Sample complexity of stochastic least squares system identification

Erik Weyer[1]

Department of Chemical Engineering
The University of Queensland
Brisbane QLD 4072, Australia
Email: erikw@cheque.uq.oz.au
Fax: +61-7-3365 4199, Phone: +61-7-3365 4551

Robert C. Williamson
Iven M. Y. Mareels

Department of Engineering
Australian National University
Canberra ACT 0200, Australia
Email: Bob.Williamson@anu.edu.au, Iven.Mareels@anu.edu.au.

November 1995

Abstract

In this paper we consider the finite sample properties of least squares identification in a stochastic framework. The problem we pose is: How many data points are required to guarantee with high probability that the expected value of the quadratic identification criterion is close to its empirical mean value. The sample sizes are obtained using risk minimisation theory which provides uniform probabilistic bounds on the difference between the expected value of the squared prediction error and its empirical mean evaluated on a finite number of data points. The bounds are very general. No assumption is made about the true system belonging to the model class, and the noise sequence is not assumed to be uniformly bounded. Further analysis shows that in order to maintain a given bound on the deviation, the number of data points needed grows no faster than quadratically with the number of parameters for FIR and ARX models. Results for general asymptotic stable linear models are also obtained by considering ARX approximations of these models. The sample sizes derived for FIR and ARX models differ dramatically from the exponential sample sizes obtained for deterministic worst case l_1 identification. However, the setting and problem considered in these two cases (stochastic least squares and deterministic l_1) are completely different, so the sample sizes are not really comparable.

Keywords: Least squares system identification, Finite sample properties, Risk minimisation theory, FIR models, ARX models, General linear models.

[1]Corresponding author

1 Introduction

In this paper we consider the finite sample properties of least squares (LS) system identification. In LS identification the estimate $\hat{\theta}_N$ is given as the value which minimises the sum of the squared errors, i.e.

$$\hat{\theta}_N = \arg \min_{\theta} V_N(\theta)$$

where the criterion function $V_N(\theta)$ is given by

$$V_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} \epsilon^2(t, \theta) \tag{1}$$

where $\epsilon(t, \theta)$ is the prediction error at time t of the model parameterised by θ, and N is the number of data points. The criterion (1) is commonly used in system identification, and the asymptotic convergence properties of $\hat{\theta}_N$ and $V_N(\theta)$ are well understood. It can be shown (Ljung (1978,1987)) that under natural conditions we have that $V_N(\theta)$ converges uniformly to its expected value with probability 1, i.e.

$$\sup_{\theta} |V_N(\theta) - V(\theta)| \to 0 \text{ as } N \to \infty \text{ w.p. } 1$$

where

$$V(\theta) = E\epsilon^2(t, \theta)$$

Here E is the expectation operator. For the parameter estimate we have similarly that

$$\hat{\theta}_N \to \arg \min_{\theta} V(\theta) \text{ as } N \to \infty \text{ w.p. } 1$$

In this paper we address the finite sample properties of LS identification. What can we say about

$$\sup_{\theta} |V_N(\theta) - V(\theta)| \tag{2}$$

for a finite N? How many data points do we need in order to guarantee that (2) is less than a prescribed value with a given (high) probability? We will mainly focus upon the properties of $V_N(\theta)$, and not the properties of $\hat{\theta}_N$. The reason for this is that the criterion $V(\theta)$ should reflect the purpose of the identification in the sense that a good model is a model which gives a small value of $V(\theta)$. Hence the important issue is the value of $V(\hat{\theta}_N)$ at the identified parameter $\hat{\theta}_N$ and not the parameter $\hat{\theta}_N$ in itself. However, the results we obtain on $V_N(\theta)$ can be translated into results about $\hat{\theta}_N$.

In order to derive bounds of the form (2) we will employ risk minimisation theory (Vapnik (1982)). This theory relates the number of data points to the bound (2), the complexity of the identification problem and the probability we can guarantee the bound with. Using the theory we obtain uniform probabilistic bounds on the discrepancy between $V(\theta)$ and $V_N(\theta)$ for a finite number N of data points. Although risk minimisation theory is not yet a common tool for analysing system identification methods, it is in frequent use in the related area of learning theory, see e.g. Haussler (1992). Concepts from learning theory have also been used in a recent paper by Dasgupta and Sontag (1995) to analyse finite time properties of system identification. The potential use of learning theory in system identification was also noted by Ljung in Ljung (1993).

The finite time properties of system identification methods have also been studied elsewhere, but in a completely different setting. Worst case l_1 identification is studied in e.g. Dahleh, Theodosopoulos and Tsitsiklis (1993), Kacewicz and Milanese (1992) and Poolla and Tikku (1994). They all use a deterministic framework and assume unknown but bounded noise and that the system is included in the model set. They consider the set S of all models which are compatible with the observed data and the noise assumptions, and show that in order to keep the diameter of S below a certain value (measured in the l_1 norm of the impulse response), the required number of data points increases exponentially with the model order. Kacewicz and Milanese (1992) also consider optimal input signals, and find that an impulse is optimal in their setting. Finite time properties of worst case deterministic identification in l_1 and H_∞ have also been studied in Zames, Lin and Wang (1994) using n-widths and metric complexity. Also here it is found that an impulse is the optimal input signal. This can hardly be said to be an intuitive results: The optimal input signal for system identification is not persistently exciting of any order!

One of the reasons for the from our perspective somewhat strange and disappointing results obtained in deterministic worst case identification, is that the worst disturbances that can occur are dependent on the applied input signal, and thus violate the nature and intent of a disturbance as being uncorrelated with the input signal. This problem does not occur in our setup.

The most interesting feature about risk minimisation theory is that it enables us to analyse the finite time properties of least squares identification in a stochastic framework without assuming bounded noise or that the system belongs to the model class under consideration. Moreover, using results from risk minimisation theory we can show that the required number of data points to maintain a bound similar to (2) at a constant level increases at most quadratically with the model order for FIR and ARX models. However, it should be noted that the constants we have in front of the quadratic terms are rather large, and the results can therefore not be used directly in practical applications. From the above discussion it is clear that the setting and the problem posed here and in the case of worst case deterministic l_1 identification are quite different, and the obtained sample sizes are therefore not really comparable.

The paper is organised as follows. In the next section we introduce the mathematical assumptions necessary and derive bounds for FIR models before we discuss the sample complexity. The results for FIR models are extended to ARX models in section 3, and further extended to general linear model in section 4 by considering

ARX approximations. Some concluding remarks are given in section 5.

2 FIR models

In this section we consider FIR models

$$y(t) = B(q^{-1})u(t) + e(t) \tag{3}$$

where $y(t)$ is the observed output, $u(t)$ the observed input signal and $e(t)$ an unobserved disturbance term required to obtain equality in equation (3). The operator $B(q^{-1})$ is given by

$$B(q^{-1}) = b_1 q^{-1} + \cdots + b_n q^{-n} \tag{4}$$

where q^{-1} is the backward shift operator ($q^{-1}u(t) = u(t-1)$). Furthermore we assume that $u(t)$ is a sequence of independent identically distributed (iid) random variables. By introducing the parameter vector

$$\theta = [b_1, \ldots, b_n]^T \tag{5}$$

and the regressor

$$\phi(t) = [u(t-1), \ldots, u(t-n)]^T \tag{6}$$

equation (3) can be written as

$$y(t) = \phi^T(t)\theta + e(t)$$

The one step ahead predictor of $y(t)$ associated with the model (3) is given by

$$\hat{y}(t,\theta) = \phi^T(t)\theta \tag{7}$$

and the corresponding prediction error is

$$\epsilon(t,\theta) = y(t) - \hat{y}(t,\theta)$$

Ideally we like to minimise the expected value of the squared prediction error

$$V(\theta) = E\epsilon^2(t,\theta)$$

with respect to θ in order to obtain a good model. However, the underlying system and probability measure are unknown so we can not perform the expectation. Instead we consider the observed data

$$[u(1), y(1), \ldots, u(N), y(N)]$$

and minimise the empirical mean

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} \epsilon^2(t,\theta) = \frac{1}{N}\sum_{t=1}^{N}(y(t) - \phi^T(t)\theta)^2 \tag{8}$$

Our hope is that $V_N(\theta)$ is a good approximation of $V(\theta)$. $\epsilon(t,\theta)$ can be computed given $y(t)$ and $\phi(t)$, and we refer to $(y(t),\phi(t))$ as a data point. In (8) we have for simplicity assumed that we have access to the initial conditions $u(-n+1),\ldots,u(0)$. The parameter estimate

$$\hat{\theta}_N = \arg\min_{\theta} V_N(\theta)$$

is the standard least squares estimate. The asymptotic properties of $V_N(\theta)$ and $\hat{\theta}_N$ have been extensively analysed in the literature, see e.g. Ljung (1987).

