Your dedication goes here

Preface

Antonio (Toni) Lepschy passed away suddendly on June 30, 2005, leaving family, friends and colleagues truly shocked. He was a man of deep intelligence and a genuine gentleman with many qualities and with diversified interests. It is not the scope of this little book to discuss and celebrate the diverse aspects of his wide cultural interests (ranging from linguistics to history, mathematical economics etc.) since this would call for contributions from many parts and would involve a colossal effort, well beyond our capabilities and expertise. May we just refer the reader to the recent commemoration of Giovanni Marchesini at the Istituto Veneto di Scienze Lettere ed Arti [1].

This book wants to be a small token of remembrance and gratitude to Toni from colleagues belonging to the control systems community. It includes some papers contributed by foreign colleagues who experienced scientific collaboration with him. The contributions are mostly technical and hence written in the language of this discipline.

Toni, in the early days of the career of many of us, was, and in many ways, has continued to be, a *master* (this is indeed the joking appellative we used for him in Padova). This, in spite of his absolutely unpretentious and unassuming style. He introduced the oldest of us to the basic ideas and taught us the basic language of Automatic Control. This was happening in the early days of Italian engineering schools when Laplace transform and the mathematics behind the Nyquist criterion were looked upon with suspicion by many colleagues. His lectures were often stuffed with (sometimes quite pedantic) historic citations. Famous, among many, has remained the disquisition about the origins of the word *Control* which, although originating from the latin expression *contra-rotulum*, the act of checking against a list on a roll, had its original meaning completely subverted, causing perennial misunderstanding in our country whenever attempting to explain what our field is about to a layman. His joint book with late Antonio Ruberti has been for several decades

[1] G. Marchesini, *Ricordo di Antonio Lepschy*, Istituto Veneto di Scienze, Lettere ed Arti, Adunanza Accademica del 26 Novembre 2005, Venezia, 2006.

a classical textbook in many Italian universities. Pushed by the great technological advances of the recent years, the field has evolved greatly beyond the classical ideas exposed in the textbook. We'd nevertheless like to imagine that the younger contributors to this book perceived the beauty and the enthusiasm for this field through his teaching, even if they have never attended a class from him.

Padova,
August 2007

Giorgio Picci
Maria Elena Valcher

Contents

List of Contributors

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
`name@e-mail.*`

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Author Name
University/Institute Name
Street No.
X - Place, Postal Code
name@e-mail.*

Eigenstructure-based model reduction for generalized RL networks

Alessandro Beghi

Dipartimento di Ingegneria dell'Informazione, Università di Padova, via Gradenigo 6/B, 35131 Padova, Italy
beghi@dei.unipd.it

To the dear memory of Toni Lepschy, maestro *and friend.*

Summary. The use of model reduction schemes based on eigenstructure analysis is shown to be useful in the study of control problems for a class of electromechanical systems, namely generalized RL networks. In particular, techniques are considered that allow to produce reduced-order models where the physical meaning of the state variables is preserved. This turns out to be useful to obtain interesting interpretations in terms of physical quantities of the simplified model, its parameters, and the approximations involved.

Keywords. Linear Systems, Electromechanical Systems, Model Reduction.

1 Introduction

When deriving model of dynamical systems starting from physical laws, one has often to face the problem of dealing with hundreds or thousands of state variables. The need of describing the system behavior with sufficient detail contrasts with the so called "curse of dimensionality," that is, the computational difficulties associated with systems of very high dimension. As is known, high dimensionality is also a drawback for most of standard linear control schemes, such as LQG, LQG/LTR, H_∞. To cope with this problem, models of reduced order are derived, which allow to simplify the design of the controller while giving a sufficiently accurate description of the system. Several model reduction techniques are available from linear state-space control theory (see e.g. [1]). A drawback of many of such techniques is that the physical meaning of the state variables is lost in the reduced model. This fact can be particularly relevant when dealing with models of electromechanical systems, where both the involved variables (currents, voltages, displacements, velocities, etc.) and the model parameters (resistances, inductances, masses, forces, etc.) do have a specific meaning that can be of paramount importance in the interpretation of the results.

The model reduction techniques that we consider in this note are based on eigenstructure analysis of the state matrix A in the standard linear state space description

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{1}$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$, and allow to provide an interpretation of the variables and parameters of the simplified models in terms of physical quantities. As a consequence, physical intuition on the system can be used to facilitate the design and tuning of the controller (for instance, the choice of the weights in an LQG setup [2]), as well as to gain some insight on the system structure. In particular, we focus on a state space truncation scheme based on Selective Modal Analysis (SMA) [3],[4],[5]. Although simple and straightforward to apply, this approach is amenable of different interpretations in terms of structural properties of the system. The state vector in the reduced model can be seen as an *aggregated* state [6] as well as as the outcome of a state decomposition based on the notion of *coherency* [7]. In both cases, the state vector, the approximation error, and the system parameters can be given an interpretation in terms of the same quantities appearing in the full-order model. We also describe a different approach, named *sub-structuring* [8], that can be seen as an approximate aggregation scheme, where the aggregation matrix, that is chosen on the basis of a priori knowledge on the system structure, is such that the resulting simplified system preserves some key modal properties of the full-order system.

