in (2.1) can be expressed in terms of the probability measure $Q \times P^{\infty}$, and becomes.

$$\begin{array}{rcl} J(\theta) & = & E[\| \ (f_{\text{true}} - h(\theta)) \cdot \mathbf{u}_t + \eta_t \ \|^2, Q \times P^{\infty}] \qquad (2.3) \\ & = & E[\| \ \tilde{h}(\theta) \cdot \mathbf{u}_t \ \|^2, P^{\infty}] + E[\| \ \eta \ \|^2, Q], \end{array}$$

where $\tilde{h}(\theta) := h(\theta) - f_{\text{true}}$. Since the second term is independent of θ, we effectively minimize only the first term. In other words, by minimizing $J(\theta)$ with respect to θ, we will find the best approximation to the true system f_{true} in the model family $\{h(\theta), \theta \in \Theta\}$. Note that it is *not* assumed the true system f_{true} belongs to $\{h(\theta), \theta \in \Theta\}$. In case there is a "true" value of θ, call it θ_{true} such that $f_{\text{true}} = h(\theta_{\text{true}})$, then *an* optimal choice of θ is θ_{true}. If in addition we impose some assumptions to the effect that the input sequence $\{u_t\}$ is sufficiently exciting, then $\theta = \theta_{\text{true}}$ becomes the *only* minimizer of $J(\cdot)$.

3 Uniform Convergence of Empirical Means

In this section, it is shown that if a particular property known as UCEM (uniform convergence of empirical means) holds, then a very natural approach of choosing θ_t to minimize the *empirical* (or cumulated) average error will lead to a solution of the system identification problem. Note that such an approach is already adopted in the paper of Ljung [10].

Theorem 1 *For each $t \geq 1$ and each $\theta \in \Theta$, define the empirical error*

$$\hat{J}_t(\theta) := \frac{1}{t} \sum_{i=1}^{t} \ell[y_t, h(\theta) \cdot \mathbf{u}_t].$$

At time t, choose θ_t^* so as to minimize $\hat{J}_t(\theta)$; that is,

$$\theta_t^* = \text{Argmin}_{\theta \in \Theta} \ \hat{J}_t(\theta).$$

Let

$$J^* := \inf_{\theta \in \Theta} J(\theta).$$

Define the quantity

$$q(t, \epsilon) := \tilde{P}_{y,\mathbf{u}}\{\sup_{\theta \in \Theta} |\hat{J}_t(\theta) - J(\theta)| > \epsilon\}. \qquad (3.1)$$

Suppose it is the case that $q(t, \epsilon) \to 0$ as $t \to \infty$. Then

$$\tilde{P}_{y,\mathbf{u}}\{\hat{J}_t(\theta_t^*) > J^* + \epsilon\} \to 0 \text{ as } t \to \infty.$$

Remark: The condition that $q(t, \epsilon) \to 0$ as $t \to \infty$ is usually referred to in the statistical learning theory as the property of **uniform convergence of empirical means (UCEM)**. Thus the theorem states that if the family of error measures $\{J(\theta), \theta \in \Theta\}$ has the UCEM property, then the natural algorithm of choosing θ_t so as to minimize the empirical estimate $\hat{J}(\theta)$ at time t is 'asymptotically optimal.' Moreover, the 'asymptotic' result can actually be used to provide finite time estimates as well.

Thus the sample complexity of ensuring that $J(\theta_t) \leq J^* + \epsilon$ is at most equal to the sample complexity of

$q(m, \epsilon/3)$. This naturally brings up the question as to what kinds of families $\{h(\theta), \theta \in \Theta\}$ have this particular UCEM property, and what their sample complexities are like. These questions are given a very simple-minded answer in the next section.

4 A UCEM Result

In this section, it is shown that the UCEM property of Theorem 1 does indeed hold in the commonly studied case where y_t is the output of a "true" system corrupted by additive noise, and the loss function ℓ is the squared error. By Theorem 1, this implies that by choosing the estimated model $h(\theta_t)$ so as to minimize the cumulated least squares error, we will eventually obtain the best possible fit to the given time series. Note that no particular attempt is made here to state or prove the 'best possible' result. Rather, the objective is to give a flavour of the the statistical learning theory approach by deriving a result whose proof is free from technicalities.

