A Learning Theory Approach to System Identification

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

[1] Tata Consultancy Services
Khan Lateefkhan Estate
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 formulate the problem of system identification as a problem in statistical learning. By doing so, we are able to derive *finite-time estimates* of the proximity of the current model to the 'true system' if any, or to the 'optimal model' in case there is no true system. The main advantage of doing so is that traditionally system identification theory provides *asymptotic* results. In contrast, statistical learning theory is devoted to the derivation of *finite time estimates*. If system identification is to be combined with robust control theory to develop a sound theory of indirect adaptive control, it is essential to have finite time estimates of the sort provided by statistical 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. In contrast with earlier results in this area, the results presented here are applicable also to nonlinear systems. Moreover, no assumptions are made about the data to which a model is to be fitted, other than that it is a stationary stochastic process. This in contrast to earlier papers which assume that the data is generated by a model of known order. In fact, in the case of linear systems, the estimates presented here are not very conservative, but are more so in the case of nonlinear systems. As there is considerable scope for improving the specific bounds presented here, the results presented here should be viewed as just the beginning of a new theoretical approach.

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 identification 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 identification, 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 [11], 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 [2, 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 [11], Lemma 3.1, or [12], 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. By doing so, we are able to derive finite-time estimates of how much the performance of the current estimate differs from the optimal performance. If these estimates can be combined with robust control theory, then it will be possible to put indirect adaptive control on a sound analytical foundation.

In statistical learning theory, the UCEM property can be established under a variety of assumptions. However, the most common assumption is that the model family $\{h(\theta), \theta \in \Theta\}$ has finite 'P-dimension,' which is an integer that reflects the 'richness' of the model family. In turn, the P-dimension equals the Vapnik-Chernonenkis (VC-) dimension of an associated family of binary-valued functions; see [13]. Moreover, the differentiability assumptions made in [11] (and commonly employed in the system identification community) in fact guarantee the finiteness of the P-dimension of the model family $\{h(\theta), \theta \in \Theta\}$. See for example [6]. Thus *in principle* it is possible to derive UCEM results very similar to (if not exactly identical to) those in [11] using statistical learning theory. However, it is well-known in the statistical learning theory community that the estimates of the (finite) VC-dimension based on differentiability assumptions are extremely conservative. Indeed, obtaining 'tight' estimates for the VC-dimension has been one of the dominant themes of statistical learning theory for the past several years. Over the years, several methods have been developed for estimating the VC-dimension of various function families. The books [14, 1, 15] contain quite complete descriptions of the known results. Many of these estimates can be applied directly to the types of model classes that are widely used in system identification theory. One of the objectives of the present paper is in fact to apply these bounds to identification.

Two distinct types of identification problems are studied here. The first might be called 'model-free' identification, in the sense that no assumptions at all are made on the data to which one is attempting to fit a model, other than that it is a bounded stationary time series. (The assumption of boundedness is purely technical and can definitely be relaxed by resorting to more careful arguments.) This model-free approach is in contrast to the usual assumption made in identification theory, namely that the data is generated by a model of known order; see for example Section 2.1 of [4]. The results derived here show that the rate of convergence of the identified model

to the 'optimal' model depend only on the richness of the *model class*, and not on the data. The second approach builds on the earlier work of [18, 17] and assumes that the data is indeed generated by an input-output stable system driven by an i.i.d. noise sequence with bounded variance. It is *not* assumed that the system is linear. By invoking recently proved results about the mixing properties of such systems and combining these with the results of [18, 17], we extend system identification theory to nonlinear systems. In principle we derive 'explicit' estimates for the rates of convergence. However, since these estimates are based on Lyapunov theory, there is considerable scope for improvement.