In this note we restrict our analysis to a particular class of electromechanical systems. For the given class, the state matrix A can be written as

$$A = L^{-1}R\,, \tag{2}$$

where $L = L^T$ is invertible and R is a diagonal positive semidefinite matrix. A further assumption is that A is diagonalizable, although the majority of the considerations still hold in the general case. Our interest for such systems is related to our activity in the field of tokamak modelling and control [9, 10]. In tokamak devices, the control action is performed by using a set of coils located outside the plasma chamber to create appropriate magnetic fields which interact with the plasma, modifying its current, position, and shape. To design efficient plasma control systems, it is necessary to describe the plasma behavior and its electromagnetic coupling with the surrounding structures. The problem is clearly continuous in nature, that is, tokamaks are appropriately described by distributed parameter systems. However, it is common practice to reduce the problem to a discrete one by using Finite Elements Analysis (FEA) methods, so that the massive structures are approximated by means of toroidally symmetric elements of finite cross section, given the substantially axisymmetric nature of a tokamak device [11]. After linearization around equilibrium points, the resulting equation is

$$L\dot{I}(t) + RI(t) = TV(t)\,, \tag{3}$$

where $I(t)$ is the vector of circuit currents, $V(t)$ is a vector of control voltages, $L = L^T > 0$ is the *modified* inductance matrix [11],[12], and $R \geq 0$ is a diagonal matrix of resistances. By taking $I(t)$ as state vector, the A matrix in the usual state space representation (1) is given by (2), and belongs to the class considered here. The model order depends on the discretization grid: The more refined the grid, the higher the number of circuits or, equivalently, of state variables. In tokamak experiments, this results in models with up to several hundreds of state variables.

The note is organized as follows. Model reduction by state space truncation complemented by SMA is described in Section 2, whereas aggregation and coherency are briefly reviewed in Section 3. The relationships among these concepts for the class of electromechanical systems considered here are reported in Section 4. The sub-structuring approach is described in Section 5, and some concluding remarks are given in Section 6.

2 State space truncation and Selective Modal Analysis

A standard model reduction technique consists in approximating the time behavior of the full-order model by neglecting the contribution of the fast modes [3]. Such technique is often referred to as *modal truncation*. The state matrix is first put in Jordan form, then the states are reordered so that they can be divided into two sets, x_1 and x_2, with x_2 being associated with a given set of modes to be neglected (typically, high frequency modes). In the following all the involved matrices will be considered as partitioned according to such partition of the state vector. By taking as reduced model the subsystem corresponding to x_1, i.e., (A_{11}, B_1, C_1, D), it is guaranteed that the modes of the reduced-order model are a subset of those of the full-order model. Moreover, perfect matching between the transfer function of the reduced model and $W(s)$ at $s = \infty$ is achieved. Instead of removing some state variables from the model by setting $x_2 = 0$, one can choose to approximate them by their steady-state values ($\dot{x}_2 = 0$). This approach is named *singular perturbation approximation* [13], and it grants that the reduced-order model exhibits zero steady-state error, which is a particularly nice property in case of feed-forward control. Singular perturbation approximation is related to state-space truncation by the frequency inversion transformation $s \rightarrow 1/s$, and it can be actually be performed by truncating a realization of $W(1/s)$. The corresponding state-space description is

$$(A_{11} - A_{12}A_{22}^{-1}A_{21}, B_1 - A_{12}A_{22}^{-1}B_1, C_1 - C_2A_{22}^{-1}A_{21}, D - C_2A_{22}^{-1}B_2)\,, \tag{4}$$

where it has been assumed that A_{22} is invertible. Both modal truncation and singular perturbation approximation are based on a preliminary eigenmode analysis of the full-order model in order to select which are the modes which can be neglected in the reduced model.

A different approach to state space truncation is the one proposed in [4],[5], which is called *Selective Modal Analysis* (SMA). Truncation can be performed without a preliminary change of basis in the state space, thus granting the preservation of the physical meaning of the state variables. It is however necessary to decide *a priori* which are the states to be retained in the reduced model. The state matrix of the reduced model is then built so that the modes of the reduced-order system are a subset of those of the full-order system, as in the modal truncation approach. The state choice is performed by considering two adimensional coefficients: The *participation factor* p_{ki} measures the contribution of the k-th state in the i-th mode (the higher the value of p_{ki}, the more relevant is state k in building mode i), and it is defined as the product of the k-th components of the left and right normalized eigenvectors corresponding to λ_i

$$p_{ki} \overset{\triangle}{=} w_{ki} v_{ki} \; ; \tag{5}$$

The *participation ratio* ρ_{ri} measures the overall contribution of a set r of states in the i-th mode (an absolute value of ρ_{ri} greater than 1 indicates that the selected states give a higher contribution in forming the i-th mode than the neglected states), and it is defined as

$$\rho_{ri} \overset{\triangle}{=} \frac{\displaystyle\sum_{k=1}^{r} p_{ki}}{\displaystyle\sum_{k=r+1}^{n} p_{ki}} \; , \tag{6}$$

where n is the model order. This is used as a guideline for the selection of the order of the simplified model. It is worth noticing that the class of systems considered here enjoy the following property [14].