We begin by listing below the assumptions regarding the family of models employed in identification, and on the time series. Recall that the symbol $\tilde{h}(\theta) \cdot \mathbf{u}_t$ denotes the function $(f_{\text{true}} - h(\theta)) \cdot \mathbf{u}_t$. Define the collection of functions $\mathcal{H}$ mapping $\mathcal{U}$ into $\mathbb{R}$ as follows:

$$g(\theta) := \mathbf{u} \mapsto \| \ (f - h(\theta)) \cdot \mathbf{u}_0 \ \|^2 : \mathcal{U} \to \mathbb{R},$$

$$\mathcal{G} := \{g(\theta) : \theta \in \Theta\}.$$

Now the various assumptions are listed.

A1. There exists a constant M such that

$$|g(\theta) \cdot \mathbf{u}_0| \leq M, \ \forall \theta \in \Theta, \mathbf{u} \in \mathcal{U}.$$

This assumption can be satisfied, for example, by assuming that the true system and each system in the family $\{h(\theta), \theta \in \Theta\}$ is BIBO stable (with an upper bound on the gain, independent of θ), and that the set U is bounded (so that $\{u_t\}$ is a bounded stochastic process).

A2. For each integer $\;\; \geq 1$, define

$$g_k(\theta) \cdot \mathbf{u}_t := g(\theta) \cdot (u_{t-1}, u_{t-2}, \ldots, u_{t-k}, 0, 0, \ldots).$$

With this notation, define

$$\mu_k := \sup_{\mathbf{u} \in \mathcal{U}} \sup_{\theta \in \Theta} |(g(\theta) - g_k(\theta)) \cdot u_0|.$$

Then the assumption is that μ_k is finite for each and approaches zero as $\;\; \to \infty$. This assumption essentially means that each of the systems in the model family has decaying memory (in the sense that the effect of the values of the input at the distant past on the current output becomes negligibly small). This assumption is satisfied, for example, if

- Each of the models $h(\theta)$ is a linear ARMA model of the form

$$y_t = \sum_{i=1}^{l} a_i(\theta) u_{t-i} + b_i(\theta) y_{t-i},$$

- The characteristic polynomials

$$\phi(\theta, z) := z^{l+1} - \sum_{i=1}^{l} b_i(\theta) z^{l-i}$$

all have their zeros inside a circle of radius $\rho < 1$, where ρ is independent of θ.

- The numbers $a_i(\theta)$ are uniformly bounded with respect to θ.

The extension of the above condition to MIMO systems is straight-forward and is left to the reader.

A3. Consider the collection of maps $\mathcal{G} = \{g_k(\theta) : \theta \in \Theta\}$, viewed as maps from U^k into $\mathbb{R}$. For each , this family has finite P-dimension, denoted by $d(\;\;)$. (See [13], Chapter 4 for a definition of the P-dimension.)

Now we can state the main theorem.

Theorem 2 *Define the quantity $q(t, \epsilon)$ as in (3.1) and suppose Assumptions A1 through A3 are satisfied. Given an $\epsilon > 0$, choose $\;\;(\epsilon)$ large enough that $\mu_k \leq \epsilon/4$ for all $\;\; \geq \;\;(\epsilon)$. Then for all $t \geq \;\;(\epsilon)$ we have*

$$\begin{aligned} q(t, \epsilon) \leq \quad & 8\;\;(\epsilon) \left(\tfrac{32e}{\epsilon} \ln \tfrac{32e}{\epsilon}\right)^{d(k(\epsilon))} \\ & \cdot \exp\left(-\lfloor t/\;\;(\epsilon) \rfloor \epsilon^2 / 512 M^2\right), \end{aligned} \tag{4.1}$$

where $\lfloor t/\;\;(\epsilon) \rfloor$ denotes the largest integer part of $t/\;\;(\epsilon)$.

Remark: From the proof of Theorem 1, it follows that the rate of convergence of the estimated model to the optimal performance can also be quantified.

