Finite Sample Properties of System Identification Methods

M. C. Campi and Erik Weyer

***Abstract*—In this note, we study the quality of system identification models obtained using the standard quadratic prediction error criterion for a general linear model class. The main feature of our results is that they hold true for a *finite data sample* and they are not asymptotic. The main theorems bound the difference between the expected value of the identification criterion evaluated at the estimated parameters and at the optimal parameters. The bound depends naturally on the model and system order, the pole locations, and the noise variance, and it shows that although these variables often do not enter in asymptotic convergence results, they do play an important role when the data sample is finite.**

***Index Terms*—Finite sample, nonasymptotic theory, prediction error methods, system identification.**

I. Introduction

In this note, we study the finite sample properties of system identification methods based on a quadratic criterion applied to a general linear model class. The asymptotic properties of these methods have been extensively studied over the last three decades (see, e.g., [8] or [14]) and are now well understood.

Frequently asked questions in system identification are: 1) what can we say about the model quality with these many data points?; 2) can we sensibly apply the asymptotic theory?; and 3) what does the applicability of the asymptotic theory depend on? All these questions involve mathematical issues of difficult treatment, and questions of this kind have remained unsolved so far.

The main result of this note (Theorem 4.2) quantitatively assesses the discrepancy between minimizing a theoretical identification cost and minimizing its empirical counterpart when only a finite number of data points is available. The estimate is given by

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

where $V_N(\theta) = (1/N)\sum_{t=1}^{N}\epsilon^2(t,\theta)$, and $\epsilon(t,\theta)$ is the prediction error at time t of the model parameterized by θ, and N is the number of data points. The best estimate is given by

$$\overline{\theta}_N = \arg\min_{\theta\in\Theta} \overline{V}_N(\theta)$$

where $\overline{V}_N(\theta) = (1/N)\sum_{t=1}^{N} E\epsilon^2(t,\theta)$ and E is the expectation operator. Theorem 4.2 gives a probabilistic bound on the difference $\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N)$ for a finite N. In this note, we focus on the identification criterion and not on the estimated parameter. In certain situations, such as when the model is going to be used for prediction, the main concern is in fact the difference $\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N)$ and not $\hat{\theta}_N - \overline{\theta}_N$. Of course, in other cases the estimated parameter can be important as well (see [19] for a discussion on this).

The difference $\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N)$ depends on a number of factors, including the variance of the noise, the model order, and the singularities of the model and data generation mechanism. All these dependencies have meaningful interpretations. A remarkable fact is that they often do not show up in asymptotic convergence results, and a finite sample theory is, therefore, needed in order to capture these dependencies. Although our results are valid for a finite number of data points, they are still mainly of conceptual interest since they are rather conservative.

The note is organized as follows. In Section I-A, we present an example which illustrate the difference between asymptotic and finite sample properties and shows that asymptotic theory can give misleading results even for an arbitrary large number of data points. In Section I-B, our results are put into perspective relative to previous results in the literature. The data generating mechanism and the model class under consideration are introduced in Section II. The identification criterion is discussed in Section III and the main results are presented in Section IV.

A. A Motivating Example

Consider the system

$$y(t) = ay(t-1) + bu(t-1) + e(t)$$

with $|a| < 1$, $b = 0$, $u(t) = 1$ for all t, and $e(t)$ a zero mean Gaussian white noise process with variance $1 - a^2$. It follows that the variance of $y(t)$ is 1. As a model, we use

$$\hat{y}(t,\theta) = \theta u(t-1).$$

The expected value of the identification criterion is

$$\begin{aligned}\overline{V}_N(\theta) &= E[(y(t) - \hat{y}(t,\theta))^2] \\ &= E[(ay(t-1) + e(t) - \theta u(t-1))^2] = 1 + \theta^2\end{aligned}$$

and, consequently, $\overline{\theta}_N = 0$. This is not a surprising result since in the true system we have $b = 0$, so there is no dependence of $y(t)$ on $u(t-1)$. The least–squares estimate is given by

$$\hat{\theta}_N = \left(\sum_{t=1}^{N} u^2(t-1)\right)^{-1}\left(\sum_{t=1}^{N} u(t-1)y(t)\right) = \frac{1}{N}\sum_{t=1}^{N} y(t)$$

and, hence, the difference between the empirical and theoretical identification cost is $\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N) = ((1/N)\sum_{t=1}^{N} y(t))^2$. After a bit of calculations, we find that

$$\begin{aligned}&E|\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N)| \\ &\qquad = \frac{1}{N^2(1-a)^2}\left((1-a^N)^2 + \sum_{t=1}^{N-1}(1-a^t)^2(1-a^2)\right).\end{aligned}$$

Now, if we let $N \to \infty$, $E|\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N)| \to 0$, regardless of the value of a, namely, $\sup_{|a|<1}\lim_{N\to\infty} E|\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N)| = 0$.

This result is consistent with the asymptotic theory according to which, given a system (i.e., a value of the parameter a), when $N \to \infty$, $\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N)$ tends to zero. On the other hand, it should be clear that the above result is totally different from saying that $\lim_{N\to\infty}\sup_{|a|<1} E|\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N)| = 0$ (i.e., $\lim$ and $\sup$ do not commute). In fact, by sending a to 1, we obtain

$$\frac{1}{N^2(1-a)^2}\left((1-a^N)^2 + \sum_{t=1}^{N-1}(1-a^t)^2(1-a^2)\right) \to 1$$

for all N

and, therefore, $\lim_{N\to\infty}\sup_{|a|<1} E|\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N)| = 1$.