Proposition 1. *Assume that the matrix $A \in \mathbb{R}^{n \times n}$ is such that*

$$A = L^{-1} R \tag{7}$$

with $L = L^T > 0$ and

$$R = \begin{bmatrix} 0 & 0 \\ 0 & R_2 \end{bmatrix} \tag{8}$$

where R_2 is a diagonal, positive definite matrix. If all the nonzero eigenvalues of A are real and distinct, then $p_{ki} \geq 0$, $i = 1, \ldots, n$, $k = 1, \ldots, n$.

When the result of Proposition 1 applies, it is particularly simple to compare the contributions of each state in a given mode. This happens for the tokamak models considered in [10, 9], where the use of FEA methods in the discretization of the original distributed parameters system yields a linear system where the only multiple eigenvalue is the one in the origin of the complex plane [11, 12].

The state-space description of the reduced order model is built as follows. The states are reordered so that the ones to be retained are the first r (i.e., $x_1 = [x_1 \dots x_r]^T$). Let V be the invertible matrix of eigenvectors $V = [v_1 \dots v_n]$, and W its inverse. By means of elementary operations on the rows and columns of V, it is possible to reorder the eigenvectors v_i's so that the first r eigenvectors correspond to the r eigenvalues $\{\lambda_1, \dots, \lambda_r\}$ to be retained in the reduced model. This implies that

$$WAV = \varLambda = \begin{bmatrix} \varLambda_1 & 0 \\ 0 & \varLambda_2 \end{bmatrix} \tag{9}$$

with $\varLambda_1 = diag(\lambda_1, \dots, \lambda_r)$. The state matrices of the reduced model are

$$\begin{aligned} A_r &= A_{11} + A_{12} V_{21} V_{11}^{-1} && (10) \\ B_r &= V_{11}(W_{11} B_1 + W_{12} B_2) && (11) \\ C_r &= C_{11} + C_{12} V_{21} V_{11}^{-1} && (12) \\ D_r &= D \, . && (13) \end{aligned}$$

A simple computation shows that the eigenvalues of A_r are actually $\{\lambda_1, \dots, \lambda_r\}$, and the corresponding eigenvectors are v_{1i}, that is, the first r components of the i-th eigenvector of A with λ_i as corresponding eigenvalue. It is also straightforward to verify that the state matrix A_r in the reduced model corresponds to the one obtained by the usual modal truncation approach, that is,

$$A_r = V_{11} \varLambda_1 V_{11}^{-1} \, . \tag{14}$$

The method can be considered essentially as modal truncation with a flavor of singular perturbation approximation. To see this, consider the system in free evolution ($u(t) = 0$). The action of the neglected dynamics on the retained ones is described by the transfer function $H(s) = (sI - A_{22})^{-1} A_{21}$ between x_1 and x_2. In the modal truncation approach, since one takes $x_2 = 0$, we have that $H(s)$ is approximated by the constant value $H(s) \simeq 0$. In the singular perturbation approach, $H(s)$ is approximated by its value at $s = 0$, $H(s) \simeq -A_{22}^{-1} A_{21}$. In the SMA-based setup, the approximation is given by

$$H(s) \simeq H = V_{21} V_{11}^{-1} \, . \tag{15}$$

This grants that the effect of x_2 on x_1 when only the selected modes are excited is completely preserved. In fact, consider without loss of generality the eigenpair $(\lambda_1, v_1 = [v_{11}^T \;\; v_{12}^T]^T)$. Then

$$\lambda_1 v_{12} = A_{21} v_{11} + A_{22} v_{12} \tag{16}$$

which implies

$$v_{12} = (\lambda_1 I - A_{22})^{-1} A_{21} v_{11} = H(\lambda_1) v_{11} \, . \tag{17}$$

Since $V_{21} V_{11}^{-1} v_{11} = v_{12}$, we have that $H v_{11} = H(\lambda_1) v_{11}$.

Finally, we observe that an easy computation shows that the choice of B_r is such that the transfer function of the reduced-order system is the same as that of a reduced-order system obtained by modal truncation, when the neglected states are associated with the discarded dynamics.

3 Aggregation and coherency based decomposition

Referring only to the state equation in (1), we say that, in the system described by

$$\dot{z}(t) = Fz(t) + Gu(t) \tag{18}$$

with $z \in \mathbb{R}^r$, $r < n$, $z(t)$ is an *aggregated* state of $x(t)$, i.e., $z(t) = Ex(t)$, $E \in \mathbb{R}^{r \times n}$ if and only if

$$FE = EA\,, \qquad G = EB \tag{19}$$

and $z(0) = Ex(0)$ [6]. In this approach, each state variable in the reduced-order model is a linear combination of states of the full-order system. Existence of the aggregation matrix E is granted when F and A have common eigenvalues [15]. We assume that the common eigenvalues correspond to the modes to be preserved in the reduced-order model, that is, those of Λ_1 in (9). The aggregated state $z(t)$ has to approximate, at least for a certain class of inputs $u(t)$, a set of r components of $x(t)$, which are retained for their physical significance. The general form of the aggregation matrix is [16]

$$E = M[I_r \;\; 0]W\,, \tag{20}$$

where $M \in \mathbb{R}^{r \times r}$ is any full rank matrix, and W is as in (9). The choice of M can be made on the basis of *a priori* knowledge on the system and physical intuition, or minimizing appropriate performance indexes. A property of aggregated models that is particularly relevant for control design is given in the following proposition [17].