Here we consider the finite sample properties of $V_N(\theta)$, and the question we address is: How many data points are needed in order to guarantee that $V_N(\theta)$ and $V(\theta)$ are close. Strictly speaking we do not consider $V_N(\theta)$ directly but a version $V_{N,L,t}(\theta)$ (defined below) computed on a reduced number of independent data points. The questions can be stated as:

How many data points $(y(t),\phi(t))$, $t=1,\ldots,N$ are required to guarantee with probability $1-\eta$ that

$$\sup_{\theta} |V_{N,L,t}(\theta) - V(\theta)| \leq K$$

or

$$\sup_{\theta} \left| \frac{V_{N,L,t}(\theta) - V(\theta)}{V_{N,L,t}(\theta)} \right| \leq K$$

In order to derive such bounds we make use of the theory put forward in Vapnik (1982) with some extensions from Weyer, Williamson and Mareels (1995). Before we apply the theory we introduce some assumptions about the underlying system which has generated the observed data. Basically we introduce assumptions of two different kinds. The first allows us to treat data separated by some time interval as independent of each other, and the second ensures that we do not run into problems with large deviations, where $V_{N,L,t}(\theta)$ can take on a very large value with a very small probability.

The risk minimisation theory is only valid for independent observations of $\epsilon(t,\theta)$, so our first assumption is that $\epsilon(t+iL,\theta)$, $i=0,1,\ldots$ are independent of each other for some known value L. The prediction error is given by

$$\epsilon(t,\theta) = y(t) - \phi^T(t)\theta$$

From (7) and the iid assumption on u it is clear that $\phi(t)$ is independent of $\phi(t+L)$ for $L \geq n$, and hence it is sufficient to assume that $y(t+iL)$, $i=0,1,\ldots$ are iid, which essentially means that the true system is time invariant and has a finite impulse response. Furthermore it implies that all the generating signals (i.e. the input and the various noise sources acting on the system) are iid. However, it is not necessary to assume that the system is linear. In section 3 we shall see that we can relax the assumption of independence of $y(t+iL)$.

The second assumption is introduced in order to avoid problems with large deviations. The problem is that the criterion function can take a very large value with a very small probability, e.g.

$$\epsilon^2(t,\theta) = \begin{cases} 0 & \text{with probability} \;\; 1-\delta \\ 1/\delta^2 & \text{with probability} \;\; \delta \end{cases}$$

The expected value is $E\epsilon^2(t,\theta) = 1/\delta$, but if δ goes to zero then it is very likely that all observations will be $\epsilon(t,\theta) = 0$, and hence the empirical value $V_N(\theta)$ will be 0 and the discrepancy will be $1/\delta$, which tends to infinity as δ goes to zero.

The problem is avoided by either assuming that $\epsilon^2(t,\theta)$ is uniformly bounded or by assuming that a ratio of means of $\epsilon^2(t,\theta)$ is bounded. In this paper we will mainly use the last assumptions which is the weaker one of the two.

We now formalise our assumptions in the following setup.

Framework for Identification

F1 The observed input is iid, and the observed output is such that for a given integer L, the variables $y(t+iL)$ $i = 0, 1, \dots$ are iid for all $t = 1, \dots, L$.

F2 Consider the class of FIR models

$$\begin{aligned} y(t) &= B(q^{-1})u(t) + e(t) \\ B(q^{-1}) &= b_1 q^{-1} + \cdots + b_n q^{-n}, \quad n < L \end{aligned}$$

with associated predictors

$$\hat{y}(t,\theta) = \phi^T(t)\theta$$

where

$$\begin{aligned} \phi(t) &= [u(t-1), \dots, u(t-n)]^T \\ \theta &= [b_1, \dots, b_n]^T \end{aligned}$$

F3 The observed data are $[u(1), y(1), \dots, u(N+n), y(N+n)]$ From the observed data we form (after a reindexing) the data points

$$D_N = [\phi(1), y(1), \dots, \phi(N), y(N)]$$

F4 The expected value of the identification criterion is

$$V(\theta) = E\epsilon^2(t,\theta) \tag{9}$$

F5 The empirical value of the identification criterion evaluated on independent data points $(\phi(t+iL), y(t+iL))$, $i = 0, 1, \dots$ is given by

$$V_{N,L,t}(\theta) = \frac{1}{l}\sum_{i=1}^{l} \epsilon^2(t+iL,\theta) \quad t = 1, \dots, L \tag{10}$$

where l is the number of independent observations and given by $l = \lfloor N/L \rfloor$ the largest integer less than or equal to N/L.

Distribution	τ
Gaussian	3
Laplacian	6
Uniform	1.8

Table 1: Values of τ for some distributions for $p = 2$.

F6 *Bounded ratio of means.* There exists a constant $p > 1$ such that the ratio of the pth order mean to the first order mean of the criterion function is bounded by a constant τ for all θ. That is

$$\frac{(E\epsilon^{2p}(t,\theta))^{\frac{1}{p}}}{E\epsilon^2(t,\theta)} = \frac{\left(E(y(t)-\phi^T(t)\theta)^{2p}\right)^{\frac{1}{p}}}{E(y(t)-\phi^T(t)\theta)^2} \leq \tau \quad \forall\theta \tag{11}$$

End Framework

In the above framework it is assumption **F6** that saves us from problems with large deviations. It is a rather weak assumption which shows the importance of the tail of the probability distribution. Values for τ in the case $\epsilon(t,\theta)$ is Gaussian, Laplacian or uniform are shown in table 1 for $p = 2$. Notice that τ is independent of the parameters (mean and variance) of the distributions.

Using assumption **F6** we can derive bounds on the relative error $\left|\frac{V_{N,L,t}(\theta)-V(\theta)}{V_{N,L,t}(\theta)}\right|$. Sometimes we want bound on the absolute error $|V_{N,L,t}(\theta) - V(\theta)|$. An alternative to assumption **F6** is then to assume a uniformly bounded criterion function, i.e.

F7 Uniformly bounded criterion

$$\sup_{t,\theta} \epsilon^2(t,\theta) < \tau$$

∎

We notice in the case that $u(t)$ and $y(t)$ are Gaussian, we do not have to require that θ belongs to a compact set since $y(t) - \phi^T(t)\theta$ is also Gaussian. As remarked above, the value $\tau = 3$ is independent of the mean and variance, and hence it is also independent of θ. However in a more general setting it may be necessary to assume that θ belongs to a compact set in order to satisfy **F6** or **F7**. This is particularly true for assumption **F7**.

An example of systems that satisfy assumptions in the setup is FIR systems given by

$$y(t) = B(q^{-1})u(t) + C(q^{-1})e(t)$$

where $u(t)$ and $e(t)$ Gaussian, iid and independent of each other, and both $B(q^{-1})$ and $C(q^{-1})$ are of orders less than L. Then **F1** is satisfied, and **F6** is satisfied with $\tau = 3$ since $y(t) - \phi^T(t)\theta$ is Gaussian. Notice that all we have assumed about the true underlying system is **F1**. In particular we have not assumed that it belongs to the class of FIR models considered in **F2**.

In the next section we return to the question: How many data points do we need to obtain a good estimate of $V(\theta)$. To obtain such results we need independent data points, and we will therefore consider the empirical functions $V_{N,L,t}(\theta)$ in **F5** which use independent prediction errors by picking every Lth data point starting at time t.

2.1 Upper and lower bounds on $V(\theta)$

In this section we give uniform (in θ) probabilistic bounds on $V(\theta)$ in terms of $V_{N,L,t}(\theta)$. The bounds are taken from Vapnik (1982) with some extensions given in Weyer, Williamson and Mareels (1995).

2.1.1 The Vapnik Chervonenkis dimension

Quite naturally the bounds are dependent on the complexity of the model class we consider; the more complex a model class is, the weaker are the bounds. One way to measure the complexity of a model class is by the Vapnik Chervonenkis (VC) dimension.

For our purposes the following result suffices.[2]

Lemma 2.1 *Given any model structure with associated predictors that are linear in the model parameters θ*

$$\hat{y}(t,\theta) = \phi^T(t)\theta \tag{12}$$

and a quadratic criterion function

$$\epsilon^2(t,\theta) = (y(t) - \hat{y}(t,\theta))^2 \tag{13}$$

Let M be the number of elements in the parameter vector θ. Then the VC dimension h, of $\{\epsilon^2(\cdot,\theta)|\ \theta \in R^M\}$ is bounded above by

$$h \leq 2(M+1)$$

∎

[2]The authors would like to thank Arne Hole, Dep. of Mathematics, University of Oslo, Norway, for pointing out an error in an earlier version of the lemma.