This result has an important interpretation; no matter how large N is, there exists a strictly stable data generation mechanism such that the asymptotic theory fails to hold even approximately. □

Manuscript received September 7, 2000; revised June 13, 2001 and February 4, 2002. Recommended by Associate Editor L. Y. Wang.

M. C. Campi is with the Department of Electrical Engineering and Automation, University of Brescia, 25123 Brescia, Italy.

E. Weyer is with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville VIC 3010, Australia.

Publisher Item Identifier 10.1109/TAC.2002.800750.

The example illustrates that the quality of a model identified on the basis of a *finite* data set cannot be assessed by means of the asymptotic theory, unless further information is provided. The reason for this is that, for any large data sample, there are systems for which such a theory does not apply. Furthermore, the example shows that the model quality will depend on variables—such as the location of the true system pole (the a parameter in the example)—that do not show up in the asymptotic theory. The goal of this note is to shed light on these dependencies.

Admittedly, in the previous example the asymptotic variance of the estimate is alerting us to the problem. The asymptotic variance of $(\hat{\theta}_N - \overline{\theta}_N)$ is $(1/N)((1-a^2)/(1-a)^2)$ which for fixed N tends to ∞ as $a \to 1$. However, it is not difficult to find another example where also the asymptotic variance give misleading results for any finite number of data points. An example is the system

$$y(t) = b_1 u_1(t-1) + b_2 u_2(t-1) + e(t)$$

where $e(t)$ is i.i.d. and zero mean with variance 1. Let $u_1(t)$ and $u_2(t)$ be generated by $u_1(t) = a_1 u_1(t-1) + e_1(t)$ and $u_2(t) = a_2 u_2(t-1) + e_2(t)$ where $e_1(t)$ and $e_2(t)$ are independent of each other, i.i.d., and zero mean with variances $1 - a_1^2$ and $1 - a_2^2$, respectively. As a model we use $\hat{y}(t) = \theta_1 u_1(t-1) + \theta_2 u_2(t-1)$.

The asymptotic theory [8] tells us that

$$\sqrt{N}\left(\begin{bmatrix}\hat{\theta}_1\\ \hat{\theta}_2\end{bmatrix} - \begin{bmatrix}b_1\\ b_2\end{bmatrix}\right)$$

is asymptotically Gaussian with zero mean and covariance matrix $I_{2\times 2}$. On the other hand, for a finite N the variance is given by $NE[(\sum_{t=1}^{N}\phi(t)\phi(t)^T)^{-1}]$ where $\phi(t) = [u_1(t-1)\ \ u_2(t-1)]^T$. The norm of this matrix tends to infinity when both a_1 and a_2 tend to one. The reason being that u_1 and u_2 tend to constant values when a_1 and a_2 tend to one, so that the system looses excitation. This entails that the asymptotic result can be completely misleading for any fixed N.

B. Putting Our Results Into Perspective Within the Existing Literature

Finite sample properties of quadratic identification methods have been studied in [20] and [18]. Results similar to our Theorem 4.1 were obtained under much more restrictive conditions using the Vapnik–Chervonenkis dimension. It was essentially assumed that the observed data were m-dependent or β-mixing, and the model class was restricted to autoregression with exogeneous variables (ARX) and finite-impulse response (FIR) models. However, the assumptions of those papers do not fit the standard identification context, as signals generated by dynamical systems are not β-mixing in general. These efforts witness the difficulty of extending to a dynamical context the results developed by the statistical learning literature (see, e.g., [15], [16], or [4]).

Other related results can be found in [7] (see also [9] and [17]). There, the problem of optimizing the choice of the model order in the context of FIR system identification is studied as the problem of balancing the approximation and the estimation errors. Nonasymptotic bounds on the accuracy of the least–squares estimate are presented in the note and they are used in order to quantify the estimation error in the aforementioned balancing problem. However, only FIR models and m-dependent regressors are considered, and the input signal and the noise is assumed to be i.i.d. These stringent conditions are necessary in order to obtain the many interesting results in that paper, but are over restrictive if we specialize to the problem considered in this note.

The results of the present note are much more general. We consider a general linear model structure (6) for identification of a general linear data generation mechanism [see (1)]. The identification criterion is the standard quadratic cost and, therefore, our results address the model quality assessment problem in a standard identification setting.

The results of this note are heavily based on exponential inequalities for stochastic processes. Identification and estimation methods have been analyzed in [2], [12], and [13] using similar techniques. However, those papers focused on the asymptotic properties when identification and estimation were carried out over an increasing model or function class, and the finite sample properties were not considered.

It should also be mentioned that finite-sample properties have been studied in the set membership and worst case identification setting, see, e.g., [17], [5], [6], and [3]. As a matter of fact, in this setting, identification algorithms are conceived so as to return all models that comply with the collected data, so that finite-sample results are automatically included in the identification result. As these are deterministic frameworks, the results are deterministic and quite different from ours. However, the results do involve entities such as gain, stability margins, size of disturbances, and unmodeled dynamics, which are related to and play a similar role as the pole and zero locations and model and system orders used in this note. Milanese and Taragna [11] consider similar problems as the above listed references and derive finite sample result in a stochastic setting starting from frequency domain measurement corrupted by i.i.d. Gaussian noise.