Proposition 2. *Let K be a $r \times m$ matrix. Then the the eigenvalues of $A - BKE$ are given by the union of those of $F - GK$ and the eigenvalues of A not retained in the aggregated model.*

Proposition 2 states that a static feedback controller K can be designed using the simplified model, then, by using K to feed back the *aggregated state* in the full order model, one is granted that a set of modes (usually, low frequency modes) are moved to the desired locations without modifying the remaining modes. In particular, if both $F - GK$ and Λ_2 in (9) are stable matrices, so is $A - BKE$.

An aggregation method based on the concept of *coherency* has been proposed in [18, 7], with particular reference to electromechanical systems. The definition of coherency is given considering the system free evolution

$(u(t) = 0)$. We say that two states x_i and x_j are *coherent* with respect to a set σ_a of r modes of A (or σ_a-*coherent*) if and only if none of these modes is observable form the difference $y_j(t) = x_j(t) - x_i(t)$ [7]. Therefore, if only the modes in σ_a are excited, for any couple of coherent states the corresponding difference $y_j(t)$ is equal to zero. By grouping states according to their coherency properties, it is possible to *decompose* the system into areas which exhibit the same behavior with respect to the given modes. Each area is represented by a *reference* state, all the other states in the area are σ_a-coherent with such state. Decomposable systems enjoy the following property [7]. Assume that the system is decomposable in exactly r areas coherent w.r.t σ_a, and reorder the state vector so that the first r components are the area references. Then, the $(n-r)$-dimensional vector y_d formed by the differences y_j relating each of the states $x_j, j = r+1, \ldots, n$ with its area reference $x_i, i = 1, \ldots, r$ form an aggregated state, $y_d = E_d x$. The corresponding dynamic equation (18) is characterized by the fact that the spectrum of F is given by the eigenvalues of A not included in σ_a. In particular, it is possible to provide the following expression for the aggregation matrix E_d

$$E_d = \left[-\Gamma \quad I_{n-r}\right], \tag{21}$$

where the area grouping information is contained in the *grouping matrix* $\Gamma \in \mathbb{R}^{(n-r)\times r}$. Observe that E_d is a matrix where the only nonzero elements are equal to one. More precisely, if the area with reference x_i has n_i states, then column i in Γ has $n_i - 1$ entries 1 defining which of the remaining $n-r$ states belong to the area i.

A general property of the grouping matrix Γ is that it can be expressed in terms of vectors forming a basis for the σ_a-eigenspace as follows: If the matrix $S \in \mathbb{R}^{n\times r}$ forms a basis for the σ_a-eigenspace, then $\Gamma = S_2 S_1^{-1}$. Furthermore, if the system is decomposable in r areas of σ_a-coherent states, then only the first r rows of S are distinct, and each of the remaining $n-r$ rows is equal to one of the first r. Such a property holds however only for systems that are exactly decomposable. In general, real systems are only approximately decomposable, in the sense that it is possible to define only sets of "nearly" coherent states. In such case, it is not possible to individuate r distinct sets of equal rows in the matrix S, and the grouping information is obtained by approximating $S_2 S_1^{-1}$ by the closest possible Γ.

4 Relations with SMA-based truncation

Let us now turn to studying the relationships existing between state space truncation in the SMA approach and coherency based aggregation and decomposition. Since all of these properties depend mainly on the modal structure of the system, in the following we will refer to systems in free evolution $(u(t) = 0)$.

Let us consider the reduced-order model obtained by state space truncation complemented by SMA. Let $x_1 = x_r$ be the r-dimensional state vector formed by the components of x to be retained in the simplified model, chosen according to the participation factors/ratios analysis, and denote by $(\sigma_a,\, \sigma_d)$ the sets of eigenvalues to be retained and neglected in the reduced model, respectively. Then, the first r columns of the eigenvector matrix V form a basis for the σ_a-eigenspace. Performing the change of basis in the state space

$$\begin{bmatrix} x_1 \\ z \end{bmatrix} = \varPi^{-1} x = \begin{bmatrix} I_r & 0 \\ -H & I_{n-r,n-r} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{22}$$

where H has been defined in (15), and using [7, Theorem 2.1], we can write

$$\begin{bmatrix} \dot{x}_1 \\ \dot{z} \end{bmatrix} = \varPi^{-1} A \varPi \begin{bmatrix} x_1 \\ z \end{bmatrix} = \begin{bmatrix} A_r & A_{12} \\ 0 & A_d \end{bmatrix} \begin{bmatrix} x_1 \\ z \end{bmatrix} , \tag{23}$$

where $A_d = A_{22} - HA_{12}$. Observe that in the SMA approach $x_2(t) \approx Hx_1(t)$, therefore $z(t) = x_2(t) - Hx_1(t)$ represents the error introduced by approximating the transfer function $H(s)$ by the constant term H. Neglecting such error (i.e., taking $z(t) = 0$), one is left with the state equation $\dot{x}_r = A_r x_r$, i.e., the dynamics of the reduced model obtained by SMA. Notice now that, since $[V_{11}^T \;\; V_{21}^T]^T$ is a basis for the σ_a-eigenspace, $H = V_{21} V_{11}^{-1}$ can be seen as an approximation for a certain grouping matrix $\varGamma$. Consequently, each of the preserved states in x_r can be considered as a reference state for a system decomposition in σ_a-coherent areas. The approximation error $z(t)$ has then the form (21) and is therefore an aggregated state

$$z(t) = E_d x(t) = [-H \quad I_{n-r}] x(t) , \tag{24}$$

and its dynamics are specified by the matrix A_d, whose eigenvalues are the elements of σ_d, given the upper-triangular form of $\varPi^{-1} A \varPi$.

