On Convergence Proofs in System Identification – A General Principle using ideas from Learning Theory

Tor A. Johansen [a] Erik Weyer [b]

[a] *Department of Engineering Cybernetics, Norwegian University of Science and Technology, N-7034 Trondheim, Norway. Email: Tor.Arne.Johansen@itk.ntnu.no*

[b] *Department of Eletrical and Electronic Engineering, The University of Melbourne, Parkville VIC 3052, Australia. Email: e.weyer@ee.mu.oz.au*

It is shown that learning theory offers convergence analysis tools that are useful in system identification problems. They allow analysis in a parameter-free context, which elevates the analysis from parameter sets to model sets and from parameter identification to model identification. When a paramterization is eventually introduced, this leads to alternative assumptions on the parameterization and parameter set. Moreover, structural identification can be analyzed within the same framework. Another advantage is that the proofs are technically and conceptually simple.

Key words: System Identification, Learning Theory, Non-linear Systems, Convergence, Stochastic Processes, Model Sets, Parameterization.

1 Introduction

Learning theory (Vapnik 1982, Pollard 1984, Valiant 1984, Haussler 1992, Vapnik 1995) and system identification theory (Ljung 1987, Söderström and Stoica 1988) have developed independently. Recently, there has been interest in exploring and exploiting the relationship between these areas, hoping for a synergy effect. Some recent results include (Ljung 1996, Campi and Kumar 1996, Weyer *et al.* 1996). In particular, Ljung (1996) poses the following challenge: "...the really interesting question is whether learning theory can help us create a theory of system identification, void of parameters, and whether this would actually change the way we estimate our models". The present work is a step towards a positive answer to the first part of this question, illustrating how ideas and analysis tools from learning theory

(Pollard 1984) give convergence results in a parameter-free setup. A condition related to uniform boundedness of the predictors generated by the model set is introduced in addition to the standard assumption of exponential stability. These conditions can be seen as a generalization from linear to nonlinear predictors of the "uniform stability" assumption made in Ljung (1987) (in this context "uniform" is with respect to the model set). When a parameterization is introduced the condition of compactness of the parameter set found in (Ljung 1978) can be relaxed, and no explicit assumption on the smoothness of the parameterization are necessary. Instead, direct assumptions on the boundedness of the predictor and data are presented. Furthermore, the paramter-free setup allows structural identification to be analysed within the same framework.

The adopted principle for convergence analysis is based on a result from (Pollard 1984) that establishes a strong uniform law of large numbers (ULLN) under two basic conditions: First, the existence of a strong law of large numbers (LLN), and second, a constraint on the complexity of the model set. The LLN is generally established by making assumptions on the correlation structure in the data generated by the system. The main problem is then to check if the model set satisfies the complexity-constraint. The convergence properties of the identification algortihm then follow immediately from the ULLN.

2 A Principle for Convergence Proofs

Let $\mathcal{M}$ be a set of models, possibly uncountable. Suppose a bounded sequence of N input/output observations from a dynamical system is available; $Z_N = (z_1, z_2, ..., z_N)$ where $z_t = (u_t, y_t)$, $u_t \in R^r$ is the input vector, and $y_t \in R^m$ is the output vector, at time t. We will introduce a stochastic framework that allows us to deal with quasi-stationary stochastic processes that may contain deterministic components and sequential correlations, similar to (Ljung 1987). In the following, let E denote expectation with respect to input/output observations, and define the operator $\overline{E}$ by $\overline{E}s_t = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} Es_t$ for any vector sequence s.

Suppose that for each model $M \in \mathcal{M}$ we are able to make a one-step-ahead prediction $\hat{y}_t(M, Z_{t-1})$ of the output y_t on the basis of the data sub-sequence Z_{t-1}. The one-step-ahead prediction error using the model M is defined by $\varepsilon_t(M, Z_t) = y_t - \hat{y}_t(M, Z_{t-1})$. In the following, we may let the dependence on Z_{t-1} be left implicit when this is more convenient, and use the notation $\hat{y}_t(M)$ and $\varepsilon_t(M)$. For later use, we introduce the definition of the squared prediction error

$$v_t(M) = ||\varepsilon_t(M)||^2$$

where $||\cdot||$ denotes Euclidean norm. Furthermore, let

$$\overline{V}(M) = \overline{E} v_t(M) = \overline{E}\, ||\varepsilon_t(M)||^2 \tag{1}$$

denote the expected value of the squared prediction error. The ultimate goal is to identify a model in $\mathcal{M}$ that minimizes $\overline{V}$, and we will in particular study the properties of the identification method as N, the number of data points approaches infinity [1].