5 Bounds on the P-Dimension

In order for the estimate in Theorem 2 to be useful, it is necessary for us to derive an estimate for the P-dimension of the family of functions defined by

$$\mathcal{G}_k := \{g_k(\theta) : \theta \in \Theta\}, \tag{5.1}$$

where $g_k(\theta) : U^k \to \mathbb{R}$ is defined by

$$g_k(\theta)(\mathbf{u}) := \| \ (f - h(\theta)) \cdot \mathbf{u}_k \ \|^2,$$

where

$$\mathbf{u}_k := (\ldots, 0, u_k, u_{k-1}, \ldots, u_1, 0, 0, \ldots).$$

Note that, in the interests of convenience, we have denoted the infinite sequence with only nonzero elements as $u_k, \ldots, u_1$ rather than $u_0, \ldots, u_{1-k}$ as done earlier. Clearly this makes no difference. In this section, we state and prove such an estimate for the commonly occuring case where each system model $h(\theta)$ is an ARMA model where the parameter θ enters linearly. Specifically, it is supposed that the model $h(\theta)$ is described by

$$x_{t+1} = \sum_{i=1}^{l} \theta_i \; \phi_i(x_t, u_t), \; y_t = x_t, \qquad (5.2)$$

where $\theta = (\theta_1, \ldots, \theta_l) \in \Theta \subseteq \mathbb{R}^l$, and each $\phi_i(\cdot, \cdot)$ is a polynomial of degree no larger than r in the components of x_t, u_t.

Theorem 3 *With the above assumptions, we have that*

$$\textit{P-dim}(\mathcal{G}_k) \leq 9l + 2l \quad \lg r. \qquad (5.3)$$

Remarks: It is interesting to note that the above estimate is *linear* in both the number of parameters l and the duration of the input sequence $\mathbf{u}$, but is only logarithmic in the degree of the polynomials ϕ_i. In the case of *linear* systems, Dasgupta and Sontag [5] have derived VC-dimension bounds that are *logarithmic* in . Their problem formulation is a little different; however, perhaps with a little effort their result can be incorporated into the present formulation as well. That is a problem for future research.

Proof: For each function $g_k(\theta) : U^k \rightarrow \mathbb{R}$ defined as in (5.1), define an associated function $g'_k : U^k \times [0, 1] \rightarrow \{0, 1\}$ as follows:

$$g'_k(\theta)(\mathbf{u}, c) := \eta[g_k(\theta)(\mathbf{u}) - c],$$

where $\eta(\cdot)$ is the Heaviside or 'step' function. In other words, $\eta(s) = 1$ if $s \geq 0$, and $\eta(s) = 0$ if $s < 0$. Let $\mathcal{G}'_k$ denote the associated family of functions $g'_k(\theta)$ as θ varies over Θ. Then it is known (see [12] or [13], Lemma 10.1) that

$$\text{P-dim}(\mathcal{G}_k) = \text{VC-dim}(\mathcal{G}'_k).$$

Next, to estimate VC-dim$(\mathcal{G}'_k)$, we use a result due to Karpinski and Macintyre [8, 9], with a refinement due to [13], Corollary 10.2. This result states that, if the condition $\eta[g_k(\theta)\mathbf{u} - c] = 1$ can be stated as a Boolean formula involving s polynomial inequalities, each of degree no larger than d, then

$$\text{VC-dim}(\mathcal{G}'_k) \leq 2l \lg(4eds). \qquad (5.4)$$

Thus the proof consists of showing that the conditions needed to apply this bound hold, and of estimating the constants d and s.

Towards this end, let us back-substitute repeatedly into the ARMA model (5.1) to express the inequality

$$\| \, (f - h(\theta))\mathbf{u}_k \, \|^2 \, - c < 0$$

as a polynomial inequality in $\mathbf{u}$ and the θ-parameters. To begin with, we have

$$\begin{aligned} x_{k+1} &= \sum_{i=1}^{l} \theta_i \; \phi_i(x_k, u_k) \\ &= \sum_{i=1}^{l} \theta_i \phi_i \left(\sum_{j=1}^{l} \theta_j \; \phi_j(x_{k-1}, u_{k-1}) \right) = (5.5) \end{aligned}$$

Thus each time one of the functions ϕ_i is applied to its argument, the degree with respect to any of the θ_j goes up by a factor of r. In other words, the total degree of x_{k+1} with respect to each of the θ_j is no larger than $1 + r + r^2 + \ldots + r^{k-1} \leq r^k$. Next, we can write

$$\| \, x_{k+1} \, \|^2 \, - c < 0 \;\Leftrightarrow\; x'_{k+1} x_{k+1} - c < 0.$$