Consider now the transformation

$$\begin{bmatrix} x_a \\ z \end{bmatrix} = P^{-1} \begin{bmatrix} x_1 \\ z \end{bmatrix} = \begin{bmatrix} I_r & Q \\ 0 & I_{n-r,n-r} \end{bmatrix} \begin{bmatrix} x_1 \\ z \end{bmatrix} . \tag{25}$$

After some algebra we get

$$\begin{bmatrix} \dot{x}_a \\ \dot{z} \end{bmatrix} = P^{-1} \varPi^{-1} A \varPi P \begin{bmatrix} x_a \\ z \end{bmatrix} = \begin{bmatrix} A_r & -A_r Q + A_{12} + QA_d \\ 0 & A_d \end{bmatrix} \begin{bmatrix} x_a \\ z \end{bmatrix} . \tag{26}$$

Observe that, if we neglect again the approximation error $z(t)$, we have that $x_a(t) = x_r(t)$ and (26) gives the same information as (23), independently of the particular Q in the definition of x_a. However, an appropriate choice of Q allows to derive an interesting form of the reduced-order model when $A = L^{-1}R$. In the following proposition we derive a result similar to [7, Theorem 3.1], with a different, more direct, technique.

Proposition 3. *If in (25)-(26)*

$$Q = L_a^{-1}(L_{12} + H^T L_{22})\,, \tag{27}$$

where

$$L_a = [I_r \;\; H^T]L \begin{bmatrix} I_r \\ H \end{bmatrix} = L_{11} + L_{12}H + H^T L_{12}^T + H^T L_{22}H\,, \tag{28}$$

then $x_a(t)$ is an aggregated state, $x_a(t) = E_a x(t)$, with

$$E_a = L_a^{-1}[I_r \;\; H^T]L\,, \tag{29}$$

and

$$\dot{x}_a(t) = A_r x_a(t) = L_a^{-1} R_a x_a(t) \tag{30}$$

where

$$R_a = [I_r \;\; H^T]R \begin{bmatrix} I_r \\ H \end{bmatrix} = R_1 + H^T R_2 H\,. \tag{31}$$

Proof. We first proof that $x_a = E_a x$. By definition,

$$x_a = x_1 + Q(x_2 - Hx_1) \;\; \Rightarrow \;\; L_a x_a = L_a x_1 + L_a Q(x_2 - Hx_1)\,. \tag{32}$$

Using the definition of Q and L_a given in (27) and (28), we get

$$\begin{aligned} L_a x_a &= L_a x_1 + (L_{12} + H^T L_{22})(x_2 - Hx_1) \\ &= (L_{11} + L_{12}H + H^T L_{12}^T + H^T L_{22}H)x_1 + (L_{12} + H^T L_{22})(x_2 - Hx_1) \\ &= (L_{11} + H^T L_{12}^T)x_1 + (L_{12} + H^T L_{22})x_2 \\ &= [I_r \;\; H^T]Lx \end{aligned} \tag{33}$$

from which one gets (29). Observe that L_a is invertible since $L > 0$ and $[I_r \, H^T]$ has full rank. To prove (30), we recall that

$$A \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix} = L^{-1}R \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix} = \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix} \Lambda_1\,, \tag{34}$$

or equivalently

$$\begin{cases} R_1 V_{11} = (L_{11}V_{11} + L_{12}V_{22})\Lambda_1 \\ R_2 V_{22} = (L_{12}^T V_{11} + L_{22}V_{22})\Lambda_1 \end{cases}. \tag{35}$$

Right-multiplying both equations in (35) by V_{11}^{-1} and taking V_{11} out of the terms in round brackets in the right-hand sides of (35) yields

$$\begin{cases} \;\;\, R_1 = (L_{11}V_{11} + L_{12}V_{22})V_{11}\Lambda_1 V_{11}^{-1} \\ R_2 H = (L_{12}^T V_{11} + L_{22}V_{22})V_{11}\Lambda_1 V_{11}^{-1} \end{cases}. \tag{36}$$

Right-multiplying the second equation in (36) by H^T and summing it to the first equation, we obtain

$$R_1 + H^T R_2 H = (L_{11} + L_{12}H + H^T L_{12}^T + H^T L_{22} H) V_{11} \Lambda_1 V_{11}^{-1} \,, \qquad (37)$$

that gives (30) recalling (14). ■

Proceeding as in [7, Theorem 3.1], one can also show that, by choosing

$$Q = -Y_1 Y_2^{-1} \,, \qquad (38)$$

where $[Y_1^T \;\; Y_2^T]^T$ is a basis for the σ_a-eigenspace of

$$\Pi^{-1} A \Pi = \begin{bmatrix} A_r & A_{12} \\ 0 & A_d \end{bmatrix} \qquad (39)$$

we have that

$$-A_r Q + A_{12} + Q A_d = 0 \,, \qquad (40)$$

thus showing more clearly how the dynamics of x_a and z in (26) are specified by σ_a and σ_d, respectively. It is not difficult to show that Q as defined in (27) can be written in the form (38) for appropriate Y_1, Y_2.