Since there is some confusion in the literature over the VC dimension of (13), the definition of the VC dimension and a very short sketch of a proof of Lemma 2.1 are given in Appendix A.1.

For easy reference we introduce the following definition of the VC dimension of a model structure. The definition follows naturally from Lemma 2.1

Definition 2.2 *Consider a model structure whose associated predictors are linear in the model parameters θ. We define the VC dimension of the model structure to be equal to the VC dimension of the set of quadratic functions given by (13).* ∎

Example. The predictor for an nth order FIR model can be written in the form (12) with $\phi(t)$ and θ given by (6) and (5), and hence the VC dimension is bounded above by $h \leq 2(n+1)$.

For the ARX model

$$\begin{aligned} A(q^{-1})y(t) &= B(q^{-1})u(t) + e(t) \\ A(q^{-1}) &= 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a} \\ B(q^{-1}) &= b_1 q^{-1} + \cdots + b_{n_b} q^{n_b} \end{aligned}$$

we have

$$\begin{aligned} \phi(t) &= [-y(t-1), \ldots, -y(t-n_a), u(t-1), \ldots, u(t-n_b)]^T \\ \theta &= [a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}]^T \end{aligned}$$

and the VC dimension is bounded above by $h \leq 2(n_a + n_b + 1)$. ∎

Remark. From the sketch of the proof of 2.1 in Appendix A.1 it follows that the VC dimension of the function class

$$\{|\epsilon(\cdot, \theta)| = |y(\cdot) - \phi^T(\cdot)\theta| \mid \theta \in R^M\} \tag{14}$$

is also bounded above by $2(M+1)$, and the results in this paper extend straightforward or with some minor modifications to the criterion function (14). ∎

2.1.2 Upper and lower bounds

We are now in the position that we can state the upper and lower bounds on $V(\theta)$ in terms of $V_{N,L,t}(\theta)$. The upper bound is given in the following theorem which is a special case of Theorem 7.6 in Vapnik (1982)

Theorem 2.3 (Upper bound) *Consider the framework* **F1-F6**. *Let the number of independent data points $l = \lfloor N/L \rfloor$ be larger than the VC dimension h of the FIR models which is bounded above by $2(n+1)$. Then with probability $1-\eta$*

$$V(\theta) \leq \left[\frac{V_{N,L,t}(\theta)}{1 - T(h,\eta,l)} \right]_\infty \tag{15}$$

is valid simultaneously for all θ. V is given by (9), and $V_{N,L,t}$ is given by (10). Here

$$T(h,\eta,l) = \begin{cases} 2\tau a(p)\sqrt{\frac{h\ln(2l)-\ln(h!)-\ln(\eta/12)}{l}} & \text{for } p > 2 \\ \tau V_p\left(2\sqrt{\frac{h\ln(2l)-\ln(h!)-\ln(\eta/12)}{l^{2-(2/p)}}}\right) & \text{for } 1 < p \leq 2 \end{cases}$$

where p and τ are given in **F6**, *and*

$$\begin{aligned} a(p) &= \left(\frac{(p-1)^{p-1}}{2(p-2)^{p-1}}\right)^{\frac{1}{p}} \\ V_p(\kappa) &= \kappa\left(1 - \frac{\ln\kappa}{p^{\frac{1}{p-1}}(p-1)}\right)^{\frac{p-1}{p}} \end{aligned}$$

$$[z]_\infty = \begin{cases} z & \text{if } z \geq 0 \\ \infty & \text{if } z < 0 \end{cases}$$

∎

Proof. See Vapnik (1982). ∎

Before we state the lower bound, we introduce an empirical functional. Let the values of the squared prediction errors be

$$z_t(i+1,\theta) = (y(t+iL) - \hat{y}(t+iL,\theta))^2, \quad i = 0,\ldots,l-1$$

Without loss of generality we can assume that $z_t(1,\theta) \leq z_t(2,\theta) \leq \cdots \leq z_t(l,\theta)$. The empirical functional $R_{N,L,t}(\theta)$ is given by

$$R_{N,L,t}(\theta) = \sum_{i=1}^{l}(z_t(i,\theta) - z_t(i-1,\theta))\sqrt{\frac{l-i+1}{l}} \geq 0 \tag{16}$$

where $z_t(0,\theta) = 0$.

A lower bound on $V(\theta)$ is now given by

Theorem 2.4 (Lower bound) *Consider the framework* **F1-F6** *and let* $l > h$. *Then with probability* $1 - \eta$

$$V(\theta) \geq \left[V_{N,L,t}(\theta) - 2\sqrt{\frac{h\ln(2l) - \ln h! - \ln\frac{\eta}{12}}{l}} R_{N,L,t}(\theta) \right]_0$$

is valid for all θ *simultaneously.* V *is given by (9),* $V_{N,L,t}$ *by (10) and* $R_{N,L,t}$ *by (16).* l *is the number of independent data points, and* h *is the VC dimension of the FIR models under consideration.*

$$[z]_0 = \begin{cases} z & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

∎

Proof. See Weyer (1992) or Weyer, Williamson and Mareels (1995). ∎

The important variables in these bounds, apart from $V_{N,L,t}(\theta)$, are the VC dimension h, and the sample size l. The functional dependence of the upper (lower) bound on h and l is quite natural in the sense that the upper (lower) bound increases (decreases) with h and decreases (increases) with l. We expect the empirical and expected values to come close when the number of observations increases, and it should not come as a surprise that the bounds (remember they are uniform) get more conservative when we consider an increasing number of parameters, i.e. we consider an increasingly complex estimation task. The bounds make a connection between the sample size, the complexity of the identification problem and the desired reliability of the estimate.

The bounds are potentially conservative since they are derived through a stochastic worst case analysis, but we must keep in mind the very weak assumptions made. The positive result is that all that is required of the unknown probability measure is that assumption **F6** is satisfied.

There are many results on system identification problems in a stochastic setting, see e.g. Ljung (1987), but most of them are asymptotic. Here we have given a result valid for a finite number of samples.

The upper bound can also be used as a criterion for model order selection similar to the more standard AIC, FPE and MDL criteria. This principle for model order selection is called structural risk minimisation in Vapnik (1982).

2.2 Sample complexity

We now consider how many samples N^*, we need in order to guarantee with probability $1-\eta$ that

$$V(\theta) \leq V_{N^*,L,t}(\theta)(1+K) \tag{17}$$

where K represents the allowable relative error. Obviously N^* is a function of the model order n, the relative error K, and the probability $1-\eta$, and hence we use the notation $N^*(n,K,\eta)$.

In order to compute $N^*(n,K,\eta)$, we first solve

$$\frac{1}{1-T(2(n+1),\eta,l)} \leq 1+K \tag{18}$$

with respect to l. In (18) we have made use of the fact that the VC dimension of an nth order FIR model is bounded above by $2(n+1)$. If we denote the smallest integer l, satisfying (18) by $l_{\min}(n,K,\eta)$, the overall number of data points we need is

$$N(n,K,\eta) = l_{\min}(n,K,\eta)L$$

since $V_{N,L,t}(\theta)$ is formed by picking every Lth sample.

In the setup we have assumed that $n < L$, which makes sense since there is no need to choose the model order larger than L when we know that outputs L time units apart are independent of each other. However, in the following we are going to study the sample complexity as the model order tends to infinity. We therefore allow the model order to be larger than L. In this case the prediction errors L samples apart are not independent since the predicted outputs are dependent. The prediction errors n samples apart however are independent, and the required overall number of data points is then

$$N(n,K,\eta) = l_{\min}(n,K,\eta)n$$

We now formalise what we mean by the sample complexity of the identification problem. The sample complexity of least squares identification of FIR models is the minimum number of samples $N^*(n,K,\eta)$, required such that (17) is satisfied with probability $1-\eta$. However, in order to generalise the sample complexity to least squares identification of ARX models we define it in a slightly different but equivalent way.

Definition 2.5 *Let $l_{\min}(n,K,\eta)$ be the smallest integer which satisfies*

$$\frac{1}{1-T(2(n+1),\eta,l)} \leq 1+K$$

where T is given in theorem 2.3. We define the sample complexity of least squares identification of FIR models as

$$N^*(n,K,\eta) = l_{\min}(n,K,\eta)\max(L,n)$$

where L is given in **F1**. ∎

There is no closed form solution of $N^*(n, K, \eta)$, but it can easily be computed numerically given values of n, η and K. We next investigate how the sample complexity depends on n, η and K, and we introduce the following variables

$$\begin{aligned} N^*_{K,\eta}(n) &= N^*(n, K, \eta) \\ N^*_{n,\eta}(K) &= N^*(n, K, \eta) \\ N^*_{n,K}(\eta) &= N^*(n, K, \eta) \end{aligned}$$

That is $N^*_{K,\eta}(n), N^*_{n,\eta}(K)$ and $N^*_{n,K}(\eta)$ is the sample complexity with respect to n, K and η respectively, when the other parameters are constants. With the notation $f(x) = O_\infty(g(x))$ we understand that $\limsup_{x\to\infty} |f(x)/g(x)|$ exists and is finite, and similarly with $f(x) = O_0(g(x))$ we understand that $\limsup_{x\to 0} |f(x)/g(x)|$ exists and is finite.