The results presented here represent only a begining at applying statistical learning theory to the longstanding problem of system identification, and 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 System Identification with No Restrictions on Data

2.1 Preliminaries

In this section, we state the problem of system identification *without* assuming anything about the nature of the data, other than that it is a stationary stochastic process assuming values in a bounded set. The assumption that the values lie in a bounded set is made purely to avoid lots of technicalities, as the assumption guarantees that the stochastic process has finite moments of all orders; perhaps, with some care, this assumption can be relaxed. Other than this assumption, there are no other assumptions about the nature of the data. For instance, it is *not* assumed, as is done in [4] for example, that the data is generated by an ARMA model whose order is bounded ahead of time. Any and all assumptions are on the *model family* used to fit the data, and not on the data itself. This seems to represent a fairly significant departure from previous work.

For the class of systems under study, the output set is some $Y \subseteq \mathbb{R}^k$, while the input set is some *bounded* $U \subseteq \mathbb{R}^\ell$ for some k and ℓ. There is also a 'loss function' $\ell : Y \times Y \to [0,1]$. The purpose of the loss function is to assign a quantitative value to the error between the actual output and the predicted output.

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 σ-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: Problem Formulation

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. The duality between two-sided infinite sequences and one-sided infinite sequences is always present in stochastic process theory, and we shall not attempt to reconcile it here. It suffices to say that the stochastic process is assumed to stretch into the infinite past, but the system identification process has a definite starting time which can be denoted as $t = 0$.

The objective of system identification is to fit the time series with a model from a specified class. 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.

Problem (System Identification): Choose the paramater $\theta \in \Theta$ so as to minimize the function $J(\theta)$.

Thus the objective of system identification is to find the model $h(\theta)$ from within the family $\{h(\theta), \theta \in \Theta\}$ that best fits the data, as represented by the stochastic process $\{(y_t, u_t)\}$. The difficulty of this problem arises from the fact that the underlying probability measure $\tilde{P}_{\mathbf{y},\mathbf{u}}$, with respect to which the expectation is being taken, is *in general unknown*. This is because, if we know the measure $\tilde{P}_{\mathbf{y},\mathbf{u}}$, then we know all the statistics of the stochastic process and there is nothing to model and nothing to identify. Thus we are obliged to resort to indirect methods to achieve this minimization.

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

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

Thus the expectation is taken with respect to the 'one-dimensional' marginal measure $\tilde{P}_{y,\mathbf{u}}$ on $Y \times \mathcal{U}$. 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 $J(\theta)$.

The theory presented here applies to *any* loss function. However, for some specific loss functions, the problem formulation becomes very natural. For instance, suppose

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

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. Suppose further that the output y_t arises from a 'true but unknown' system driven by i.i.d. noise, corrupted by i.i.d. measurement noise. Specifically, suppose

$$y_t = f_{\text{true}} \cdot \mathbf{u}_t + \eta_t, \ \forall t, \tag{2.2}$$

where the input sequence $\{u_t\}_{-\infty}^{\infty}$ is i.i.d. according to some law P, $\{\eta_t\}_{-\infty}^{\infty}$ is a measurement noise sequence that is zero mean and i.i.d. with law Q, and in addition, u_i, η_j are independent for each i, j. 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.

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

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 $\{h(\theta), \theta \in \Theta\}$. Here by 'best approximation' we mean the system $h(\theta)$ that minimizes the second moment of the output error. With suitable assumptions on u_t (e.g., white noise), this quantity can be readily related to the H_∞-norm of the error transfer function $\tilde{h}(\theta)$. 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 [11]. Note also that the result given here is not by any means the most general possible. In particular, it is possible to show that if θ_t is chosen so as to 'nearly' minimize the empirical error 'most of the time,' then the resulting algorithm will still be asymptotically optimal. For an exposition of this approach to the standard PAC learning problem, see [14], Section 3.2.

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

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

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, as the proof below makes clear, the 'asymptotic' result can actually be used to provide finite time estimates as well.