We conclude this Section with some comments on the results derived above.

1. The fundamental element in both the SMA-based truncation scheme and coherency-based decomposition is the matrix H defined in (15). In the SMA approach, it approximates a matrix transfer function, and it provides a way for representing the action of the neglected states in the simplified model by means of the preserved states. In the context of coherency, the matrix H gives an approximation to a grouping matrix Γ, thus providing a clue on how far from being exactly decomposable the considered system is.
2. In the analysis of the coherency properties of the system, two quantities have been introduced, $z(t)$ and $x_a(t)$, that have been both proved to be aggregated states with dynamics given by A_d and A_r, respectively. As far as $z(t)$ is concerned, it has a natural interpretation in the SMA-based truncation approach as an approximation error, and its expression (24) as an aggregated state is meaningful independently of the properties of the system in terms of coherency. On the other hand, the expression (30) of the dynamics of $x_a(t)$ crucially depends on the particular choice of Q in (25).
3. As shown in [16], the state in the reduced model of [3] can be seen as an aggregated state, and the corresponding aggregation matrix E_s is given by choosing $M = V_{11}$ in (20). It is interesting to compare E_s with E_a in (29) that defines the aggregated state x_a in the coherency based approach. In fact, we observe that the main difference existing between the truncation approach and coherency based aggregation consists in the fact that in the first case no trace is left in the simplified model of the neglected variables x_2, whereas their effect is considered in the second approach in

the definitions of both $z(t)$ and $x_a(t)$. Such difference is highlighted in the comparison of E_s and E_a. After some algebra we get

$$E_s - E_a = L_a^{-1}[I_r \;\; H^T]\left(L\begin{bmatrix}V_{11} & 0\\ V_{21} & 0\end{bmatrix} - LV\right)W \,, \tag{41}$$

which clearly shows that the difference between the two aggregation matrices is actually introduced by the operation of truncating the state vector to its first r components.

4. The expression obtained in (30) for the state matrix of the reduced model lends itself to a very clear interpretation in terms of physical quantities. Indeed, the factorization of A_r as $L_a^{-1}R_a$ shows that the reduced model maintains the same nature of the full-order model. In the case of generalized electrical networks, the elements of L and R are inductances and resistances, respectively, then equation (30) allows to see the reduced model as a second network with a reduced number of circuits, where some of the inductances and resistances have been grouped together according to the modal properties of the system. We stress that the matrix $R_a > 0$ is *not* a diagonal matrix if the system is not perfectly decomposable in r areas. Its departure from diagonality depends on how far the matrix H is from a real grouping matrix.

5 Sub-Structuring Method

The rationale of the sub-structuring approach to modal reduction presented in [8] is that of splitting the physical system into sub-structures. Then in the simplified model a single state (or few states) is used to describe the dynamic of each sub-structure and its effect on the output. It turns out that the order of the reduced model is equal (or proportional) to the number of the sub-structures forming the system. The main difference with the above approaches is that typically only few selected eigenvalues from the full order system are retained in the low-order model.

To characterize the modal properties of the full- and reduced- order models, in particular as far as stability is concerned, we need to recall some definition and properties of the inertia of a symmetric matrix.

Definition 1. *Let X be a real symmetric matrix. Then the* inertia *of X, $\mathbb{I}n(X)$, is the ordered triple*

$$(p^X, z^X, n^X)\,, \tag{42}$$

where p^X, z^X, n^X are the number of positive, zero, and negative eigenvalues of X, respectively.

If X is non-singular, then $z^X = 0$ and we will omit to indicate it in $\mathbb{I}n(X)$. A relevant property of the inertia of hermitian matrices is that it is invariant under congruence transformations, as stated in the following theorem.

Theorem 1 (Sylvester's law of inertia [19]). *Let $X, Y \in \mathbb{C}^{n\times n}$ be Hermitian matrices. There is a non-singular matrix $M \in \mathbb{C}^{n\times n}$ such that $Y = M^H X M$ if and only if*

$$\mathbb{I}n(X) = \mathbb{I}n(Y)\ . \tag{43}$$

For the class of systems considered in this paper, i.e. those described by the state-space model (3), the following result holds.

Theorem 2. *Let L be a non-singular symmetric matrix, $R > 0$, and $A = -L^{-1}R$. Then A has only real eigenvalues, it is diagonalizable, and*

$$\mathbb{I}n(A) = \mathbb{I}n(-L)\ . \tag{44}$$

Proof. see [19, Theorem 7.6.3]. ■

Remark 1: If $R \geq 0$, with $\mathrm{rank}(R) = q < n$, then A has $z^A = z^R = n - q$ zero eigenvalues. If this is the case, then, by taking the change of basis specified by the unitary matrix U that diagonalizes R,

$$\bar{R} = U^T R U = \begin{bmatrix} \bar{R}_1 & 0 \\ 0 & 0 \end{bmatrix}\ , \tag{45}$$

$$\bar{L} = U^T L U = \begin{bmatrix} \bar{L}_{11} & \bar{L}_{12} \\ \bar{L}_{12}^T & \bar{L}_{22} \end{bmatrix}\ , \tag{46}$$

with $\bar{R}_1 \in \mathbb{R}^{q\times q}$, $\bar{R}_1 > 0$, it is easy to show that

$$\mathbb{I}n(A) = (p^{\tilde{L}}, n - q, n^{\tilde{L}})\ , \tag{47}$$

where $\tilde{L}$ is the Schur complement of $\bar{L}_{11}$ in $\bar{L}$, i.e., $\tilde{L} = \bar{L}_{11} - \bar{L}_{12}^T \bar{L}_{22}^{-1} \bar{L}_{12}$.