We have the following theorem.

Theorem 2.6 *Assume that $p \geq 2$ (**F6**), then*

$$N^*_{K,\eta}(n) = O_\infty(n^2) \tag{19}$$

$$N^*_{n,K}(\eta) = O_0(\ln \eta) \tag{20}$$

and if $p > 2$ then

$$N^*_{n,\eta}(K) = O_0(\ln K/K^2) \tag{21}$$

∎

Proof. See Appendix A.2 ∎

Eq. (19) implies that the required number of samples in order to maintain the bound (17) with a given probability increases no more than quadratically with the model order. We also see that the number of samples increases at most as $\ln \eta$ as η tends to 0, i.e. the probability $1 - \eta$ tends to 1.

A similar analysis can also be performed for the case $1 < p < 2$, and it can be shown that in this case all of $N^*_{K,\eta}(n), N^*_{n,K}(\eta)$ and $N^*_{n,\eta}(K)$ grow faster than in the $p > 2$ case.

For the lower bound we could similarly study the number of samples required to guarantee with probability $1 - \eta$ that

$$V(\theta) \geq V_{N,L,t}(\theta) - K R_{N,L,t}(\theta)$$

and we will end up with the same sample complexities as given in Theorem 2.6.

Notice that the required number of data points to satisfy a certain upper bound bound (15) or (17) can be computed before any data are collected. This is not the case for the lower bound since it is dependent on the value of $R_{N,L,t}(\theta)$ evaluated on the observed data. In fact the derivation of the upper bound (15) for $p > 2$, makes use of the expected value of $R_{N,L,t}(\theta)$ and upper bounds it in terms of $V(\theta)$ using assumption **F6**.

2.3 Remarks

$V(\theta)$ is dependent on the input signal. The expected value $V(\theta)$ is given by

$$V(\theta) = \int (y(t) - B(q^{-1})u(t))^2 P(y,u)dydu$$

which can also be written as

$$V(\theta) = \int (y(t) - B(q^{-1})u(t))^2 P(y|u)P(u)dydu \tag{22}$$

Obviously the value of $V(\theta)$ is dependent on the probability distribution of the input signal, reflecting that some input signals are better for identification purposes than others. In the system identification literature it is common to assume the input signal to be persistently exciting, see e.g. Söderström and Stoica (1988), however such an assumption is not necessary here.

Bounds for a uniformly bounded criterion function. If, instead of assumption **F6** (bounded ratio of means), we assume a uniformly bounded criterion function (**F7**) we could use the following result (Vapnik (1982), Theorem 7.3)

Theorem 2.7 *Consider the framework* **F1**-**F5** *and* **F7**. *Let the number of independent observations l be larger than the VC dimension h of the class of FIR models which is bounded above by $2(n+1)$. Then with probability $1-\eta$*

$$\begin{aligned} V_{N,L,t}(\theta) - 2\tau\sqrt{\tfrac{h\ln(2l)-\ln(h!)-\ln(\eta/9)}{l}} &\leq V(\theta) \leq \\ V_{N,L,t}(\theta) + 2\tau\sqrt{\tfrac{h\ln(2l)-\ln(h!)-\ln(\eta/9)}{l}} & \end{aligned} \tag{23}$$

is valid simultaneously for all θ. V is given by (9), and $V_{N,L,t}$ is given by (10). Now τ is the constant by which the criterion function is uniformly bounded (assumption **F7***).* ∎

See also Haussler (1992) and the references therein for other bounds and further results in the case of uniformly bounded criterion functions.

Equation (23) can also be written as

$$|V(\theta) - V_{N,L,t}(\theta)| \leq 2\tau\sqrt{\frac{h\ln(2l) - \ln(h!) - \ln(\eta/9)}{l}}$$

and the sample complexity $N^*(n, K, \eta)$ required to guarantee with probability $1-\eta$ that

$$|V(\theta) - V_{N^*,L,t}(\theta)| < K$$

is equal to the sample complexities given in Theorem 2.6.

Confidence sets for θ. By combining the upper and lower bounds we can derive a confidence set for the parameter minimising the expected value of the criterion function. Denote the upper bound given in Theorem 2.3 by $V^{\mathrm{u}}(\theta)$ and the lower one given in Theorem 2.4 by $V^{\mathrm{lo}}(\theta)$. Let

$$\hat{\theta} = \arg\min_{\theta} V^{\mathrm{u}}(\theta)$$

be a parameter which minimises the upper bound on the expected value of the criterion function. Then the set

$$S = \{\theta |\ V^{\mathrm{lo}}(\theta) \leq V^{\mathrm{u}}(\hat{\theta})\}$$

is a confidence set for the parameter $\theta^* = \arg\min_\theta V(\theta)$ which minimises the expected value of the criterion function. We can assert that $\theta^* \in S$ with probability at least $1 - 2\eta$.

The reason for the probability $1-2\eta$ is as follows. The set of data points Ω_1, such that $V(\theta) > V^{\mathrm{u}}(\theta)$ has probability less than η. Similarly the set of data points Ω_2 such that $V(\theta) < V^{\mathrm{lo}}(\theta)$ has also probability less than η. S will therefore contain the minimising element for all data points which are not in $\Omega_1 \cup \Omega_2$, and hence θ^* belongs to S with probability at least $1-2\eta$.

3 ARX models

In this section we relax the assumption that $y(t+iL)$, $i = 0, 1, \ldots$ are iid, and we extend the results from the previous section to ARX models. The assumption on $y(t+iL)$ is relaxed by assuming that there exist good approximations $\bar{y}(t)$ of $y(t)$ and that the approximations $\bar{y}(t+iL), i = 0, 1, \ldots$ are iid, i.e.

F1a The observed input $u(t)$ is iid, and there exist approximations $\bar{y}(t)$ of the output $y(t)$ such that

$$|y(t) - \bar{y}(t)| < \delta_1 \quad \forall t \tag{24}$$

and $\bar{y}(t)$ and $\bar{y}(s)$ are iid for $|t-s| \geq L_1$ where L_1 is known. ∎

We now also consider the more general model class of ARX models, and **F2** is changed to

F2a Consider the class of ARX models given by

$$\begin{aligned} A(q^{-1})y(t) &= B(q^{-1})u(t) + e(t) \\ A(q^{-1}) &= 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a} \\ B(q^{-1}) &= b_1 q^{-1} + \cdots + b_{n_b} q^{-n_b} \end{aligned}$$

with predictors

$$\begin{aligned} \hat{y}(t,\theta) &= \phi^T(t)\theta \\ \phi^T(t) &= [-y(t-1), \ldots, -y(t-n_a), u(t-1), \ldots, u(t-n_b)] \\ \theta^T &= [a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}] \end{aligned}$$

∎

The special class of FIR models is obtained by fixing $A(q^{-1}) = 1$, equivalently $n_a = 0$.

In the previous section we did not make any assumptions on the parameter vector θ such that it should be bounded or belonging to a compact set. Here we introduce the following commonly used assumption

F8 Assumption

$$||\theta||_\infty \leq C_1 \tag{25}$$

where, for a vector $x = (x_1, \ldots, x_m)$ in R^m

$$||x||_\infty = \max_{1 \leq i \leq m} |x_i|$$

Let D_{C_1} be the set of parameter vectors such that (25) is satisfied. ∎

Assumptions of the type **F1a** or similar are common in the analysis of system identification methods, confer the exponential forgetting processes of Ljung (1978) and Hjalmarson (1993), and the L-mixing processes of Gerencsér (1989). The variables $\bar{y}(t)$ are not used in the computations, so it is sufficient to assume the existence of them.