Proof: Suppose $q(t,\epsilon) \to 0$ as $t \to \infty$. Given any number $\delta > 0$, choose t_0 large enough that

$$\tilde{P}_{y,\mathbf{u}}\{\sup_{\theta\in\Theta} |\hat{J}_t(\theta) - J(\theta)| > \epsilon/3\} < \delta. \tag{3.2}$$

This number is often referred to as the 'sample complexity' corresponding to the accuracy $\epsilon/3$ and confidence δ. Select a $\theta_\epsilon \in \Theta$ such that $J(\theta_\epsilon) \leq J^* + \epsilon/3$. Such a θ_ϵ exists in view of the definition of J^*. Then, in view of (3.2), we can say with confidence $1-\delta$ that

$$\hat{J}(\theta_t) \geq J(\theta_t) - \epsilon/3, \text{ and } \hat{J}(\theta_\epsilon) \leq J(\theta_\epsilon) + \epsilon/3.$$

By definition,

$$\hat{J}(\theta_t) \leq \hat{J}(\theta_\epsilon).$$

Combining these two inequalities shows that

$$\begin{aligned} J(\theta_t) &\leq \hat{J}(\theta_t) + \epsilon/3 \leq \hat{J}(\theta_\epsilon) + \epsilon/3 \leq J(\theta_\epsilon) + 2\epsilon/3 \\ &\leq J^* + 2\epsilon/3 + \epsilon/3 = J^* + \epsilon. \end{aligned} \tag{3.3}$$

This statement holds with confidence $1-\delta$. ■

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 $k \geq 1$, 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 k and approaches zero as $k \to \infty$. 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 k, this family has finite P-dimension, denoted by $d(k)$. (See [14], 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 $k(\epsilon)$ large enough that $\mu_k \leq \epsilon/4$ for all $k \geq k(\epsilon)$. Then for all $t \geq k(\epsilon)$ we have*

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

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

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.

Proof: Write $g(\theta) = g_k(\theta) + (g(\theta) - g_k(\theta))$, and define

$$f_k(\theta) := g_k(\theta)\cdot \mathbf{u}_i,\ \tilde{f}_k(\theta) := (g(\theta) - g_k(\theta))\cdot \mathbf{u}_i.$$

Next, define

$$q_1^k(t,\epsilon) := \Pr\{\sup_{\theta\in\Theta}\left|\frac{1}{t}\sum_{i=1}^{t} f_k(\theta) - E[f_k(\theta),\tilde{P}]\right| > \epsilon\},$$

$$q_2^k(t,\epsilon) := \Pr\{\sup_{\theta\in\Theta}\left|\frac{1}{t}\sum_{i=1}^{t} \tilde{f}_k(\theta) - E[\tilde{f}_k(\theta),\tilde{P}]\right| > \epsilon\}.$$

Then it is easy to see that

$$q(t,\epsilon) \leq q_1^k(t,\epsilon/2) + q_2^k(t,\epsilon/2). \quad (4.2)$$

Now observe that if k is sufficiently large that $\mu_k \leq \epsilon/4$, then $q_2^k(t,\epsilon) = 0$. This is because, if $|(g(\theta) - g_k(\theta))\cdot \mathbf{u}_i|$ is always smaller than $\epsilon/4$, then its expected value is also smaller than $\epsilon/4$, so that their difference can be at most equal to $\epsilon/2$. Since this is true for all $\mathbf{u}$ and all θ, the above observation follows. Thus it follows that if $k(\epsilon)$ is chosen large enough that $\mu_k \leq \epsilon/4$ for all $k \geq k(\epsilon)$, then

$$q(t,\epsilon) \leq q_1^{k(\epsilon)}(t,\epsilon/2)\ \forall t \geq k(\epsilon),\ \forall \epsilon. \quad (4.3)$$

Hence the rest of the proof consists of estimating $q_1^{k(\epsilon)}(t,\epsilon)$ when $t \geq k(\epsilon)$.