Remark 2: As a consequence of Theorem 2, if $R > 0$, then the system is asymptotically stable (i.e. all the eigenvalues in the system spectrum $\sigma(L, R)$ are in $\mathbb{C}^-$, where $\mathbb{C}^-$ is the open left half plane of $\mathbb{C}$) if and only if $L > 0$.

We proceed now to describe the sub-structuring approach to model reduction. To begin with we must identify and group the currents flowing in the different structural elements (sub-structures). In general we can re-order and split the inductance and resistance matrix in N-blocks describing the different currents flowing in the N sub-structures and write

$$\begin{bmatrix} L_{11} & L_{12} & \dots & L_{1N} \\ L_{21} & L_{22} & \dots & L_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ L_{N1} & L_{N2} & \dots & L_{NN} \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_N \end{bmatrix} + \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1N} \\ R_{21} & R_{22} & \dots & R_{2N} \\ \vdots & \vdots & \ddots & \dots \\ R_{N1} & \dots & \dots & R_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_N \end{bmatrix} v \tag{48}$$

$$y = \begin{bmatrix} C_1 & C_2 & \dots & C_N \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}\ . \tag{49}$$

Then, as in the modal truncation approach, it is important to identify the $r < n$ most relevant modes to be used to well approximate the input-output mapping. For clarity of exposition, we first assume that the relevant system dynamics are associated with one eigenpair, $(\lambda_k, x^{(k)})$. These have to be faithfully reproduced in the reduced model. By solving the eigenvalue problem

$$\lambda_k \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1N} \\ L_{21} & L_{22} & \dots & L_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ L_{N1} & L_{N2} & \dots & L_{NN} \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_N^{(k)} \end{bmatrix} = - \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1N} \\ R_{21} & R_{22} & \dots & R_{2N} \\ \vdots & \vdots & \ddots & \dots \\ R_{N1} & \dots & \dots & R_{NN} \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_N^{(k)} \end{bmatrix} \tag{50}$$

we can assemble the block-diagonal sub-structuring matrix $S^{(k)}$ for the k-th eigenpair as:

$$S^{(k)} \triangleq \begin{bmatrix} x_1^{(k)} & 0 & \dots & 0 \\ 0 & x_2^{(k)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_N^{(k)} \end{bmatrix} . \tag{51}$$

It is worth noticing that the matrix $S^{(k)}$ has dimension $n \times N$ where $n = \dim(L)$ whereas N is the number of sub-structures which have to be retained in the reduced model. We can now reduce the system order by pre/post multiplying the system matrices as follows:

$$L^{(k)} = S^{(k)T} L S^{(k)} \tag{52a}$$

$$R^{(k)} = S^{(k)T} R S^{(k)} \tag{52b}$$

$$T^{(k)} = S^{(k)T} T \tag{52c}$$

$$C^{(k)} = C S^{(k)} . \tag{52d}$$

The square matrices $L^{(k)*}$ and $R^{(k)}$ have dimension N and are, respectively, symmetric and symmetric, positive semi-definite. This allows us to give a physical interpretation to the parameters of the reduced-order model, whose state variables represent currents flowing in the sub-structures. The dynamics of the reduced order model are given by

$$\begin{cases} L^{(k)}\dot{z}(t) + R^{(k)}z(t) = T^{(k)}v(t) \\ \hfill y(t) = C^{(k)}z(t) \end{cases} , \tag{53}$$

and they can be thought as obtained from (3) by premultiplying the first equation by $S^{(k)T}$ and assuming that the state vector z in the reduced order system is related to x by means of the following approximate relation

$$x(t) \simeq S^{(k)} z(t) . \tag{54}$$

Relation (54) can be solved for $z(t)$ in a least square sense giving, if $S^{(k)}$ is full rank,

$$z(t) = (S^{(k)T}S^{(k)})^{-1}S^{(k)T}x(t)\ , \tag{55}$$

and the matrix

$$E \triangleq (S^{(k)T}S^{(k)})^{-1}S^{(k)T} \tag{56}$$

could be considered as an (approximate) aggregation matrix. However, in general the state and input matrices in the full and reduced order models do not satisfy relations (19), so that the sub-structuring approach does not yield exactly aggregated models. Observe that, given its structure, $S^{(k)}$ is not full rank if and only if there exists at least an index i such that $x_i^{(k)} = 0$. Therefore, the full rank assumption on $S^{(k)}$ implies that all the sub-structures are "involved" by excitation of the mode associated with λ_k.

It is easy to show that the eigenvalue λ_k is exactly preserved in the reduced order model, as stated in the following proposition.