Example. One class of systems that satisfy **F1a** is the class of asymptotically stable ARMAX systems. Assume the data have been generated by

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t)$$

where $A(q^{-1})$ and $C(q^{-1})$ are asymptotically stable filters, $u(t)$ and $e(t)$ are iid and bounded by a constant K. Then

$$y(t) = \frac{B(q^{-1})}{A(q^{-1})}u(t) + \frac{C(q^{-1})}{A(q^{-1})}e(t)$$

Let

$$\begin{aligned}\frac{B(q^{-1})}{A(q^{-1})} &= \sum_{i=1}^{\infty} \beta_i q^{-i} \\ \frac{C(q^{-1})}{A(q^{-1})} &= 1 + \sum_{i=1}^{\infty} \gamma_i q^{-i}\end{aligned}$$

Since $A(q^{-1})$ is asymptotically stable we can find an L_1 such that

$$\begin{aligned}\sum_{i=L_1}^{\infty} |\beta_i| &\leq \frac{\delta_1}{2K} \\ \sum_{i=L_1}^{\infty} |\gamma_i| &\leq \frac{\delta_1}{2K}\end{aligned}$$

Now let

$$\bar{y}(t) = \left(\sum_{i=1}^{L_1 - 1} \beta_i q^{-i} \right) u(t) + \left(1 + \sum_{i=1}^{L_1 - 1} \gamma_i q^{-i} \right) e(t)$$

Obviously $\bar{y}(t)$ and $\bar{y}(s)$ are independent for $|t - s| \geq L_1$, and

$$|y(t) - \bar{y}(t)| \leq \delta_1$$

which shows that data generated by asymptotically stable ARMAX systems with bounded iid input signals satisfy **F1a**. ∎

The prediction errors are given by

$$\epsilon(t, \theta) = y(t) - \phi^T(t)\theta$$

but under the current setup we can not find an L such that $\epsilon(t + iL, \theta), i = 0, 1, \ldots$ are iid. However, if we replace $\epsilon(t, \theta)$ with

$$\bar{\epsilon}(t, \theta) = \bar{y}(t) - \bar{\phi}^T(t)\theta$$

where

$$\bar{\phi}^T(t) = [-\bar{y}(t-1), \ldots, -\bar{y}(t - n_a), u(t-1), \ldots, u(t - n_b)]$$

we see that $\bar{\epsilon}(t + iL, \theta),\ i = 1, \ldots$ are iid for $L = L_1 + n_a$. The main idea is to derive bounds on

$$\bar{V}(\theta) = E\bar{\epsilon}^2(t, \theta)$$

in terms of

$$\bar{V}_{N,L,t}(\theta) = \frac{1}{l} \sum_{i=1}^{l} \bar{\epsilon}^2(t + iL, \theta)$$

using the theorems from section 2, and then using these bounds to obtain bounds on $V(\theta)$ in terms of $V_{N,L,t}(\theta)$. To obtain the bounds on $\bar{V}(\theta)$ the assumption of finite ratio of means is accordingly changed to

F6a Assumption. *Bounded ratio of means.* There exists a constant $p > 1$ such that

$$\frac{\left(E(\bar{y}(t) - \bar{\phi}^T(t)\theta)^{2p}\right)^{\frac{1}{p}}}{E(\bar{y}(t) - \bar{\phi}^T(t)\theta)^2} \leq \tau \quad \forall \theta \in D_{C_1} \tag{26}$$

∎

We then have the following theorem.

Theorem 3.1 *Consider the framework* **F1a**, **F2a**, **F3**, **F4**, **F5**, **F6a** *and* **F8**. *Let the number of data points $l = \lfloor N/L \rfloor = \lfloor N/(L_1 + n_a) \rfloor$ be larger than the VC dimension h of the ARX models which is bounded above by $2(n_a + n_b + 1)$. Then with probability $1 - \eta$*

$$V(\theta) \leq \delta_2^2 + 2\delta_2\sqrt{\tilde{V}_{N,L,t}(\theta)} + \tilde{V}_{N,L,t}(\theta) \tag{27}$$

is valid simultaneously for all $\theta \in D_{C_1}$. Here

$$\begin{aligned} \tilde{V}_{N,L,t}(\theta) &= \left[\frac{\delta_2^2 + 2\delta_2\sqrt{V_{N,L,t}(\theta)} + V_{N,L,t}(\theta)}{1 - T(h,\eta,l)}\right]_\infty & (28) \\ \delta_2 &= \delta_1 + \sqrt{n_a}\delta_1 C_1 & (29) \end{aligned}$$

where $T(h,\eta,l)$ and the interpretation of $[\cdot]_\infty$ is given in Theorem 2.3. δ_1 is given in **F1a**, *C_1 in* **F8** *and n_a is the order of the $A(q^{-1})$ polynomial (***F2a***). V is given by (9), and $V_{N,L,t}$ is given by (10).* ∎

Proof. See Appendix A.3. ∎

Before we give the lower bound, we modify the empirical functional $R_{N,L,t}(\theta)$ (16). By rewriting (16) we have

$$R_{N,L,t}(\theta) = \sum_{i=1}^{l} ||y(t_i) - \phi^T(t_i)\theta||^2 \frac{\sqrt{l-i+1} - \sqrt{l-i}}{\sqrt{l}}$$

where $y(t_i), \phi(t_i),\ i = 1, \ldots, l$ are the l observations sorted in increasing order in terms of the value of $||y(t_i) - \phi^T(t_i)\theta||^2$. We now modify $R_{N,L,t}(\theta)$ to

$$\tilde{R}_{N,L,t}(\theta) = \sum_{i=1}^{l} \left(\delta_2 + ||y(t_i) - \phi^T(t_i)\theta||\right)^2 \frac{\sqrt{l-i+1} - \sqrt{l-i}}{\sqrt{l}} \tag{30}$$

We now have the following theorem

Theorem 3.2 *Consider the framework* **F1a**, **F2a**, **F3**, **F4**, **F5**, **F6a** *and* **F8**. *Let the number of data points* $l = \lfloor N/L \rfloor = \lfloor N/(L_1 + n_a) \rfloor$ *be larger than the VC dimension* h *of the ARX models. Then with probability* $1 - \eta$

$$V(\theta) \geq \left(\left[\sqrt{V_{N,L,t}(\theta) - 2\delta_2 \sqrt{V_{N,L,t}(\theta)} - 2\sqrt{\frac{h \ln(2l) - \ln h! - \ln \frac{\eta}{12}}{l}} \tilde{R}_{N,L,t}(\theta)} - \delta_2 \right]_0 \right)^2$$

is valid for all $\theta \in D_{C_1}$ *simultaneously.* V *is given by (9),* $V_{N,L,t}$ *by (10) and* $\tilde{R}_{N,L,t}$ *by (30).* δ_2 *is given by (29) and* C_1 *in* **F8**. *The interpretation of* $[\cdot]_0$ *is given in Theorem 2.4.* ∎

Proof. See Appendix A.3. ∎

Remark. Assumption **F1a** can be relaxed so that we only require (24) to hold with probability $1 - \beta$. The difference in the above theorems is that the probability we assert the bounds with is reduced from $1 - \eta$ to $1 - \eta - \beta$. Assumption **F1a** together with **F8** will often lead to the conclusion that $y(t)$ and $\hat{y}(t, \theta)$ are uniformly bounded. The criterion function will then also be uniformly bounded, and upper and lower bounds for the ARX models can in this case be derived utilising Theorem 2.7.

3.1 Sample complexity

We now discuss the sample complexity of least squares identification of ARX models. The situation for ARX models is a bit different from FIR models, because of the presence of δ_2 in the bounds and that the overall number of samples is now given by

$$N = lL = l(L_1 + n_a)$$

where L_1 depends on δ_2 and n_a on the model order. We therefore modify the definition sample complexity as follows

Definition 3.3 *Let* $l_{\min}(n, K_1, \eta)$ *be the smallest integer which satisfies*

$$\frac{1}{1 - T(2(2n+1), \eta, l)} \leq 1 + K_1 \tag{31}$$

and let $L_1^*(n, K_2)$ *be the smallest integer such that*

$$\delta_1(L_1)(1 + \sqrt{n} C_1) \leq K_2 \tag{32}$$

We define the sample complexity of least squares identification of ARX models as

$$N^*(n, K_1, K_2, \eta) = l_{\min}(n, K_1, \eta)(L_1^*(n, K_2) + n)$$

$\blacksquare$

This definition implies that provided we have $N^*(n, K_1, K_2, \eta)$ samples, the upper bound in Theorem 3.1 is valid with probability $1-\eta$ for an nth order system and $1/(1-T(h,\eta,l)) \leq 1+K_1$ and $\delta_2 \leq K_2$.

As in the analysis of the FIR models we now introduce the variables

$$
\begin{aligned}
N^*_{K_1,K_2,\eta}(n) &= N^*(n, K_1, K_2, \eta) \\
N^*_{n,K_2,\eta}(K_1) &= N^*(n, K_1, K_2, \eta) \\
N^*_{n,K_1,\eta}(K_2) &= N^*(n, K_1, K_2, \eta) \\
N^*_{n,K_1,K_2}(\eta) &= N^*(n, K_1, K_2, \eta)
\end{aligned}
$$

Since $L_1^*(n, K_2)$ does not depend on η and K_1, we have from Theorem 2.6

Corollary 3.4 *Assume $p > 2$ (**F6a**) then*

$$N^*_{n,K_1,K_2}(\eta) = O_0(\ln \eta) \tag{33}$$
$$N^*_{n,K_2,\eta}(K_1) = O_0(\ln K_1/K_1^2) \tag{34}$$

$\blacksquare$

For $N^*_{K_1,K_2,\eta}(n)$ and $N^*_{n,K_1,\eta}(K_2)$ we introduce an extra assumption.