II. The Identification Setting

A. The Data Generation Mechanism

We assume that the observed data are generated by a linear system

$$y(t) = G_0 u(t) + H_0 e(t) \tag{1}$$

where $e(t)$ is a sequence of independent Gaussian random variables with zero mean and variance σ^2. The system is assumed to operate in open-loop. Correspondingly, $u(t)$ is seen as a deterministic signal. We also assume that $u(t)$ is bounded according to $|u(t)| \le U$. G_0 and H_0 are transfer functions in the backward shift operator q^{-1}, i.e., $q^{-1}y(t) = y(t-1)$; however, for the sake of readability, we omit throughout to explicitly indicate the dependence on q^{-1}. G_0 and H_0 can be written as $G_0 = B_0/A_0$ and $H_0 = C_0/D_0$ where

$$A_0 = 1 + a_{01}q^{-1} + \cdots + a_{0n_0}q^{-n_0} \tag{2}$$

$$B_0 = b_{01}q^{-1} + \cdots + b_{0n_0}q^{-n_0} \tag{3}$$

$$C_0 = 1 + c_{01}q^{-1} + \cdots + c_{0n_0}q^{-n_0} \tag{4}$$

$$D_0 = 1 + d_{01}q^{-1} + \cdots + d_{0n_0}q^{-n_0} \tag{5}$$

and n_0 is an upper bound on the degrees. Moreover, we assume that the zeros of A_0, C_0, and D_0 are inside a circle of radius $\eta_0 < 1$, i.e., we assume stability of the system and also that H_0 has a stable inverse.

B. Model Class

The model class considered is

$$y(t) = G(\theta)u(t) + H(\theta)w(t) \tag{6}$$

where $w(t)$ is a sequence of independent Gaussian random variables with zero mean, $G(\theta) = B(\theta)/A(\theta)$ and $H(\theta) = C(\theta)/D(\theta)$. The order of $A(\theta)$, $B(\theta)$, $C(\theta)$, and $D(\theta)$ are upper bounded by n_1, but otherwise similar to (2)–(5). It should be noted that we do not assume that the model class contains the true system. θ contains the unknown coefficients of these polynomials, and we assume that θ belongs to a set Θ, such that $A(\theta)$, $C(\theta)$, and $D(\theta)$ have all their zeros inside a circle of radius $\eta_1 < 1$, and $B(\theta)$ has all zeros inside acircle of radius μ_1 and the leading coefficient b_1 is bounded according to $|b_1| \le B_1$.

For future use, we define

$$\begin{aligned}\eta &:= \max\{\eta_0,\, \eta_1\}\\ \mu &:= \max\{\text{magnitude of the zeroes of } B_0,\, \mu_1\}\\ B &:= \max\{b_{01},\, B_1\}\\ \tilde{n} &:= \max\{n_0,\, n_1\}.\end{aligned}$$

The model quality depends on μ, η, $\tilde{n}$, and B, and in this note, we make these dependencies explicit.

III. The Identification Criterion

From a system identification perspective, the most important feature of the aforementioned model is its associated predictor, which is given by

$$\hat{y}(t,\,\theta) = (1 - H^{-1}(\theta))y(t) + H^{-1}(\theta)G(\theta)u(t). \tag{7}$$

As we only have data available for $t \geq 1$, it is common to use the computational scheme

$$A(\theta)C(\theta)\hat{y}(t,\,\theta) = (A(\theta)C(\theta) - A(\theta)D(\theta))y(t) + B(\theta)D(\theta)u(t) \tag{8}$$

with $\hat{y}(t,\,\theta) = 0$, $y(t) = 0$, and $u(t) = 0$ for $t \leq 0$. On the other hand, if we assume that the system is initially at rest (as we certainly do in order to avoid cumbersome computations due to initial condition issues), there is no difference between $\hat{y}(t,\,\theta)$ in (7) and (8).

The corresponding prediction error is given by

$$\epsilon(t,\,\theta) = y(t) - \hat{y}(t,\,\theta) = H^{-1}(\theta)y(t) - H^{-1}(\theta)G(\theta)u(t). \tag{9}$$

Ideally, one would like to choose θ such that

$$\overline{V}_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} E\epsilon^2(t,\,\theta) \tag{10}$$

is minimized, where N is the number of data points. Since the data generation mechanism is unknown, one cannot compute the expected value, and, in place of (10), its sample version

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} \epsilon^2(t,\,\theta) \tag{11}$$

is used. Clearly, the minimization of $V_N(\theta)$ can only be expected to be equivalent to minimization of $\overline{V}_N(\theta)$ when $N \to \infty$ (and, this is indeed the case under mild assumptions, see, e.g., [8]). One question that arises naturally is to quantify the deterioration in the model quality due to the finiteness of the data sample. In a formal way, answering this question requires quantitative bounds for

$$\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N) \tag{12}$$

where $\hat{\theta}_N = \arg\min_{\theta\in\Theta} V_N(\theta)$ and $\overline{\theta}_N = \arg\min_{\theta\in\Theta} \overline{V}_N(\theta)$. Equation (12) quantifies the discrepancy between the ideal identification result [measured by $\overline{V}_N(\overline{\theta}_N)$] and the achieved result $\overline{V}_N(\hat{\theta}_N)$. The expected value $\overline{V}_N(\theta)$ depends on the input signal, as does the optimal value $\overline{\theta}_N$ and the estimate $\hat{\theta}_N$. As we are considering the difference between two expected values of the squared prediction errors (12), issues such as identifiability and persistence of excitation do not enter the picture, and we do not have to impose any conditions in this regard. Needless to say, such issues are of major importance when the focus shifts to properties of the estimated parameters.