Minimization of the expected value of the squared prediction error $\overline{V}$ will in many cases reflect the desired properties of the model, and $\overline{V}$ is therefore often suited as a criterion for model identification. However, the underlying probability measure is almost always unknown, and consequently, (1) cannot be computed. Instead we consider the average of the empirical squared prediction error

$$V_N(M) = \frac{1}{N}\sum_{t=1}^{N} ||\varepsilon_t(M)||^2 \tag{2}$$

and assume that V_N has a global minimum:

Assumption A1. For every N the following global minimum exists with probability one

$$\hat{M}_N = \arg\min_{M\in\mathcal{M}} V_N(M)$$

$\square$

Notice that we do not assume that the minimum is unique, but use the convention that if more than one model in $\mathcal{M}$ minimizes V_N, then $\hat{M}_N$ is an arbitrary minimum.

For a given model M a convergence result such as

$$\left|\overline{V}(M) - V_N(M)\right| \to 0 \text{ w.p. } 1 \quad \text{as } N \to \infty \tag{3}$$

is referred to as a strong law of large numbers (LLN). However, in order to argue that minimization of V_N will asymptotically give a model with similar

[1] Notice that the results in this section can be easily reformulated for multi-step predictors as well.

prediction performance as minimization of $\overline{V}$, a stronger condition is needed, namely a strong uniform law of large numbers (ULLN):

$$\sup_{M \in \mathcal{M}} \left| \overline{V}(M) - V_N(M) \right| \to 0 \text{ w.p. } 1 \quad \text{as } N \to \infty \tag{4}$$

This property is often referred to as strong uniform convergence. Its main implication is that V_N will be a close approximation to $\overline{V}$ for all models in $\mathcal{M}$, for large N. Hence, minimization of V_N and $\overline{V}$ will almost surely give a model with similar prediction performance, asymptotically.

It follows immediately from the ULLN and assumption A1 that the expected value of the squared prediction error for the identified model converges towards the best possible one (in the sense that $\overline{V}$ will be as small as possible):

$$\overline{V}(\hat{M}_N) \to \inf_{M \in \mathcal{M}} \overline{V}(M) \quad \text{w.p. } 1 \text{ as } N \to \infty \tag{5}$$

Let

$$\mathcal{M}^\star = \left\{ M^\star \in \mathcal{M} \,\middle|\, \overline{V}(M^\star) = \inf_{M \in \mathcal{M}} \overline{V}(M) \right\}$$

and suppose $\mathcal{M}^\star$ is non-empty. Clearly, $\mathcal{M}^\star$ will depend on several factors, including the model set, the true system, and the operating conditions. Conditions under which $\mathcal{M}^\star$ will contain a single model are typically related to identifiability of the model set and operating conditions that give persistently exciting data, e.g. (Ljung 1978, Ljung 1987, Söderström and Stoica 1988). Hence, the convergence result (5) can be used to derive a consistency result, i.e. convergence of $\hat{M}_N$ towards $\mathcal{M}^\star$ in some sense, under additional assumptions.

In order to prove a ULLN, it is necessary to limit the complexity of the model set. In the standard system identification context, a commonly used tool is to assume continuous parameterization and compactness of the parameter set (Ljung 1978) or "uniform stability" ("uniform" with respect to the model set) of the set of predictors (Ljung 1987). However, learning theory offers a wide range of more sophisticated complexity measures, including ϵ-nets, covering numbers and the Vapnik-Chervonenkis dimension (Vapnik and Chervonenkis 1971, Vapnik 1982, Pollard 1984), to mention a few. In this work, we will apply a basic convergence result of (Pollard 1984) that we later will relate to the existence of a finite ϵ-net:

Theorem 1 (Uniform Law of Large Numbers.) *Let*

$$\mathcal{H} = \{(v_t(M)) \mid M \in \mathcal{M}\}$$

be the set of squared prediction error sequences generated by the model set $\mathcal{M}$. Suppose for each $\epsilon > 0$ there exists a finite set of models $\mathcal{M}_\epsilon^0$ that generates a finite set of squared prediction error sequences

$$\mathcal{H}_\epsilon = \left\{ (v_t(M)) \mid v_t(M) = ||y_t - \hat{y}_t(M, Z^{t-1})||^2, \; M \in \mathcal{M}_\epsilon^0 \right\}$$

for which a law of large numbers hold, and $\mathcal{H}_\epsilon$ contains lower and upper approximations $v^{\epsilon,L} \in \mathcal{H}_\epsilon$ and $v^{\epsilon,U} \in \mathcal{H}_\epsilon$ to each $v \in \mathcal{H}$ such that

$$v_t^{\epsilon,L} \leq v_t \leq v_t^{\epsilon,U} \quad \forall t \;\; \textit{and} \;\; \overline{E}\left(v_t^{\epsilon,U} - v_t^{\epsilon,L}\right) < \epsilon \tag{6}$$

Then

$$\sup_{M \in \mathcal{M}} \left| \overline{V}(M) - V_N(M) \right| \rightarrow 0 \; w.p. \; 1 \; as \;\; N \rightarrow \infty \tag{7}$$

The convergence proof principle is now apparent: To establish a LLN and to check that there exist a finite set of models $\mathcal{M}_\epsilon^0$ such that (6) holds. Eq. (6) is, however, not a very transparent and easily verifiable condition. In the next section we therefore introduce a more convenient set of conditions that is sufficient for the conditions of Theorem 1 to hold.

So far, parameterization of the model set is not mentioned. We will continue this development in a parameter-free system identification context in section 3. Parameter and structure identification are discussed in section 4.

3 Parameter-free System Identification

This section describes how the assumptions of Theorem 1 can be satisfied in a parameter-free system identification setup. As an alternative to eq. (6) we introduce another characterization of the complexity of the model set, namely finite ϵ-nets (Vapnik 1982). Thereafter we introduce the class of exponentially forgetting sequences which allows the establishment of a LLN, with some additional stability conditions on the predictor. It will be shown that this is sufficient for Theorem 1 to hold. We start by assuming the existence of a finite ϵ-net:

Assumption M1. For any $\epsilon > 0$ the model set $\mathcal{M}$ can be covered by a finite ϵ-net with $N(\epsilon)$ elements, i.e. there exists a finite set of models $\mathcal{M}_\epsilon = \left\{ M_1, M_2, ..., M_{N(\epsilon)} \right\}$ (not necessarily a subset of $\mathcal{M}$) such that for any $M \in \mathcal{M}$ there exists an $M_i(M) \in \mathcal{M}_\epsilon$ with the property

$$\sup_{t, Z_{t-1}} ||\hat{y}_t(M, Z_{t-1}) - \hat{y}_t(M_i(M), Z_{t-1})|| \leq \epsilon \tag{8}$$

Please apologize for possible confusion by using ϵ and ε in two different meanings.

□

This assumption states a key characterization of the model and data sets. It essentially limits the complexity of the model set, in the sense that there must exist a finite number of good approximations over the data range to the possibly uncountable number of models in the model set.

Next, we want to establish a LLN. This requires a constraint on the sequential correlations in the data sequences and a constraint on the predictors' sensitivity to past data. The quite general framework of (Ljung 1978, Hjalmarsson 1993) based on exponentially forgetting sequences is adopted. Let the γ-norm of a vector $r \in R^n$ be defined by $||r||_\gamma = (|r_1|^\gamma + ... + |r_n|^\gamma)^{1/\gamma}$ for all $\gamma \geq 1$.

Definition (Exponential forgetting). A sequence (x_t) is said to be exponentially forgetting of order γ if there exists a doubly indexed sequence (x_t^s) where x_t^s is independent of the subsequences $(x_0, x_1, ..., x_s)$ and $(x_p^p, x_{p+1}^p, ..., x_s^p)$ for all $p \leq s$, $x_t^t = 0$, and for every $t \geq s$

$$E\, ||x_t - x_t^s||_\gamma^\gamma \leq C\lambda^{t-s} \tag{9}$$

for some constants $C \geq 0$ and $\lambda \in (0, 1)$.