Theorem 3.5 *Assume that the underlying system is a finite dimensional asymptotically stable linear system. Then for $p > 2$ (**F6a**)*

$$N^*_{n,K_1,\eta}(K_2) = O_0(\ln K_2) \tag{35}$$
$$N^*_{n,K_1,K_2}(n) = O_\infty(n^2) \tag{36}$$

$\blacksquare$

Proof. We first consider $N^*_{n,K_1,\eta}(K_2)$. In this case $l_{\min}(n,\eta,K_1)$ and n are constant, so it is sufficient to consider how L_1 depends on δ_1. Assume that the magnitude of the system's slowest pole is a_s, then δ_1 is proportional to $a_s^{L_1}$, hence $a_s^{L_1} = CK_2$ which implies

$$L_1 = C\frac{\ln K_2}{\ln a_s} \tag{37}$$

and (35) is proved. Notice that (37) is dependent upon the dynamics of the system through the slowest pole a_s, confirming the intuition that we need a longer data sequence to identify a slow system than we need to identify a fast system.

We now consider $N^*_{n,K_1,K_2}(n)$. From Theorem 2.6 it follows that $l_{\min}$ is proportional to n for constant K_1 and η. From (32) it follows that to keep K_2 constant, δ_1 must be proportional to $1/\sqrt{n}$, and hence $a_s^{L_1} = C/\sqrt{n}$ and

$$L_1 = C\frac{-\ln\sqrt{n}}{\ln a_s} \tag{38}$$

Since $N^*(n, K_1, K_2, \eta) = l_{\min}(n, K_1, \eta)(L_1^*(n, K_2) + n)$ we have that

$$N^*_{K_1,K_2,\eta}(n) = O_\infty(n)(O_\infty(\ln n) + O_\infty(n)) = O_\infty(n^2) \tag{39}$$

and (36) is proved. Again we note that (38) is dependent on the system dynamics. ■

A similar analysis for Theorem 3.2 is also possible and leads to the same growth rates on the number of samples required.

4 General linear models

In this section we extend the previous results to the class of general linear models (Ljung(1987), Söderström and Stoica (1988))

$$y(t) = G(q^{-1}, \theta)u(t) + H(q^{-1}, \theta)e(t) \tag{40}$$

where $G(q^{-1}, \theta)$ and $H(q^{-1}, \theta)$ are asymptotically stable filters. $H(q^{-1}, \theta)$ is monic ($H(0, \theta) = 1$) and the inverse $H^{-1}(q^{-1}, \theta)$ is also asymptotically stable. $e(t)$ is a sequence of iid random variables. The optimal one step ahead predictor is given by

$$\begin{aligned}\hat{y}(t, \theta) &= H^{-1}(q^{-1}, \theta)G(q^{-1}, \theta)u(t) + (1 - H^{-1}(q^{-1}, \theta))y(t) \\ &= W_u(q^{-1}, \theta)u(t) + W_y(q^{-1}, \theta)y(t)\end{aligned}$$

where the filters

$$\begin{aligned}W_u(q^{-1}, \theta) &= H^{-1}(q^{-1}, \theta)G(q^{-1}, \theta) \\ W_y(q^{-1}, \theta) &= 1 - H^{-1}(q^{-1}, \theta)\end{aligned}$$

are asymptotically stable under the above assumptions.

For the general model class

$$A(q^{-1})y(t) = \frac{B(q^{-1})}{F(q^{-1})}u(t) + \frac{C(q^{-1})}{D(q^{-1})}e(t)$$

we have that

$$\begin{aligned} G(q^{-1},\theta) &= \frac{B(q^{-1})}{A(q^{-1})F(q^{-1})} \\ H(q^{-1},\theta) &= \frac{C(q^{-1})}{A(q^{-1})D(q^{-1})} \end{aligned}$$

and for most of the commonly used special model classes we have that $W_u(q^{-1},\theta)$ and $W_y(q^{-1},\theta)$ have infinite impulse responses. These classes include ARMAX models ($D(q^{-1}) = F(q^{-1}) = 1$), output error models ($A(q^{-1}) = C(q^{-1}) = D(q^{-1}) = 1$) and Box-Jenkins models ($A(q^{-1}) = 1$). The exception is the ARX models ($F(q^{-1}) = C(q^{-1}) = D(q^{-1}) = 1$) treated in the last section, for which $W_u(q^{-1},\theta)$ and $W_y(q^{-1},\theta)$ are FIR filters. So our main idea in this section is to approximate the class of general linear models by ARX models and then apply the results from section 3. We have the following change in the setup

F2b Consider the model class given by

$$y(t) = G(q^{-1},\theta)u(t) + H(q^{-1},\theta)e(t) \quad \theta \in D_\theta$$

with associated predictors

$$\hat{y}(t,\theta) = H^{-1}(q^{-1},\theta)G(q^{-1},\theta)u(t) + (1 - H^{-1}(q^{-1},\theta))y(t)$$

where $G(q^{-1},\theta)$ and $H(q^{-1},\theta)$ are uniformly asymptotically stable filters for $\theta \in D_\theta$. $H(q^{-1},\theta)$ is monic ($H(0,\theta) = 1$) and the inverse $H^{-1}(q^{-1},\theta)$ is also asymptotically stable. D_θ is a compact set. Assume that for every model $\theta \in D_\theta$ there exist an Mth order ARX model

$$\begin{aligned} \alpha(q^{-1})y(t) &= \beta(q^{-1})u(t) + e(t) \\ \alpha(q^{-1}) &= 1 + \alpha_1 q^{-1} + \cdots + \alpha_M q^{-m} \\ \beta(q^{-1}) &= \beta_1 q^{-1} + \cdots + \beta_M q^{-M} \end{aligned}$$

such that

$$||H^{-1}(q^{-1},\theta) - \alpha(q^{-1})||_1 \leq \delta_3 \tag{41}$$

$$||H^{-1}(q^{-1},\theta)G(q^{-1},\theta) - \beta(q^{-1})||_1 \leq \delta_3 \tag{42}$$

where

$$||P(q^{-1})||_1 = ||\sum_{i=0}^{\infty} p_i q^{-i}||_1 = \sum_{i=0}^{\infty} |p_i|$$

∎

Hence we approximate the transfer function $H(q^{-1},\theta)$ from the disturbance to the output by $1/\alpha(q^{-1})$ and the transfer function from the input to the output by

$\beta(q^{-1})/\alpha(q^{-1})$. Since $G(q^{-1},\theta)$ and $H^{-1}(q^{-1},\theta)$ are uniformly asymptotically stable, finite order ARX models can always be found such that (41) and (42) are satisfied.

Let

$$\begin{aligned}\bar{\theta} &= [\alpha_1,\ldots,\alpha_M,\beta_1,\ldots,\beta_M]^T \\ \phi(t) &= [-y(t-1),\ldots,-y(t-M),u(t-1),\ldots,u(t-M)]^T \\ \bar{\phi}(t) &= [-\bar{y}(t-1),\ldots,-\bar{y}(t-M),u(t-1),\ldots,u(t-M)]^T\end{aligned}$$

Assumptions **F8** and **F6a** are accordingly changed to

F8a

$$\bar{\theta} \in D_{C_1}$$

That is $||\bar{\theta}||_\infty \leq C_1$. ∎

and

F6b *Bounded ratio of means.* There exists a constant $p > 1$ such that

$$\frac{\left(E(\bar{y}(t)-\bar{\phi}^T(t)\bar{\theta})^{2p}\right)^{\frac{1}{p}}}{E(\bar{y}(t)-\bar{\phi}^T(t)\bar{\theta})^2} \leq \tau \quad \forall\bar{\theta}\in D_{C_1} \tag{43}$$

∎

Finally we make the additional assumption that $y(t)$ and $u(t)$ are bounded.