From here onwards, let us replace $k(\epsilon)$ by k in the interests of notational clarity. When $t \geq k$, define $l := \lfloor t/k\rfloor$, and $r = t - kl$. Partition $\{1,\ldots,t\}$ into k intervals, as follows.

$$I_j := \{i, i+k, \ldots, i+lk\} \text{ for } 1 \leq j \leq r, \text{ and}$$

$$I_j := \{i, i+k, \ldots, i+(l-1)k\} \text{ for } r+1 \leq j \leq k.$$

Then we can write

$$\frac{1}{t}\sum_{i=1}^{t} g_k(\theta)\cdot\mathbf{u}_i = \frac{1}{t}\sum_{j=1}^{k}\sum_{i\in I_j} g_k(\theta)\cdot\mathbf{u}_i.$$

Now define

$$\alpha_j := \frac{1}{|I_j|}\left|\sum_{i\in I_j}\left(g_k(\theta)\cdot\mathbf{u}_i - E[g_k(\theta)\cdot\mathbf{u}_i,\tilde{P}]\right)\right|.$$

Then, noting that $E[g_k(\theta)\cdot\mathbf{u}_i,\tilde{P}]$ is independent of i due to the stationarity assumption, we get

$$\left|\frac{1}{t}\sum_{i=1}^{t} g_k(\theta)\cdot\mathbf{u}_i - E[g_k(\theta)\cdot\mathbf{u}_i,\tilde{P}]\right| \leq \left|\sum_{j=1}^{r}\frac{l+1}{t}\alpha_j + \sum_{j=r+1}^{k}\frac{l}{t}\alpha_j\right|.$$

It follows that if $\alpha_j \leq \epsilon$ for each j, then the left side of the equality is also less than ϵ. So the following containment of events holds:

$$\left\{\sup_{\theta\in\Theta}\left|\frac{1}{t}(g_k\cdot\mathbf{u}_i - E[g_k\cdot\mathbf{u}_i,\tilde{P}]\right| > \epsilon\right\} \subseteq \bigcup_{j=1}^{k}\{\alpha_j > \epsilon\}.$$

Hence

$$q_1^k(t,\epsilon) \leq \sum_{j=1}^{k}\Pr\{\alpha_j > \epsilon\}. \quad (4.4)$$

Now note that each $g_k\cdot\mathbf{u}_i$ depends on only u_{i-1} through u_{i-k}. Hence, in the summation defining each of the α_j, the various quantities being summed are independent. Since it is assumed that the family $\{g_k(\theta), \theta\in\Theta\}$ has finite P-dimension $d(k)$, standard results from statistical learning theory can be used to bound each of the probabilities on the right side of (4.4). A small adjustment is necessary, however. The results stated in [14] for example assume that all the functions under study assume values in the interval $[0,1]$, whereas in the present instance the functions $h(\theta)\cdot\mathbf{u}_i$ all assume values in the interval $[-M,M]$.

Thus the range of values now has width $2M$ instead on one. With this adjustment, Equation (7.1) of [14] implies that

$$\Pr\{\alpha_j > \epsilon\} \leq 8\left(\frac{16e}{\epsilon}\ln\frac{16e}{\epsilon}\right)^{d(k)} \exp(-|I_j|^2\epsilon^2/128M^2)\phi_i.$$

Clearly $\exp(-|I_j|^2) \leq \exp(-l^2)$. Hence

$$q_1^k(t,\epsilon) \leq 8k\left(\frac{16e}{\epsilon}\ln\frac{16e}{\epsilon}\right)^{d(k)} \exp(-l^2\epsilon^2/128M^2). \tag{4.5}$$

Finally, the conclusion (4.1) is obtained by replacing ϵ by $\epsilon/2$ in the above expression, and then applying (4.3). ■

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

$$g_k(\theta)(\mathbf{u}) := \| (f - h(\theta)) \cdot \mathbf{u}_k \|^2, \tag{5.2}$$