□

This means that it must be possible to approximate the sequence (x_t) with a sequence (x_t^s) that is independent of the remote past. Moreover, the approximation must be so good that the "remote past is forgotten at an exponential rate". It is assumed the data are generated by a system with this property:

Assumption S1. The sequences (u_t) and (y_t) are exponentially forgetting of order 4.

□

On $\mathcal{M}$ and the ϵ-net $\mathcal{M}_\epsilon$ an additional assumption is needed to ensure that the squared prediction error sequence inherits the exponentially forgetting property when S1 holds.

Assumption M2. Define the sequence $v(M) = (v_t(M)) = (||\varepsilon_t(M)||^2)$, $M \in \mathcal{M} \cup \mathcal{M}_\epsilon$ and suppose

(i) The function $\hat{y}_t(M, Z_{t-1})$ is differentiable with respect to Z_{t-1} for all $M \in \mathcal{M} \cup \mathcal{M}_\epsilon$, and for all $t, k \geq 0$ there exists a convex function $g_{t,k}$ such that for all models $M \in \mathcal{M} \cup \mathcal{M}_\epsilon$

$$\left|\left|\frac{\partial}{\partial z_{t-k}} v_t(M, z_t, z_{t-1}, ..., z_1)\right|\right| \leq g_{t,k}(z_t, z_{t-1}, ..., z_1)$$

(ii) There exist constants C and $\lambda \in (0,1)$ such that for all $t \geq 0$ the set of functions $\{g_{t,k} \mid k \geq 0\}$ has the property

$$E||z_t||^4 < \infty \quad \text{for all } t \text{ implies } E(g_{t,k}(z_t, ..., z_1))^4 \leq C\lambda^k$$

(iii) Moreover, $(v_t(M, 0, 0, ..., 0))$ is a bounded sequence for all $M \in \mathcal{M} \cup \mathcal{M}_\epsilon$.

□

This assumption requires that the sensitivity of the prediction to past data decays exponentially, which basically means that the predictor must be exponentially stable. We now have the following result from (Hjalmarsson 1993), pp. 46-48.

Lemma 1 *Suppose Assumptions S1 and M2 hold, then for all $M \in \mathcal{M} \cup \mathcal{M}_\epsilon$ the sequences $(v_t(M))$ are exponentially forgetting of order 2. Moreover the sequences $(\sup_{M \in \mathcal{M}_\epsilon} v_t(M))$ and $(\sup_{M \in \mathcal{M}} v_t(M))$ are also exponentially forgetting of order 2.*

□

Exponentially forgetting sequences are closely related to mixing processes, and in particular L-mixing processes, (Gerencsér 1989). However, as noted in (Hjalmarsson 1993), an equivalent to Lemma 1 appears to be more difficult to establish for such processes.

Exponential forgetting of order 2 is sufficient for a strong law of large numbers to hold:

Lemma 2 (Law of Large Numbers.) *Let $M \in \mathcal{M}_\epsilon$ be given, and suppose the scalar sequence $v(M)$ is exponentially forgetting of order 2. Then*

$$\left|\overline{V}(M) - V_N(M)\right| \to 0 \quad \textit{w.p. 1 as} \quad N \to \infty$$

Proof. Follows from a somewhat stronger result in (Hjalmarsson 1993), p. 66.

□

Next, we state a main result of the report, which relates assumption M1 to the conditions of Theorem 1:

Theorem 2 *(Uniform Law of Large Numbers) Suppose S1, M1 and M2 hold, then the ULLN (4) holds.*

Proof. The proof is inspired by a result in (Weyer 1992). Define the set of squared prediction error sequences $\mathcal{H} = \{v(M) \mid M \in \mathcal{M}\}$ generated by the model set. The proof is basically to show that $\mathcal{H}$ satisfies (6).