IV. A Bound on the Model Quality

Deriving exact expressions for the probability distribution of (12) is truly overwhelming. Instead, bounds for (12) are derived. The final expressions are very interesting as they reveal the dependence of the identification result on a number of variables which often disappear in the asymptotic analysis. The bound for (12) is given in Theorem 4.2 below. The proof of the bound is immediate from the more general result presented first as Theorem 4.1.

Theorem 4.1: Assume the data has been generated by the process (1) as described in Section II-A, and let the model class be given by (6) as described in Section II-B. Let the prediction error be computed according to (9). Given any two real numbers $\epsilon_0 > 0$ and $\nu_0 > 0$, let $\epsilon = \epsilon_1 + \epsilon_2$ and $\delta = \delta_1 + \delta_2$ with

$$\epsilon_1 = 2^{4\tilde{n}}\epsilon_0\,\frac{4\tilde{n}\eta + (1-\eta)}{(1-\eta)^{4\tilde{n}+1}} \tag{13}$$

$$\epsilon_2 = 2^{3\tilde{n}+2}\nu_0 B\left(1 + \frac{\mu}{\eta}\right)^{\tilde{n}-1}\frac{(4\tilde{n}-1)\eta + 2(1-\eta)}{(1-\eta)^{4\tilde{n}+1}} \tag{14}$$

$$\delta_1 = 4\,\frac{e^{-(N\epsilon_0^2/4\sigma^2(4\sigma^2+\epsilon_0^2))}}{\left(1 - e^{-(N\epsilon_0/4\sigma^2(4\sigma^2+\epsilon_0))}\right)^2} \tag{15}$$

$$\delta_2 = 2\,\frac{e^{-(N\nu_0^2/2U\sigma(2U\sigma+\nu_0))}}{\left(1 - e^{-(N\nu_0^2/2U\sigma(2U\sigma+\nu_0))}\right)^2}. \tag{16}$$

Then, the following bound holds:

$$\Pr\left(\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{t=1}^{N}\epsilon^2(t,\,\theta) - \frac{1}{N}\sum_{t=1}^{N}E\epsilon^2(t,\,\theta)\right| \geq \epsilon\right) \leq \delta.$$

Proof: See Appendix A. □

Theorem 4.1 provides a *uniform* bound for the difference $V_N(\theta) - \overline{V}_N(\theta)$ between the empirical and the theoretical identification cost. This bound is the key to establish the following main result

Theorem 4.2: With ϵ and δ as in Theorem 4.1, we have $\overline{V}_N(\hat{\theta}_N) - \overline{V}_N(\overline{\theta}_N) \leq 2\epsilon$ with probability at least $1 - \delta$.

Proof: By applying Theorem 4.1 twice, it follows with probability at least $1 - \delta$ that $\overline{V}_N(\hat{\theta}_N) \leq V_N(\hat{\theta}_N) + \epsilon \leq V_N(\overline{\theta}_N) + \epsilon \leq \overline{V}_N(\overline{\theta}_N) + 2\epsilon$. □

Theorem 4.2 is believed to be the first result addressing the problem of quantifying the identification performance obtained by minimization of the criterion (11) for a general linear model class (6). It is important to stress that the result holds true for any *finite* data sample of size N, that is, it is not asymptotic in N and, hence, as can be seen from (15) and (16), ϵ and δ depend on N. Also notice that the assumption that $e(t)$ is Gaussian is not crucial. It is only used to establish (24) in the Appendix and other cases can be considered as well.

For a fixed ϵ, the number of data points required in order to guarantee the bound in Theorem 4.1 increases only as $\log(1/\delta)$ as $\delta \to 0$. On the other hand for a fixed δ, the number of data points required increases roughly as $1/\epsilon^2$ as $\epsilon \to 0$. This is often popularly phrased as "confidence is cheap, but accuracy is expensive."

We see that even though the dependence on σ^2, η, μ, $\tilde{n}$, and B often disappears in asymptotic results, they play a fundamental role in the finite sample properties.

The bound tells us that, with a certain confidence $1 - \delta$, minimizing the empirical cost (11) corresponds to minimizing the theoretical cost (10) to within an accuracy 2ϵ. The presence of a confidence $1 - \delta$ is a natural ingredient stemming from the stochastic nature of the problem. When the data sample is finite, there is always a nonzero (even though possibly small) probability that the noise plays against the identification objective, resulting in a deterioration of the accuracy. Suppose we want to decrease δ. This corresponds to increase ϵ_0 and ν_0. In turn, this entails that ϵ increases, and not surprisingly, ϵ tends to infinity in the limit as $\delta \to 0$. This is in agreement with the fact that no level of accuracy can be guaranteed with probability 1 for a finite data sample.

This is in contrast with the asymptotic theory, where the assumption $N \to \infty$ leads to a result valid with probability 1.

When $\sigma^2 \to 0$, the stochastic nature of the problem disappears result can be guaranteed with probability 1 for any ϵ: for arbitrary small ϵ_0 and ν_0 in (13) and (14), δ_1 and δ_2 given by (15) and (16) tend to zero as $\sigma^2 \to 0$.

Suppose now we fix a certain confidence level δ. Then, we can find $\epsilon_0 = \epsilon_0(N, \sigma^2, \delta)$ and $\nu_0 = \nu_0(N, \sigma^2, \delta)$ such that the desired confidence level is achieved. Substituting such ϵ_0 and ν_0 in (13) and (14), we see that the accuracy depends on $\tilde{n}$, η, μ, B, and N, besides δ and σ^2. These dependencies are now analyzed.