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 k 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

$$x_{t+1} = \sum_{i=1}^{l} \theta_i \, \phi_i(x_t, u_t), \; y_t = x_t, \tag{5.3}$$

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*

$$\begin{aligned} \textit{P-dim}(\mathcal{G}_k) &\leq 9l + 2l\lg[2(r^{k+1}-1)/(r-1)] \\ &\approx 9l + 2lk\lg(2r) \text{ if } r > 1. \end{aligned} \tag{5.4}$$

In case $r = 1$ so that each system is linear, the above bound can be simplified to

$$\textit{P-dim}(\mathcal{G}_k) \leq 9l + 2l\lg(2k). \tag{5.5}$$

Remark: It is interesting to note that the above estimate is *linear* in both the number of parameters l and the duration k of the input sequence $\mathbf{u}$, but is only logarithmic in the degree of the polynomials ϕ_i. In the practically important case of linear ARMA models, even k appears inside the logarithm.

Proof: For each function $g_k(\theta) : U^k \to \Re$ defined as in (5.2), define an associated function $g_k' : U^k \times [0,1] \to \{0,1\}$ as follows:

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

where $\eta(\cdot)$ is the Heaviside or 'step' function. Then it follows from [14], 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 [14], Corollary 10.2, which 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

$$\text{VC-dim}(\mathcal{G}_k') \leq 2l\lg(4eds). \tag{5.6}$$

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 (5.3) 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

$$\begin{aligned} x_{k+1} &= \sum_{i=1}^{l} \theta_i \, \phi_i(x_k, u_k) \\ &= \sum_{i=1}^{l} \theta_i \phi_i \left(\sum_{j=1}^{l} \theta_j \, \phi_j(x_{k-1}, u_{k-1}) \right) \\ &= \ldots \end{aligned} \tag{5.7}$$

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 $1 + r + r^2 + \ldots + r^k = (r^{k+1} - 1)/(r-1)$. If $r = 1$, then the degree is simply 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. Moreover, the degree of this polynomial in the components of θ is at most $2(r^{k+1} - 1)/(r-1)$ if $r > 1$, and $2k$ if $r = 1$.

Thus we can apply the bound (5.6) with and $s = 1$, and

$$d = \begin{cases} \frac{2(r^{k+1}-1)}{r-1} & \text{if} \quad r > 1, \\ 2k & \text{if} \quad r = 1. \end{cases}$$

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

6 Mixing Properties of Dynamical Systems

In [17] it is shown that if the time series to be fit with a suitable model is β-mixing, then conventional least-squares identification methods will possess the UCEM property. This result generalizes an earlier result in [18] in which the same conclusion is reached under the much stronger assumption that the time series to be fitted with a model exhibits finite dependence (that is, $(y_{t+k}, u_{t+k}$ is independent of (y_t, u_t) whenever k exceeds some finite number k_0. In [4], this very promising approach to system identification is dismissed with the off-hand observation that 'signals generated by dynamical systems are not β-mixing in general.' On the contrary, recent results from [8] show that input-output stable dynamical system *do* generate β-mixing sequences. The main result, taken from [8], is as follows:

Throughout, we consider Markov chains described by the recursion relation

$$x_{t+1} = f(x_t, e_t), \tag{6.1}$$

where $x_t \in \mathbb{R}^k, e_t \in \mathbb{R}^m$ for some integers k, m, and $\{e_t\}$ is a stationary noise sequence. It is assumed that the following assumptions are satisfied:

A1. The function $f : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}^k$ is 'smooth,' i.e., is C^∞, and in addition, f is globally Lipschitz continuous. Thus there exist constants L and K such that

$$|f(x,u) - f(y,v)| \leq L|x-y| + K|u-v|. \tag{6.2}$$

A2. The noise sequence $\{e_t\}$ is i.i.d., has finite variance, and has a continuous multivariate density function $\phi(\cdot)$ that is positive in some neighbourhood Ω of the origin in $\mathbb{R}^m$.