Pick an arbitrary $\epsilon > 0$ and $M \in \mathcal{M}$, and let $M_i(M) \in \mathcal{M}_{\epsilon'}$ denote the model in the ϵ'-net $\mathcal{M}_{\epsilon'}$ that is closest to M (in the sense of (8)), where ϵ' is yet unspecified. Now

$$v_t(M_i(M)) - 2\epsilon'\sqrt{v_t(M_i(M))} \leq v_t(M) \;\leq\; \left(\sqrt{v_t(M_i(M))} + \epsilon'\right)^2 \tag{10}$$

and since $\sqrt{x} \leq 1 + x$ for any $x \geq 0$, lower and upper approximations to the sequence $v(M)$ are given by

$$\begin{aligned} v_t^{\epsilon,L} \;&=\; v_t(M_i(M)) - 2\epsilon'(1 + v_t(M_i(M))) \leq v_t(M) \\ &\qquad\qquad\qquad \leq\; v_t(M_i(M)) + 2\epsilon'(1 + v_t(M_i(M))) + (\epsilon')^2 \;=\; v_t^{\epsilon,U} \end{aligned}$$

Taking the expectation, we get

$$\overline{E}\left(v_t^{\epsilon,U} - v_t^{\epsilon,L}\right) = (\epsilon')^2 + 4\epsilon' + 4\epsilon'\overline{V}(M_i(M)) \tag{11}$$

From Lemma 1 it follows that $(\sup_{M\in\mathcal{M}} v_t(M))$ is exponentially forgetting of order 2. Hence there exists a K such that

$$\sup_{M\in\mathcal{M}} \overline{V}(M) \leq K \tag{12}$$

and it therefore follows that

$$\max_{M'\in\mathcal{M}_{\epsilon'}} \overline{V}(M') \leq K + (\epsilon')^2 + 2\epsilon'(1 + K) \tag{13}$$

Choosing ϵ' sufficiently small thus ensures

$$\overline{E}\left(v_t^{\epsilon,U} - v_t^{\epsilon,L}\right) < \epsilon \tag{14}$$

Hence, the conditions of Theorem 1 are fulfilled, and the ULLN follows.

□

A finite ϵ-net covering the "predictors' output" is required by assumption 1. This is in contrast to the assumption of a compact parameter set made in (Ljung 1978), which guarantees the existence of a finite ϵ-net that covers the parameter set. These apparently different assumptions are related since the additional assumption made in (Ljung 1978) of uniform continuity of the predictors in the parameters can be applied to show that a finite ϵ-net for the parameter set is sufficient to construct a finite ϵ-net covering the "predictors' output", cf. section 4. By using an ϵ-net covering the "predictors' output" we do not need to assume that the predictors depend smoothly on the parameters. In fact, at this point we have not even introduced a parameterization of the model set.

In this section we continue with a parameter-free characterization of a set of predictors that guarantees the existence of a finite ϵ-net, as required by assumption M1. It is interesting to observe that the complexity constraint on the model set imposed by assumption M1 reappears as uniform boundedness of the data and the predictor, together with a regularity condition.

Theorem 3 *Suppose*

(i) There exits a compact set $\mathcal{Z}$ such that $z_t \in \mathcal{Z}$ for all t.

(ii) There exists a constant K such that

$$\sup_{t,Z_{t-1},M} ||\hat{y}_t(M, Z_{t-1})|| \leq K \tag{15}$$

(iii) For every $k \geq 1$ there exists a uniformly continuous (in $\mathcal{Z}^k$) function $f_k : \mathcal{M} \times \mathcal{Z}^k \to R^m$ with the property

$$\sup_{t,Z_{t-1},M} ||\hat{y}_t(M, z_{t-1}, ..., z_1) - f_k(M, z_{t-1}, ..., z_{t-k})|| \to 0$$

as $k \to \infty$.

Then $\mathcal{M}$ can be covered by a finite ϵ-net.

Proof. Let $\epsilon > 0$ be arbitrary. Choose k so large that

$$\sup_{t,Z_{t-1},M} ||\hat{y}_t(M, z_{t-1}, ..., z_1) - f_k(M, z_{t-1}, ..., z_{t-k})|| \leq \epsilon/2$$