$\epsilon \to \infty$, both when the system and/or the model complexity (as measured by $\tilde{n}$) tend to infinity and when $\eta \to 1$. This behavior can easily be understood. Increasing $\tilde{n}$ or sending η to 1 leads to a prediction error process (8) with a long correlation tail (the prediction error transfer functions increase in size and their poles get close to the unit circle). When this happens, the averaging effect on the noise is decreased and a large number of data points is necessary to guarantee a certain accuracy.

When μ or $B \to \infty$, ϵ_2 in (14) tends to infinity too, so that $\epsilon \to \infty$. This is also expected since a large value of μ and/or B inflates the contribution to the error due to the input signal u. Finally, ϵ_0 and ν_0 tend to zero as $N \to \infty$, and hence $\epsilon \to 0$ when $N \to \infty$, as it is expected because of averaging effects.

V. Conclusion and Discussion

In this note, we have studied the issue of model quality for system identification methods based on the standard quadratic prediction error criterion for a general linear model class. The main feature of our results is that they hold true for a finite data sample and they are not asymptotic in nature. Theorem 4.2 bounds the difference between the expected value of the identification criterion evaluated at the estimated parameter value and at the optimal value. The bound depends naturally on the model and system order, the pole locations and the noise variance, and it shows that although these variables often do not enter in asymptotic convergence results, they do play an important role when the data sample is finite which of course is always the situation in practice.

The results in this note can certainly not be considered as final. For example, we have made no attempts of optimizing the bounds. The main goal is that of stimulating research activity in the field of model quality assessment. It is our belief that certain recent advances in the statistical literature (some of which have been used in the appendices) give us today the opportunity to attack this important problem, and our results should be considered as a first step in this direction.

While we credit our results for being a significant step in the direction of clarifying the model quality dependencies, we also recognize that they are not tight enough to compute the needed amount of data points in real situations. More effort has to be devoted in the direction of working out applicable bounds.

Appendix A
Proof of the Main Theorem

Theorem 4.1 follows by combining Lemmas A.1 and A.2. In Lemma A.1, the difference between the expected and empirical values of the identification criterion is expressed in terms of the noise sequence $e(t)$ and the input sequence $u(t)$. The probability that these expressions exceed a certain value is then bounded in Lemma A.2 using exponential inequalities.

Lemma A.1:

$$\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{t=1}^{N}\epsilon^2(t,\theta)-\frac{1}{N}\sum_{t=1}^{N}E\epsilon^2(t,\theta)\right| \le \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}s_k s_l\left|\frac{1}{N}\sum_{t=1}^{N}(e(t-k)e(t-l) - Ee(t-k)e(t-l))\right| + \sum_{k=0}^{\infty}\sum_{l=1}^{\infty}s_k r_l\left|\frac{2}{N}\sum_{t=1}^{N}e(t-k)u(t-l)\right|$$

with

$$r_k = 2^{n_1+1}B\left(1+\frac{\mu}{\eta}\right)^{\tilde{n}-1}\frac{k\cdots(k+n_1+\tilde{n}-2)}{(n_1+\tilde{n}-1)!}\,\eta^{k-1} \tag{17}$$

$$s_k = 2^{n_1+n_0}\,\frac{(k+1)\cdots(k+n_1+n_0-1)}{(n_1+n_0-1)!}\,\eta^k. \tag{18}$$

Proof: By introducing

$$R(\theta) := H^{-1}(\theta)(G_0 - G(\theta)) = r_{1,\theta}q^{-1} + r_{2,\theta}q^{-2} + \cdots \tag{19}$$

$$S(\theta) := H^{-1}(\theta)H_0 = 1 + s_{1,\theta}q^{-1} + s_{2,\theta}q^{-2} + \cdots \tag{20}$$

the prediction error can be written as $\epsilon(t,\theta) = R(\theta)u(t) + S(\theta)e(t)$. The difference between the empirical and theoretical value of the identification criterion is given by

$$\frac{1}{N}\sum_{t=1}^{N}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}s_{k,\theta}s_{l,\theta}(e(t-k)e(t-l) - Ee(t-k)e(t-l)) + \frac{2}{N}\sum_{t=1}^{N}\sum_{k=0}^{\infty}\sum_{l=1}^{\infty}s_{k,\theta}r_{l,\theta}e(t-k)u(t-l).$$

Let $s_k := \sup_{\theta\in\Theta}|s_{k,\theta}|$ and $r_k := \sup_{\theta\in\Theta}|r_{k,\theta}|$. The Lemma follows by interchanging the summations. The bounds on r_k and s_k are given by Lemma C.2. ∎

Lemma A.2:

$$\Pr\left\{\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}s_k s_l\left|\frac{1}{N}\sum_{t=1}^{N}(e(t-k)e(t-l) - Ee(t-k)e(t-l))\right| > \epsilon_1\right\} \le \delta_1 \tag{21}$$

$$\Pr\left\{\sum_{k=0}^{\infty}\sum_{l=1}^{\infty}s_k r_l\left|\frac{2}{N}\sum_{t=1}^{N}e(t-k)u(t-l)\right| > \epsilon_2\right\} \le \delta_2 \tag{22}$$

where ϵ_1, ϵ_2, δ_1 and δ_2 are given by (13)–(16).