This is a single polynomial inequality in the components of θ of degree at most $2r^k$. Thus we can apply the bound (5.5) with $d = 2r^k$ and $s = 1$. This leads to

$$\text{VC-dim}(\mathcal{G}'_k) \leq 2l \lg(8er^k).$$

The desired estimate now follows on noting that $\lg e < 1.5$, so that $\lg(8e) < 4.5$. ■

6 Conclusions

In this paper, a beginning has been made towards showing that it is possible to use the methods of statistical learning theory to derive *finite time* estimates for use in system identification theory. Obviously there is a great deal of room for improvement in the *specific results* presented here. For instance, in Sections 4 and 5, it would be desirable to combine the fading memory argument and the ARMA model into a single step. This would require new results in statistical learning theory, whereby one would have to compute the VC-dimension of mappings whose range is an infinite-dimensional space. This has not been the practice thus far.

As it stands, the bound derived in Theorem 3 is *linear* in the length . In [5], an improved bound for VC-dimension is derived for *linear* systems; however, that problem formulation differs slightly from the present one. It is an interesting (and perhaps not very difficult) problem for future research to use their approach in the present context and to improve the bound in Theorem 3 to be *logarithmic* in .

In summary, the message of the paper is that both system identification theory and statistical learning theory can enrich each other. Much work remains to be done to take advantage of this potential.

References

[1] [AB99] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge, UK, 1999.

[2] [Caines76] P. E. Caines, "Prediction error identification methods for stationary stochastic processes," *IEEE Trans. Auto. Control*, AC-21(4), 500-505, Aug. 1976.

[3] [Caines78] P. E. Caines, "Stationary linear and nonlinear system identification and predictor set completeness," *IEEE Trans. Auto. Control*, AC-23(4), 583-594, Aug. 1978.

[4] [CW02] M. C. Campi and E. Weyer, "Finite sample properties of system identification methods," *IEEE Trans. Auto. Control*, to appear.

[5] [DS96] B. Dasgupta and E. D. Sontag, "Sample complexity for learning recurrent perceptron mappings," *IEEE Trans. Info. Thy.*, **42**, 1479-1487, 1996.

[6] [Haussler92] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Information and Computation*, **100**, 78-150, 1992.

[7] [KV01] R. L. Karandikar and M. Vidyasagar, "Rates of uniform convergence of empirical means with mixing processes,' to appear in *Statistics and Probability Letters*.

[8] [KM95] M. Karpinski and A.J. Macintyre, "Polynomial bounds for VC dimension of sigmoidal neural networks," *Proc. 27th ACM Symp. Thy. of Computing*, pp. 200-208, 1995.

[9] [KM97] M. Karpinski and A.J. Macintyre, "Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks," *J. Comp. Sys. Sci.*, **54**, pp. 169-176, 1997.

[10] [Ljung78] L. Ljung, "Convergence analysis of parametric identification methods," *IEEE Trans. Auto. Control*, AC-23(5), 770-783, Oct. 1978.

[11] [Ljung99] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, U.S.A., 1999.

[12] [MS93] A.J. Macintyre and E.D. Sontag, "Finiteness results for sigmoidal neural networks," *Proc. 25th ACM Symp. Thy. of Computing*, pp. 325-334, 1993.

[13] [Vidyasagar97] M. Vidyasagar, *A Theory of Learning and Generalization*, Springer-Verlag, London, 1997.

[14] [VK01] M. Vidyasagar and R. L. Karandikar, "A learning theory approach to system identification and stochastic adaptive control," *IFAC Symp. on Adaptation and Learning*, Como, Italy, Aug. 2001.

[15] [VK02] M. Vidyasagar and R. L. Karandikar, "System identification: A learning theory approach," *Proc. IEEE Conf. on Decision and Control*, Orlando, FL, 2001-2006, Dec. 2001.

[16] [Weyer00] E. Weyer, "Finite sample properties of system identification of ARX models under mixing conditions," *Automatica*, 36(9), 1291-1299, Sept. 2000.