Provided the set of functions

$$\mathcal{F}_k = \{f_k(M, \cdot) : \mathcal{Z}^k \to R^m \mid M \in \mathcal{M}\} \tag{16}$$

can be covered with a finite $\epsilon/2$-net, this $\epsilon/2$-net will also be an ϵ-net for $\mathcal{M}$. From the uniform continuity of f_k and the compactness of $\mathcal{Z}$ there exists a finite number of sets $\mathcal{Z}_1, \mathcal{Z}_2, ..., \mathcal{Z}_n$

$$\mathcal{Z} \subset \bigcup_i \mathcal{Z}_i \quad \text{and} \quad \mathcal{Z}_i \cap \mathcal{Z}_j = \emptyset \quad \text{for } i \neq j \tag{17}$$

such that

$$Z^1, Z^2 \in \mathcal{Z}_i \quad \text{implies} \quad \sup_{M \in \mathcal{M}} ||f_k(M, Z^1) - f_k(M, Z^2)|| < \epsilon/4 \tag{18}$$

Let μ_i be the characteristic function for $\mathcal{Z}_i$. From (15) and (18) it is clear that there exists indices $\{l(M,i) \mid i = 1, 2, ..., n\}$ such that

$$\sup_{M \in \mathcal{M}} \sup_{Z \in \mathcal{Z}} \left|\left| f_k(M, Z) - \sum_{i=1}^{n} \beta_{l(M,i)} \mu_i(Z) \right|\right| < \epsilon/4 \tag{19}$$

where the constant vectors $\beta_{l(M,i)} \in R^m$ are chosen from a finite set of values, determined by K and ϵ. Hence, the approximation $\sum_{i=1}^{n} \beta_{l(M,i)} \mu_i(Z)$ can be used to define a finite $\epsilon/2$-net for $\mathcal{F}_k$.

Remark. It is required that the ϵ-net for $\mathcal{M}$ consists of predictors that are smooth functions of past data, cf. M2. This may be achieved by replacing the μ_i-functions by smoothed versions. It should be mentioned that the use of a piecewise constant approximation in the proof is only one possibility, and several other approximators are expected to work as well.

□

The third assumption requires that the dependence of the predictor on past data must vanish asymptotically. Roughly speaking, this means that the predictor must be asymptotically stable. Moreover, the predictor must be time-invariant.

We would like to stress that Theorem 3 is only an example of sufficient conditions for the existence of an ϵ-net that covers $\mathcal{M}$, and weaker conditions may very well exist. In particular the first and second assumptions are more restrictive compared to (Ljung 1978, Ljung 1987). It must also be emphasized that it is sufficient to prove the existence of an ϵ-net. However, as we have seen, the proofs are typically constructive.

In (Ljung 1987), only linear systems are considered. The main assumption on the set of linear predictors

$$\hat{y}_t(M, Z^{t-1}) = \sum_{k=1}^{\infty} g_k(M) y_{t-k} + \sum_{k=1}^{\infty} h_k(M) u_{t-k} \tag{20}$$

is "uniform stability", i.e. the impulse response coefficients $h_k(M)$ and $g_k(M)$ satisfy

$$\sup_{M \in \mathcal{M}} |h_k(M)| \leq h_k$$
$$\sup_{M \in \mathcal{M}} |g_k(M)| \leq g_k$$

for some sequences h and g that satisfies $\sum_{k=1}^{\infty} |h_k| < \infty$ and $\sum_{k=1}^{\infty} |g_k| < \infty$. It is straightforward to see that a set of uniformly stable predictors satisfies the assumptions of the present paper. Hence, the present result provides a natural generalization of the convergence result in (Ljung 1987) to a class of exponentially stable non-linear systems and models. Note however, that the assumptions on the data set used in this paper are different from the ones used in (Ljung 1987). The present conditions may be more restrictive, particularly if we assume, as in Theorem 3, that z_t belongs to a compact set.

4 Parameter and Structure Identification

Parameterization of the model set is introduced in this section. The model set is decomposed into structural and parametric levels, which allows structure identification to be studied in the same framework. Furthermore, we will show how the well known convergence result in (Ljung 1978) now follows directly.