Proof: First, we prove (21). Suppose that

$$\left|\frac{1}{N}\sum_{t=1}^{N}(e(t-k)e(t-l) - Ee(t-k)e(t-l))\right| \le \epsilon(k,l) \tag{23}$$

where $\epsilon(k,l) = (k+l+1)\epsilon_0$. Then (we use the bounds of r_k and s_k in Lemma A.1)

$$\begin{aligned}&\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}s_k s_l\left|\frac{1}{N}\sum_{t=1}^{N}(e(t-k)e(t-l) - Ee(t-k)e(t-l))\right| \\ &\le \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}2^{2n_1+2n_0}\frac{(k+1)\cdots(k+n_1+n_0-1)}{(n_1+n_0-1)!} \\ &\quad\cdot\eta^k\,\frac{(l+1)\cdots(l+n_1+n_0-1)}{(n_1+n_0-1)!}\,\eta^l(k+l+1)\epsilon_0 \\ &\le 2^{4\tilde{n}}\epsilon_0\,\frac{4\tilde{n}\eta+(1-\eta)}{(1-\eta)^{4\tilde{n}+1}} = \epsilon_1\end{aligned}$$

where we have used $\sum_{k=0}^{\infty}((k+1)\cdots(k+n-1)/(n-1)!)\eta^k = 1/(1-\eta)^n$ and $\sum_{k=0}^{\infty}(k(k+1)\cdots(k+n-1)/(n-1)!)\eta^k = n\eta/(1-\eta)^{n+1}$.

Next, we compute the probability that (23) holds true simultaneously for all $k,\ l \geq 0$. For a fixed pair of k and l this probability is bounded by Corollary B.2. Hence, using (25) and (26) we find that this probability is at least $1-\delta'$ where

$$\begin{aligned}\delta' &= \sum_{k=0}^{\infty} 2e^{-(N(2k+1)^2\epsilon_0^2/4\sigma^2(2\sigma^2+(2k+1)\epsilon_0))} \\ &\quad + \sum_{k=0}^{\infty}\sum_{l=0,\,l\neq k}^{\infty} 4e^{-(N(k+l+1)^2\epsilon_0^2/4\sigma^2(4\sigma^2+(k+l+1)\epsilon_0))} \\ &\leq \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} 4e^{-(N(k+l+1)^2\epsilon_0^2/4\sigma^2(4\sigma^2+(k+l+1)\epsilon_0))} \\ &= \sum_{m=0}^{\infty}(m+1)4e^{-(N(m+1)^2\epsilon_0^2/4\sigma^2(4\sigma^2+(m+1)\epsilon_0))} \\ &\leq \sum_{m=0}^{\infty}(m+1)4e^{-(N(m+1)\epsilon_0^2/4\sigma^2(4\sigma^2+\epsilon_0))} = \delta_1\end{aligned}$$

where we have used $\sum_{m=0}^{\infty}(m+c_1)a^{m+c_2} = (a^{c_2}/(1-a)^2)(1+(c_1-1)(1-a))$. Equation (22) follows along similar lines. ∎

Appendix B
Exponential Inequalities

Theorem B.1 (Bernstein's Inequality, [1, Th. 1.2]): Let $X_1, \ldots, X_N$ be independent zero-mean random variables and let $S_N = \sum_{t=1}^{N} X_t$. Assume there exists a $c > 0$ such that

$$E|X_t|^p \leq c^{p-2}p!EX_t^2 < \infty, \qquad t = 1, \ldots, N;\ p = 3, 4, \ldots. \tag{24}$$

Then $\Pr\{|S_N| \geq \epsilon\} \leq 2\exp(-(\epsilon^2/4\sum_{t=1}^{N} EX_t^2 + 2c\epsilon))$.

Corollary B.2: Let $e(t), t \in \mathbb{Z}$ be zero mean i.i.d. Gaussian variables with variance σ^2, and let $u(t), t \in \mathbb{Z}$ be deterministic variables satisfying $|u(t)| \leq U$. Then

$$\Pr\left\{\left|\frac{1}{N}\sum_{t=1}^{N} e^2(t-k) - \sigma^2\right| \geq \epsilon(k,\,k)\right\} \leq 2e^{-(N\epsilon^2(k,k)/4\sigma^2(2\sigma^2+\epsilon(k,k)))} \tag{25}$$

$$\Pr\left\{\frac{1}{N}\left|\sum_{t=1}^{N} e(t-k)e(t-l)\right| \geq \epsilon(k,\,l)\right\} \leq 4e^{-(N\epsilon^2(k,l)/4\sigma^2(4\sigma^2+\epsilon(k,l)))} \qquad k \neq l \tag{26}$$

$$\Pr\left\{\frac{1}{N}\left|\sum_{t=1}^{N} e(t-k)u(t-l)\right| \geq \nu(k,\,l)\right\} \leq 2e^{-(N\nu^2(k,l)/2U\sigma(2U\sigma+\nu(k,l)))}. \tag{27}$$

Proof: Equations (25) and (27) follow directly from Theorem B.1 with $X_t = e^2(t-k)-\sigma^2, c = 2\sigma^2$ and $X_t = e(t-k)u(t-l), c = U\sigma$, respectively. For (26), we can group the time indexes $\{1, 2, \ldots, N\}$ into two sets A_1 and A_2 such that $e(t-k)e(t-l)$ and $t \in A_1$ are i.i.d. random variables and $e(t-k)e(t-l)$ and $t \in A_2$ are i.i.d. random variables. Equation (26) then follows with $X_t = e(t-k)e(t-l)$ and $c = \sigma^2$. ∎

Appendix C
Bounds on Coefficients

We first present a general result bounding the magnitude of the coefficients of certain polynomials. Then, we use this result to bound the magnitude of the coefficients of $R(\theta)$ and $S(\theta)$.