A3. When $e_t = 0\ \forall t$, the 'unforced' system

$$x_{t+1} = f(x_t, 0)$$

is globally exponentially stable with the origin as the unique globally attractive equilibrium. This means that there exist constants M' and $\lambda < 1$ such that

$$|x_t| \leq M'|x_0|\lambda^t,\ \forall t \geq 1,\ \forall x_0.$$

By taking $M := \max\{M', 1\}$, one can write the above inequality as

$$|x_t| \leq M|x_0|\lambda^t,\ \forall t \geq 0,\ \forall x_0.$$

A4. The associated deterministic control system

$$x_{t+1} = f(x_t, u_t) \tag{6.3}$$

is 'globally forward accessible' from the origin with the control set Ω. In other words, for every $y \in \mathbb{R}^k$, there exist a time N and a control sequence $\{u_0, \ldots, u_{N-1}\} \subseteq \Omega$ such that, with $x_0 = 0$ we have $x_N = y$.

A5. The associated deterministic control system (6.3) is 'locally controllable' to the origin with the control set Ω. This means that there exists a neighbourhood $\mathcal{B}$ of the origin in $\mathbb{R}^k$ such that, for every $y \in \mathcal{B}$ there exist a time N and a control sequence $\{u_0, \ldots, u_{N-1}\} \subseteq \Omega$ such that, with $x_0 = y$ we have $x_N = 0$.

Now we can state the main result.

Theorem 4 *Suppose assumptions A1 through A5 hold. Then the state sequence $\{\mathcal{X}_t\}$ is geometrically β-mixing.*

The next result shows that if a Markov chain is (geometrically) β-mixing, so is any hidden Markov model generated from the Markov chain. Actually, the result is more general than that.

Theorem 5 *Suppose $\{\mathcal{X}_t\}_{t\geq 0}$ is a stationary stochastic process assuming values in a set X with associated σ-algebra $\mathcal{S}$. Suppose Y is a complete separable metric space, and let $\mathcal{B}(Y)$ denote the Borel σ-algebra on Y. Suppose $\mu : X \times \mathcal{B}(Y) \to [0,1]$ is a transition probability function. Thus for each $x \in X$, $\mu(x, \cdot)$ is a probability measure on Y, and for each $A \in \mathcal{B}(Y)$, $\mu(\cdot, A)$ is a measurable function on $(X, \mathcal{S})$. Finally, suppose $\{\mathcal{Y}_t\}_{t\geq 0}$ is a Y-valued stochastic process such that*

$$\Pr\{\mathcal{Y}_t \in A | \mathcal{Y}_i, i \leq t-1, \mathcal{X}_j, j \leq t\} = \mu(\mathcal{X}_t, A).$$

Under these assumptions, if $\{\mathcal{X}_t\}$ is β-mixing, so is $\{\mathcal{Y}_t\}$.

Proofs of both theorems can be found in [8].

Thus, in summary, the approach of [17] can be applied very fruitfully to problems of identifying time series when the data is generated by an input-output stable system driven by a noise signal with bounded variance. Needless to say, this includes the standard situation of a stable linear system driven by i.i.d. Gaussian noise.

7 Conclusions

In this paper, a general approach has been outlined on using 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.

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

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

[3] 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] M. C. Campi and E. Weyer, "Finite sample properties of system identification methods," *IEEE Trans. Auto. Control*, 47, 1329-1334, 2002.

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

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

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

[8] R. L. Karandikar and M. Vidyasagar, "Probably approximately correct learning with beta-mixing input sequences," submitted for publication.

[9] 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.

[10] 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.

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

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

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

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

[15] M. Vidyasagar, *Learning and Generalization with Applications to Neural Networks*, Springer-Verlag, London 2003.

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

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

[18] E. Weyer, R. C. Williamson and I. Mareels, "Finite sample properties of linear model identification," *IEEE Trans. Auto. Control*, 44(7), 1370-1383, July 1999.