Let $\mathcal{S}$ be a set of model structures. A model structure $S \in \mathcal{S}$ is a set of possibly non-linear equations, which may be algebraic, difference, differential, or a combination. The equations may contain some unknown parameters θ that belong to a parameter set Θ_S, which is a subset of a finite-dimensional Euclidean vector space, by assumption. This defines the model set

$$\mathcal{M} = \{M = (S, \theta) \mid S \in \mathcal{S}, \theta \in \Theta_S\}$$

4.1 Parameter Identification

The parameterization allows us to derive an alternative set of conditions that guarantees that assumption M1 hold. It appears that the complexity constraint on the model set that is implicit in assumption M1 can be related to "boundedness" of the parameter and structure sets, together with a regularity condition:

Theorem 4 *Suppose*

(i) $\mathcal{S}$ *contains only a finite number of model structures.*
(ii) For every $S \in \mathcal{S}$ *the parameter set* Θ_S *is compact.*
(iii) For every $S \in \mathcal{S}$ *the predictor* $\hat{y}_t(S, \theta, Z_{t-1})$ *is uniformly continuous in* θ *on* Θ_S*, i.e. for every* $\epsilon > 0$ *there exists a* $\delta > 0$ *that depends only on* ϵ*, with the property that*

$$||\theta_1 - \theta_2|| < \delta \textit{ implies} \sup_{t, Z_{t-1}} ||\hat{y}_t(S, \theta_1, Z_{t-1}) - \hat{y}_t(S, \theta_2, Z_{t-1})|| < \epsilon \quad (21)$$

Then $\mathcal{M}$ *can be covered by a finite* ϵ*-net.*

Proof. Let $S \in \mathcal{S}$ be arbitrary. Since Θ_S is compact, it can be covered by a finite number of δ-balls. The parameter vectors at the centre of these δ-balls define an ϵ-net for the model set $\mathcal{M}_S = \{M = (S, \theta) \mid \theta \in \Theta_S\}$, cf. (21). Since $\mathcal{S}$ contains only a finite number of model structures, the result follows.

□

The uniform law of large numbers that underlies the convergence result of (Ljung 1978), which applies to a fixed model structure and some slightly different other assumptions, follows immediately from Theorems 2 and 4.

Corollary 1 *Suppose S1 and M2 hold. Consider a fixed model structure* $S \in \mathcal{S}$*, suppose* Θ_S *is compact and* $\hat{y}_t(S, \theta, Z_{t-1})$ *is uniformly continuous in* θ *on* Θ_S *(in the sense of (21)). Then*

$$\sup_{\theta \in \Theta_S} \left| V_N(S, \theta) - \overline{V}(S, \theta) \right| \to 0 \quad \textit{w.p. } 1 \quad \textit{as} \quad N \to \infty$$

□

4.2 Structure Identification

The model set is now hierarchically decomposed into parametric and structural levels. This allows us to study joint structural and parametric identification. In the sequel we assume that a data sequence Z_{2N} of length $2N$ is available. The first half is applied for parameter identification, while the second half is applied for structure identification. We define $\hat{M}_{S,N} = (S, \hat{\theta}_{S,N})$. This is the (not necessarily unique) model within a fixed structure S that on the basis of the N observations in the parameter identification data sequence is believed to be the best. Next, we need to choose among the models in the set $\{\hat{M}_{S,N} \mid S \in \mathcal{S}\}$. We define a new criterion on the basis of the second half of the data sequence Z_{2N}

$$V'_N(S,\theta) = \frac{1}{N} \sum_{t=N+1}^{2N} ||\varepsilon_t(S,\theta)||^2 \tag{22}$$

and assume that the minima exists:

Assumption A2. For every N and $S \in \mathcal{S}$, the global minima

$$\hat{\theta}_{S,N} = \arg\min_{\theta \in \Theta_S} V_N(S,\theta)$$
$$\hat{S}_{2N} = \arg\min_{S \in \mathcal{S}} V'_N(S, \hat{\theta}_{S,N})$$

exists with probability one.

□

These minima need not be unique, and we use the convention that in the case of non-uniqueness, $\hat{S}_{2N}$ and $\hat{\theta}_{S,N}$ are any global minima. In summary, the model estimate $\hat{M}_{2N}$ is defined by the following hierarchical optimization procedure

$$\text{Parameter Identification: } \hat{\theta}_{S,N} = \arg\min_{\theta \in \Theta_S} V_N(S,\theta) \quad \text{for all } S \in \mathcal{S} \tag{23}$$

$$\text{Structure Identification: } \hat{S}_{2N} = \arg\min_{S \in \mathcal{S}} V'_N(S, \hat{\theta}_{S,N}) \tag{24}$$

$$\text{Identified Model: } \hat{M}_{2N} = (\hat{S}_{2N}, \hat{\theta}_{\hat{S}_{2N},N}) \tag{25}$$