Lemma C.1: Let $M(q^{-1}) = 1 + m_1q^{-1} + \cdots m_{n_m}q^{-n_m}$ and $P(q^{-1}) = 1 + p_1q^{-1} + \cdots p_{n_p}q^{-n_p}$ be polynomials with all zeros inside a circle of radius $\eta < 1$. Furthermore, let $W(q^{-1}) = w_1q^{-1} + \cdots w_{n_w}q^{-n_w}$ be a polynomial with all zeros inside a circle of radius μ and leading coefficient bounded by $|w_1| < B$. Then, the coefficients of the polynomials

$$M^{-1}(q^{-1}) = 1 + \overline{m}_1q^{-1} + \overline{m}_2q^{-2} + \cdots \tag{28}$$

$$M^{-1}(q^{-1})P(q^{-1}) = 1 + \overline{p}_1q^{-1} + \overline{p}_2q^{-2} + \cdots \tag{29}$$

$$M^{-1}(q^{-1})P(q^{-1})W(q^{-1}) = \overline{w}_1q^{-1} + \overline{w}_2q^{-2} + \cdots \tag{30}$$

are bounded by

$$|\overline{m}_k| \leq \frac{(k+1)\cdots(k+n_m-1)}{(n_m-1)!}\,\eta^k \tag{31}$$

$$|\overline{p}_k| \leq 2^{n_p}\,\frac{(k+1)\cdots(k+n_m-1)}{(n_m-1)!}\,\eta^k \tag{32}$$

$$|\overline{w}_k| \leq 2^{n_p}B\left(1+\frac{\mu}{\eta}\right)^{n_w-1}\frac{k\cdots(k+n_m-2)}{(n_m-1)!}\,\eta^{k-1}. \tag{33}$$

Proof: First, we bound the coefficients of $M^{-1}(q^{-1})$. $1/M(q^{-1}) = \prod_{j=1}^{n_m}(1/(1-z_jq^{-1}))$. $1/(1-z_jq^{-1}) = 1+z_jq^{-1}+z_j^2q^{-2}+z_j^3q^{-3}+\cdots$ with $|z_j| < \eta$ and, hence, the coefficient in front of q^{-k} is less than η^k in absolute value. With two roots z_1 and z_2, we have $1/(1-z_1q^{-1})(1-z_2q^{-1}) = 1+\overline{m}_1q^{-1}+\overline{m}_2q^{-2}\cdots$ where $|\overline{m}_k| \leq (k+1)\eta^k$. Continuing recursively in the number of roots, we end up with (31). The other bounds follow similarly. ∎

Lemma C.2: Let $R(\theta)$ and $S(\theta)$ be the polynomials given by (19) and (20), respectively. Then, for all $\theta \in \Theta$ the coefficients of the polynomials are bounded by

$$|r_k| \leq 2^{n_1+1}B\left(1+\frac{\mu}{\eta}\right)^{\tilde{n}-1}\frac{k\cdots(k+n_1+\tilde{n}-2)}{(n_1+\tilde{n}-1)!}\,\eta^{k-1} \tag{34}$$

$$|s_k| \leq 2^{n_1+n_0}\,\frac{(k+1)\cdots(k+n_1+n_0-1)}{(n_1+n_0-1)!}\,\eta^k. \tag{35}$$

References

[1] D. Bosq, *Nonparametric Statistics for Stochastic Processes—Estimation and Prediction*. New York: Springer-Verlag, 1998, Lecture Notes in Statistics 110.

[2] M. C. Campi and P. R. Kumar, "Learning dynamical systems in a stationary environment," *Syst. Control Lett.*, vol. 34, pp. 125–132, 1998.

[3] J. Chen and G. Gu, *Control Oriented System Identification. An H_∞ Approach*. New York: Wiley, 2000.

[4] V. Cherkassky and F. Mulier, *Learning From Data*. New York: Wiley, 1998.

[5] L. Giarre', B. Kacewicz, and M. Milanese, "Model quality evaluation in H_2 identification," *Automatica*, vol. 33, pp. 1133–1139, 1997.

[6] L. Giarre' and M. Milanese, "Model quality evaluation in set membership identification," *IEEE Trans. Automat. Contr.*, vol. 42, pp. 691–698, May 1997.

[7] A. Goldenshluger, "Nonparametric estimation of transfer functions: Rates of convergence and adaptation," *IEEE Trans. Inform. Theory*, vol. 44, pp. 644–658, Mar. 1998.