F9 There exists a constant C_2 such that

$$|y(t)| < C_2 \quad \forall t \quad \text{and} \quad |u(t)| < C_2 \quad \forall t$$

∎

We now have the following upper bound

Theorem 4.1 *Consider the framework* **F1a**, **F2b**, **F3**, **F4**, **F5**, **F6b**, **F8a** *and* **F9**. *Let the number of data points* $l = \lfloor N/L \rfloor = \lfloor N/(L_1+M) \rfloor$ *be larger than the VC dimension* h *of the approximate ARX models which is bounded above by* $2(M+1)$. *Then with probability* $1-\eta$

$$V(\theta) \leq \delta_4^2 + 2\delta_4\sqrt{\tilde{V}_{N,L,t}(\theta)} + \tilde{V}_{N,L,t}(\theta) \tag{44}$$

is valid simultaneously for all $\theta \in D_\theta$. Here

$$\tilde{V}_{N,L,t}(\theta) = \left[\frac{\delta_4^2 + 2\delta_4\sqrt{V_{N,L,t}(\theta)} + V_{N,L,t}(\theta)}{1 - T(h,\eta,l)}\right]_\infty \tag{45}$$

$$\delta_4 = \delta_1 + \sqrt{M}\delta_1 C_1 + 2\delta_3 C_2 \tag{46}$$

where $T(h,\eta,l)$ and the interpretation of $[\cdot]_\infty$ is given in Theorem 2.3. δ_1 is given in **F1a***, δ_3 in* **F2b***, C_1 in* **F8a***, C_2 in* **F9** *and M is the order of the approximate ARX models (***F2b***). V is given by (9), and $V_{N,L,t}$ is given by (10).* ∎

Proof. See Appendix A.4. ∎

For the lower bound we make the obvious modification of $\tilde{R}_{N,L,t}(\theta)$ and arrive at

$$\tilde{R}_{N,L,t}(\theta) = \sum_{i=1}^{l}\left(\delta_4 + ||(y(t_i) - \phi(t_i)\theta||\right)^2 \frac{\sqrt{l-i+1} - \sqrt{l-i}}{\sqrt{l}} \tag{47}$$

A lower bound is given by

Theorem 4.2 *Consider the framework* **F1a***,* **F2b***,* **F3***,* **F4***,* **F5***,* **F6b***,* **F8a** *and* **F9***. Let the number of data points $l = \lfloor N/L \rfloor = \lfloor N/(L_1 + M) \rfloor$ be larger than the VC dimension h of the approximate ARX models. Then with probability $1 - \eta$*

$$V(\theta) \geq \left(\left[\sqrt{V_{N,L,t}(\theta) - 2\delta_4\sqrt{V_{N,L,t}(\theta)} - 2\sqrt{\frac{h\ln(2l) - \ln h! - \ln\frac{\eta}{12}}{l}}\tilde{R}_{N,L,t}(\theta)} - \delta_4\right]_0\right)^2$$

is valid for all $\theta \in D_\theta$ simultaneously. V is given by (9), $V_{N,L,t}$ by (10) and $\tilde{R}_{N,L,T}$ by (47). δ_4 is given by (46), δ_3 in **F2b***, C_1 in* **F8a***, C_2 in* **F9** *and M is the order of the approximate ARX models (***F3b***). The interpretation of $[\cdot]_0$ is given in Theorem 2.4.* ∎

Proof. See Appendix A.4. ∎

Remark. In principle it is again straight forward to define and calculate the sample complexities. However, the computations become rather involved because of the levels of approximations, so we omit these calculations. ∎

5 Conclusions

In this paper we have analysed the finite sample properties of least squares identification in a stochastic framework. Using risk minimisation theory we have derived uniform probabilistic bounds on the discrepancy between the expected value of the squared prediction error $V(\theta)$, and the empirical value $V_{N,L,t}(\theta)$ evaluated on $l = \lfloor N/L \rfloor$ data points. The bounds are valid under very general conditions. In particular, no assumption is made requiring the true system to belong to the model class considered, and moreover, for the FIR and ARX cases the noise sequence is not assumed to be uniformly bounded. Further analysis showed that the sample complexity for stochastic least squares identification is quadratic in the model order for FIR and ARX models. It was also shown that the upper and lower bounds on $V(\theta)$ could be used to establish a confidence set for the parameter vector minimising $V(\theta)$.

Acknowledgement: This work was supported by the Australian Research Council.

References

[1] Dahleh M.A., T.V. Theodosopoulos, and J.N. Tsitsiklis (1993). "The sample complexity of worst-case identification of FIR linear systems," *System and Control Letters*, Vol. 20, pp. 157-166.

[2] Dasgupta D., and E. D. Sontag (1995). "Sample Complexity for Learning Recurrent Perceptron Mappings." Preprint.

[3] Gerencsér L. (1989). "On a class of mixing processes," *Stochastics*, Vol. 26, pp. 165-191.

[4] Haussler, D. (1992). "Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications." *Information and Computation.* Vol. 100, pp. 78-150.

[5] Hjalmarson, H. (1993). "Aspects on Incomplete Modeling in System Identification", Ph.D. thesis. Department of Electrical Engineering, Linköping University. Linköping, Sweden.

[6] Hole, A. (1995). "Vapnik-Chervonenkis generalization bounds for real valued neural networks." Preprint, Department of Mathematics, The University of Oslo, Norway.

[7] Kacewicz B. and M. Milanese (1992). "Optimal finite-sample experiment design in worst-case l_1 system identification" *Proc. of the 31st CDC*, Tucson, Arizona, US, pp. 56-61.

[8] Ljung, L. (1978). "Convergence Analysis of Parametric Identification Methods." *IEEE Trans. Automatic Control*, Vol. 23, pp. 770-783.

[9] Ljung, L. (1987). *System Identification - Theory for the User.* Prentice Hall.

[10] Ljung, L. (1993). "Perspectives on the process of identification" *Preprints of the 12th IFAC World Congress 1993*, Sydney, Australia, Vol. 5, pp. 197-205.

[11] Poolla K., and A. Tikku (1994). "On the Time Complexity of Worst-Case System Identification," *IEEE Trans. on Automatic Control*, Vol. 39, no. 5, pp. 944-950.

[12] Söderström, T. and P. Stoica (1988). *System Identification.* Prentice Hall.

[13] Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data.* Springer Verlag.

[14] Weyer, E. (1992). "System Identification in the Behavioural Framework," Ph. D. thesis. The Norwegian Institute of Technology, Trondheim, Norway.

[15] Weyer, E., R.C. Williamson and I.M.Y. Mareels (1995). "System identification in the behavioral framework via risk minimisation," In preparation.

[16] Zames, G., L. Lin, and L.Y. Wang (1994). "Fast Identification n-widths and Uncertainty Principles for LTI and Slowly Varying Systems," *IEEE Trans. on Automatic Control*, Vol. 39, no. 9, pp. 1827-1838.

A Proofs

A.1 Lemma 2.1

The VC dimension is defined as follows (Vapnik (1982)).

Definition A.1 *Let F be a set of functions from a set X to $\{0,1\}$. The VC dimension of F is the maximal size of a subset E of X such that for every $S \subseteq E$, there is $f_S \in F$ with $f_S(x) = 1$ if $x \in S$ and $f_S(x) = 0$ if $x \in E \setminus S$. If no maximal size set exists, the VC dimension is said to be infinite.*

Let G be a set of functions from a set Z to R. The VC dimension of G is the VC dimension of the set of indicator functions $\chi(g(z)+\beta)$ where $g \in G$ and $\beta \in R$, and $\chi(s) = 1$ for $s > 0$ and $\chi(s) = 0$ for $s \leq 0$. ∎

We now give a brief sketch of a proof of Lemma 2.1 based upon a proof in Hole (1995). The VC dimension of $(y(t) - \phi^T(t)\theta)^2$ is by definition equal to the VC dimension of the indicator functions $\chi((y(t)-\phi^T(t)\theta)^2+\beta)$. Let $\chi_0(s)$ denote the indicator function for which $\chi_0(s)=1$ for $s\geq 0$ and $\chi_0(s)=0$ for $s<0$. Then

$$\chi((y(t)-\phi^T(t)\theta)^2+\beta)=\chi(y(t)-\phi^T(t)\theta-\sqrt{|\beta|})\oplus(1-\chi_0(y(t)-\phi^T(t)\theta+\sqrt{|\beta|}))$$

where $\oplus$ denotes logical OR. This implies that the VC dimension of $(y(t)-\phi^T(t)\theta)^2$ is bounded above by twice the VC dimension of $y(t)-\phi^T(t)\theta$, see Hole (1995) for details. Moreover the VC dimension of $y(t)-\phi^T(t)\theta$ is equal to the pseudo dimension of $\phi^T(t)\theta+\beta$ which is equal to $M+1$ where M is the number of elements in the θ vector, see Haussler (1992) pp. 108-109. Hence the VC dimension of $(y(t)-\phi^T(t)\theta)^2$ is bounded above by $2(M+1)$.