The following result relies mainly on the existence of a ULLN and assumption A2 (the proof and additional details can be found in (Johansen and Weyer 1995)):

Theorem 5 *Suppose Assumptions A2, S1, M1 and M2 hold. Then the hierarchical identification procedure (23)-(25) has the property*

$$\overline{V}(\hat{M}_{2N}) \to \inf_{M \in \mathcal{M}} \overline{V}(M) \quad w.p.\ 1 \ \ as \ \ N \to \infty \tag{26}$$

The theorem tells us that the expected squared prediction error for the model $\hat{M}_{2N}$ will tend to the smallest possible one w. p. 1 as $N \to \infty$. Notice that the same convergence result also holds when V'_N is replaced by V_N, i.e. the same data is used for both structure and parameter identification.

An interesting fact is that assumption M1 is a sufficient characterization of the model set also in this case. Also notice that even though parameterization has been introduced, the parameter-free characterization of the model set in terms of predictor properties (Theorem 3) are still applicable.

5 Conclusions

The adopted principle for convergence proofs in system identification is shown to extend some well known convergence results to more general model sets. Compared to (Ljung 1978), compactness of the parameter set and continuity of the parameterization is not required. Compared to (Ljung 1987), the present results can be seen as a generalization to non-linear predictors. Complexity constraints on the model set are related to the existence of a finite ϵ-net (Vapnik 1982) and the use of a result from (Pollard 1984). This allows the model set to be characterized in a parameter-free context, and explicit conditions on the predictor rather than the parameterization and parameter set, are introduced.

However, learning theory contains more powerful complexity measures than ϵ-nets, such as Vapnik-Chervonenkis dimensions and covering numbers (Vapnik and Chervonenkis 1971, Vapnik 1982, Pollard 1984, Haussler 1992). These concepts are not easily extended to the system identification setup with sequentially correlated data, although there are some promising results in that direction (Campi and Kumar 1996, Weyer *et al.* 1996).

In conclusion, it appears that ideas from learning theory can indeed be applied to extend the system identification theory to be valid for more general model sets, but the full potential of this is not explored yet. Whether this would acutally lead to more efficient algorithms for system identication, as questioned in (Ljung 1996), is still an open issue.

References

Campi, M. C. and P. R. Kumar (1996). Learning dynamical systems in a stationary environment. In: *Proc. 35th IEEE Conf. Decision and Control, Kobe, Japan.* pp. 2308–2311.

Gerencsér, L. (1989). On a class of mixing processes. *Stochastics* **26**, 165–191.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computations* **100**, 78–150.

Hjalmarsson, H. (1993). Aspects on Incomplete Modeling in System Identification. PhD thesis. University of Linköping.

Johansen, T. A. and E. Weyer (1995). Model structure identification using separate validation data - asymptotic properties. In: *Proc. European Control Conference, Rome*. Vol. 1. pp. 833–838.

Ljung, L. (1978). Convergence analysis of parametric identification methods. *IEEE Trans. Automatic Control* **23**, 770–783.

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Ljung, L. (1996). PAC-learning and asymptotic system identification theory. In: *Proc. 35th IEEE Conference on Decision and Control, Kobe, Japan*. pp. 2303–2307.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

Söderström, T. and P. Stoica (1988). *System Identification*. Prentice Hall, Englewood Cliffs, NJ.

Valiant, L. G. (1984). A theory of the learnable. *Communication of the ACM* **27**, 1134–1142.

Vapnik, V. N. (1982). *Estimation of Dependences based on Empirical Data*. Springer-Verlag, New York.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Vapnik, V. N. and A. Ya. Chervonenkis (1971). On the uniform convergence of relative frequencies to their probabilities. *Theor. Probability Appl.* **56**, 269–280.

Weyer, E. (1992). System Identification in the Behavioural Framework. PhD thesis. Norwegian University of Science and Technology.

Weyer, E., R. C. Williamson and I. M. Y. Mareels (1996). Sample complexity of least squares identification of FIR models. In: *Proceedings of IFAC World Congress, San Fransisco*. Vol. J. pp. 239–244.