[8] L. Ljung, *System Identification—Theory for the User*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[9] L. Ljung and Z. D. Yuan, "Asymptotic properties of the least squares method for estimating transfer functions," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 514–530, June 1985.
[10] L. Ljung and B. Wahlberg, "Asymptotic properties of the least squares method for estimating transfer functions and disturbance spectra," *Ad. Appl. Probab.*, vol. 24, pp. 412–440, 1978.
[11] M. Milanese and M. Taragna, "H_∞ identification of 'soft' uncertainty models," *Syst. Control Lett.*, vol. 37, pp. 217–228, 1999.
[12] D. S. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2133–2145, Nov. 1996.
[13] ——, "Memory-universal prediction of stationary random processes," *IEEE Trans. Inform. Theory*, vol. 44, pp. 117–133, Jan 1998.
[14] T. Söderström and P. Stoica, *System Identification*. Upper Saddle River, NJ: Prentice-Hall, 1988.
[15] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[16] M. Vidyasagar, *A Theory of Learning and Generalization*. New York: Springer-Verlag, 1997.
[17] B. Wahlberg and L. Ljung, "Hard frequency-domain model error bounds from least-squares like identification techniques," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 900–912, July 1992.
[18] E. Weyer, "Finite sample properties of system identification of ARX models under mixing conditions," *Automatica*, vol. 36, no. 9, pp. 1291–1299, 2000.
[19] E. Weyer and M. C. Campi, "Non-asymptotic confidence ellipsoids for the least squares estimate," in *Proc. 39th IEEE Conf. Decision Control*, Sydney, Australia, Dec. 2000, pp. 2688–2693.
[20] E. Weyer, R. C. Williamson, and I. M. Y. Mareels, "Finite sample properties of linear model identification," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 1370–1383, July 1999.

A Globally Stabilizing Hybrid Variable Structure Control Strategy

A. Ferrara, L. Magnani, and R. Scattolini

***Abstract*—A hybrid variable structure control strategy for a class of second order systems is presented in this note. It relies on a system state decomposition into regions, and on a suitable event-driven switching among the corresponding control laws. By enforcing conventional and unconventional sliding-mode behaviors, as well as avoiding the generation of limit cycles, the proposed strategy proves to globally asymptotically stabilize the origin of the system state space.**

***Index Terms*—Continuous-time systems, hybrid systems, variable structure systems.**

I. Introduction

During recent years, an extensive literature has been devoted to the subject of hybrid systems [1]–[6]. Even if the term "hybrid" allows various interpretations, the definition which is becoming the conventional one is that a hybrid system is a system the evolution of which is characterized by the interlacing of continuous-time and discrete-valued signals. Since plenty of computer supervised continuous-time controlled systems satisfy this definition, a great number of practical applications of hybrid systems can be envisaged, such as traffic or intelligent vehicle control systems, communication networks, chemical process control, robotic systems [5], [6].

As far as hybrid system modeling is concerned, various approaches have been pursued [7]. One of the most commonly used relies on a state space decomposition into regions delimited by borders which can be interpreted as the switching boundaries of the switched controlled systems. These systems can be modeled by equations with event or transition-dependent parts. Alternatively or contemporarily, it could be the control law to be switched on the crossing of the switching boundaries in a sort of gain-scheduling fashion. The crucial problem of the existence of solutions of the event-driven equation modeling the hybrid system requires to be faced by making reference to the theory of differential inclusions [8]. On the other hand, stability issues are not trivial at all, since it is sensible to foresee that letting the controller switch among different control laws one may obtain an unstable closed-loop system [9], [10].

Also variable structure control (VSC) systems, to which a large numbers of works have been devoted during the past two decades [11]–[14], comply with the aforementioned definition of hybrid systems. They are "hybrid" in the sense that the control design is still based on a state space decomposition through a border which is a linear or nonlinear function of the full system state, so that the control law is switched on crossing it. Yet, they do not fit the intuitive idea one has of hybrid systems, since the key point in the theory of VSC systems is to force the state trajectory not to instantaneously cross the commutation manifold as expected in classical hybrid systems, but to slide on it. Indeed, in this way, the desired dynamical features turn out to be assigned to the controlled system (accordingly, the switching boundary is called sliding manifold). Moreover, since the control is designed so that, once the sliding manifold is reached, the state trajectory is maintained on it featuring a regime called sliding mode, the origin of the state space, regarded as the equilibrium point to be asymptotically stabilized, must belong to the sliding manifold.

The aim of the present note is to design and analyze a truly hybrid VSC strategy for a class of second-order systems which relies on a peculiar system state decomposition into countable regions by means of a grid of conventional sliding manifolds, and commutation manifolds not including the origin. Each region is a "block" in the sense used in [16], and a block invariant control law is associated with it. On the whole, the control laws corresponding to the blocks included between two commutation manifolds (note that also infinity and the origin of the state space can be interpreted in this way) concur to the attainment of either the sliding mode objective to reach a particular sliding manifold, or to the aim of crossing the commutation manifold closer to the origin. The overall hybrid VSC strategy proves to globally asymptotically stabilize the origin of the system state space, even if unconventional sliding modes can be generated on the commutation manifolds. Indeed, these behaviors, in contrast to what happens in classical hybrid systems exhibiting chattering solutions, the so-called "Zeno systems" [17], turns out to be of finite time duration.

The motivation for using VSC to design a hybrid strategy mainly relies on the appreciable features of the VSC methodology, such as simplicity and robustness versus matched uncertainties and disturbancies, which are naturally inherited by the proposed control approach. Note that, the combination of VSC with hybrid control has already been investigated in [18] and [19]. Yet, the control strategies proposed in such papers are characterized by a continuous adaptation of the control gain, and switching is driven by a logic relying on the decomposition of the

Manuscript received March 10, 2000; revised December 11, 2000 and October 17, 2001. Recommended by Associate Editor P. Voulgaris.

The authors are with the Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, 27100 Pavia, Italy (e-mail: ferrara@conpro.unipv.it; magnani@conpro.unipv.it; scatto@conpro.unipv.it).

Publisher Item Identifier 10.1109/TAC.2002.800749.