A.2 Theorem 2.5

Since $N^*(n,K,\eta)=l_{\min}(n,K,\eta)\max(L,n)$ and L is constant, we can first consider $l_{\min}(n,K,\eta)$. Suppose K is fixed , this means that $T(2(n+1),\eta,l)$ must be kept below a constant C. (In the sequel C will denote a constant which may change from equation to equation.) For the case $p\geq 2$ this gives that

$$\sqrt{\frac{(2(n+1))\ln(2l)-\ln((2(n+1))!)-\ln(\eta/12)}{l}}\leq C \tag{48}$$

Assuming n constant we see that this will be achieved if $-\ln\eta/l_{\min}=C$ which proves (20).

Assume now that η is constant. Utilising that

$$m!>\left(\frac{m}{e}\right)^m$$

where e is the base of the natural logarithm, and substituting in (48) we have that

$$\sqrt{\frac{(2(n+1))(\ln\frac{2l}{2(n+1)}+1)-\ln(\eta/12)}{l}}\leq C \tag{49}$$

and we see that a constant level will be maintained provided $l_{\min}=Cn$, and hence (19) follows since $N^*(n,K,\eta)=l_{\min}(n,K,\eta)\max(L,n)$ and L is constant.

We now consider the case where n and η are constant. Then we have for $p>2$ that

$$\frac{1}{1-C\sqrt{\frac{(2(n+1))\ln(2l)-\ln((2(n+1))!)-\ln(\eta/12)}{l}}}\leq 1+K$$

which implies that

$$C\sqrt{\frac{(2(n+1))\ln(2l)-\ln((2(n+1))!)-\ln(\eta/12)}{l}}\leq\frac{K}{K+1}$$

which leads to the condition

$$(\frac{1}{K}+1)\sqrt{\frac{\ln l}{l}} = C$$

Hence

$$\frac{l_{\min}}{\ln l_{\min}} = O_0(1/K^2) \tag{50}$$

and it can be shown that (50) implies (21).

A.3 Theorem 3.1 and 3.2

Before we give the proofs, we derive some inequalities.

Let $||\cdot||$ denote the Euclidian norm on R. First we note that

$$||\bar{y}(t)-\bar{\phi}^T(t)\theta|| \le ||\bar{y}(t)-y(t)|| + ||y(t)-\phi^T(t)\theta|| + ||(\phi^T(t)-\bar{\phi}^T(t))\theta||$$

by the triangle inequality. We have that

$$\phi^T(t)-\bar{\phi}^T(t) = [\bar{y}(t-1)-y(t-1),\ldots,\bar{y}(t-n_a)-y(t-n_a),0,\ldots 0]$$

Using (24) and (25) we find that

$$||\bar{y}(t)-\bar{\phi}^T(t)\theta|| \le \delta_1+\sqrt{n_a}\delta_1 C_1 + ||y(t)-\phi^T(t)\theta|| = \delta_2 + ||y(t)-\phi^T(t)\theta|| \tag{51}$$

Similarly we find that

$$||y(t)-\phi^T(t)\theta|| \le \delta_2 + ||\bar{y}(t)-\bar{\phi}^T(t)\theta|| \tag{52}$$

From (51) it follows that

$$\begin{aligned}
\bar{V}_{N,L,t}(\theta) &= \frac{1}{l}\sum_{i=1}^{l} ||\bar{y}(t+iL)-\bar{\phi}^T(t+iL)\theta||^2 \\
&\le \frac{1}{l}\sum_{i=1}^{l}\Big(\delta_2 + ||y(t+iL)-\phi^T(t+iL)\theta||\Big)^2 \\
&\le \delta_2^2 + 2\delta_2\sqrt{V_{N,L,t}(\theta)} + V_{N,L,t}(\theta)
\end{aligned} \tag{53}$$

where we have used

$$\frac{1}{l}\sum_{i=1}^{l} ||y(t+iL)-\phi^T(t+iL)\theta|| \le \left(\frac{1}{l}\sum_{i=1}^{l} ||y(t+iL)-\phi^T(t+iL)\theta||^2\right)^{\frac{1}{2}}$$

which follows from Schwarz inequality on R^l. Similarly we can show that

$$V_{N,L,t}(\theta) \le \delta_2^2 + 2\delta_2\sqrt{\bar{V}_{N,L,t}(\theta)} + \bar{V}_{N,L,t}(\theta) \tag{54}$$

From (52) it follows that

$$\begin{aligned} V(\theta) &= E\epsilon^2(t,\theta) \le E\left(\delta_2 + ||\bar{y}(t+iL) - \bar{\phi}^T(t+iL)\theta||\right)^2 \\ &\le \delta_2^2 + 2\delta_2\sqrt{\bar{V}(\theta)} + \bar{V}(\theta) \end{aligned} \tag{55}$$

and similarly

$$\bar{V}(\theta) \le \delta_2^2 + 2\delta_2\sqrt{V(\theta)} + V(\theta) \tag{56}$$

Proof. (Theorem 3.1). From Lemma 2.1 it follows that the VC dimension of the ARX models is bounded above by $2(n_a + n_b + 1)$ (see example in section 2.1.1). Furthermore from Theorem 7.6 in Vapnik (1982) we have similar to the result in Theorem 2.3 that

$$\bar{V}(\theta) \le \left[\frac{\bar{V}_{N,L,t}(\theta)}{1 - T(h,\eta,l)}\right]_\infty \tag{57}$$

By substituting (53) in (57) we obtain

$$\bar{V}(\theta) \le \left[\frac{\delta_2^2 + 2\delta_2\sqrt{V_{N,L,t}(\theta)} + V_{N,L,t}(\theta)}{1 - T(h,\eta,l)}\right]_\infty \tag{58}$$

and we recognise the right hand side as $\tilde{V}_{N,L,t}(\theta)$, and the inequality in the theorem follows by substituting (58) in (55). ∎

Proof. (Theorem 3.2) From Theorem 2.4 it follows that

$$\bar{V}(\theta) \ge \bar{V}_{N,L,t}(\theta) - 2\sqrt{\frac{h\ln(2l) - \ln h! - \ln\frac{\eta}{12}}{l}}\bar{R}_{N,L,t}(\theta) \tag{59}$$

where (by rewriting (16)) $\bar{R}_{N,L,t}(\theta)$ is given by

$$\bar{R}_{N,L,t}(\theta) = \sum_{i=1}^{l} ||(\bar{y}(t_i) - \bar{\phi}(t_i)\theta||^2 \frac{\sqrt{l-i+1} - \sqrt{l-i}}{\sqrt{l}}$$

From (51) and (30) we have that

$$\bar{R}_{N,L,t}(\theta) \le \tilde{R}_{N,L,t}(\theta) \tag{60}$$

Moreover from (52) it follows that

$$||y(t) - \phi^T(t)\theta|| - \delta_2 \le ||\bar{y}(t) - \bar{\phi}^T(t)\theta|| \tag{61}$$

and hence

$$V_{N,L,t}(\theta) - 2\delta_2\sqrt{V_{N,L,t}(\theta)} \le \bar{V}_{N,L,t}(\theta) \tag{62}$$

where we on the left hand side have omitted the term δ_2^2 in case the left hand side of (61) is negative. Rewriting (56) we have

$$\bar{V}(\theta) \leq \left(\delta_2 + \sqrt{V(\theta)}\right)^2 \tag{63}$$

By substituting (60), (62) and (63) in (59) we obtain

$$\left(\delta_2 + \sqrt{V(\theta)}\right)^2 \geq V_{N,L,t}(\theta) - 2\delta_2\sqrt{V_{N,L,t}(\theta)} - 2\sqrt{\frac{h\ln(2l) - \ln h! - \ln\frac{\eta}{12}}{l}\tilde{R}_{N,L,t}(\theta)} \tag{64}$$

Then the inequality in the theorem is obtained by taking the square root of both sides of (64), moving δ_2 to the right, and squaring both sides again.

∎

A.4 Theorem 4.1 and 4.2

Proof. (Theorem 4.1) We have that

$$\begin{aligned} ||\bar{y}(t) - \bar{\phi}^T(t)\bar{\theta}|| &\leq ||\bar{y}(t) - y(t)|| + ||y(t) - \hat{y}(t,\theta)|| + \\ &\quad ||\hat{y}(t,\theta) - \phi^T(t)\bar{\theta}|| + ||(\phi^T(t) - \bar{\phi}^T(t))\bar{\theta}|| \\ &\leq \delta_1 + 2\delta_3 C_2 + \sqrt{M}\delta_1 C_1 + ||y(t) - \hat{y}(t,\theta)|| = \delta_4 + ||y(t) - \hat{y}(t,\theta)|| \end{aligned}$$

Similarly

$$||y(t) - \hat{y}(t,\theta)|| \leq \delta_4 + ||\bar{y}(t) - \bar{\phi}^T(t)\bar{\theta}||$$

and the rest of the proof follows the proof of Theorem 3.1 with δ_4 replacing δ_2. ∎

Proof. (Theorem 4.2) Similar to the proof of Theorem 3.2 ∎