Proposition 4. $\lambda_k \in \sigma(L^{(k)}, R^{(k)})$.

Proof. From (51), we have that $x^{(k)}$ belongs to the range of $S^{(k)}$. Let $\alpha^{(k)}$ be such that

$$x^{(k)} = S^{(k)}\alpha^{(k)}\ . \tag{57}$$

Then from (50)

$$\lambda_k L S^{(k)}\alpha^{(k)} = -RS^{(k)}\alpha^{(k)} \tag{58}$$

and by premultiplication of (58) by $S^{(k)T}$ we get

$$\lambda_k L^{(k)}\alpha^{(k)} = -R^{(k)}\alpha^{(k)}\ , \tag{59}$$

showing that $(\lambda_k, \alpha^{(k)})$ is an eigenpair for the reduced order model. ■

We remark that the remaining eigenvalues of the reduced system do not in general coincide with eigenvalues of the full order model, that is,

$$\sigma(L^{(k)}, R^{(k)}) \not\subset \sigma(L, R)\ . \tag{60}$$

However,

$$\sigma(L^{(k)}, R^{(k)}) \cap \sigma(L, R) \neq \emptyset\ , \tag{61}$$

and, as is known, this implies that there exist non-trivial solutions E to the first matrix equation in (19) [15]. However, these solutions yield trivial aggregated systems.

The modal properties relevant to stability of the reduced system can be derived as follows. It is a simple exercise to prove the following result (see also [19]):

Lemma 1. *Let the non-singular symmetric matrix X be partitioned as*

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^T & X_{22} \end{bmatrix}\ . \tag{62}$$

Then, if $\mathbb{I}n(X) = (p^X, n^X)$ *and* $\mathbb{I}n(X_{11}) = (p^{X_{11}}, n^{X_{11}})$,

$$p^{X_{11}} \leq p^X\ , \qquad n^{X_{11}} \leq n^X\ . \tag{63}$$

Proof. Consider the non-singular matrix P defined as

$$P = \begin{bmatrix} I & -X_{11}^{-1}X_{12} \\ 0 & I \end{bmatrix} . \tag{64}$$

Then

$$P^T X P = \bar{X} = \begin{bmatrix} X_{11} & 0 \\ 0 & X_{22} - X_{12}^T X_{11}^{-1} X_{12} \end{bmatrix} . \tag{65}$$

By Sylvester's theorem, $\mathbb{I}n(X) = \mathbb{I}n(\bar{X})$, and since $\sigma(\bar{X}) = \sigma(X_{11}) \cup \sigma(X_{22} - X_{12}^T X_{11}^{-1} X_{12})$, we have that

$$p^{X_{11}} \leq p^{\bar{X}} = p^X , \qquad n^{X_{11}} \leq n^{\bar{X}} = n^X . \tag{66}$$

■

We can prove now the following theorem, that relates the inertia of the system matrices of the full- and reduced-order models.

Theorem 3. *Let $L^{(k)}$ be defined as in (52a), with $S^{(k)}$ of full rank N. Then, if $\mathbb{I}n(L) = (p^L, n^L)$ and $\mathbb{I}n(L^{(k)}) = (p^{L^{(k)}}, n^{L^{(k)}})$, we have that*

$$p^{L^{(k)}} \leq p^L , \qquad n^{L^{(k)}} \leq n^L . \tag{67}$$

Proof. By computing the Singular Value Decomposition of $S^{(k)}$, we can find a unitary matrix $U \in \mathbb{R}^{n \times n}$ such that

$$U^T S^{(k)} = \begin{bmatrix} S_1 \\ 0 \end{bmatrix} , \tag{68}$$

where $S_1 \in \mathbb{R}^{N \times N}$ is full rank. Let $\bar{L} = U^T L U$. Since $\bar{L}$ and L are similar, $\mathbb{I}n(\bar{L}) = \mathbb{I}n(L)$. Now, we have that

$$\begin{aligned} L^{(k)} &= S^{(k)T} L S^{(k)} = S^{(k)T} U \bar{L} U^T S^{(k)} \\ &= [S_1^T \ \ 0] \begin{bmatrix} \bar{L}_{11} & \bar{L}_{12} \\ \bar{L}_{12}^T & \bar{L}_{22} \end{bmatrix} \begin{bmatrix} S_1 \\ 0 \end{bmatrix} \\ &= S_1^T \bar{L}_{11} S_1 . \end{aligned} \tag{69}$$

From Sylvester's theorem, $\mathbb{I}n(L^{(k)}) = \mathbb{I}n(\bar{L}_{11})$, and from Lemma 1,

$$p^{L^{(k)}} = p^{\bar{L}_{11}} \leq p^{\bar{L}} = p^L , \tag{70}$$

$$n^{L^{(k)}} = n^{\bar{L}_{11}} \leq n^{\bar{L}} = n^L . \tag{71}$$

■

As a consequence of Theorem 3, if the full-order system is stable, so is the reduced-order system, as stated in the following Corollary.

Corollary 1. *If* $\sigma(L,R) \subset \mathbb{C}^-$, *then* $\sigma(L^{(k)}, R^{(k)}) \subset \mathbb{C}^-$.

The procedure can be easily extended to deal with the retention of $r > 1$ modes in the reduced model. In this case, we solve r eigenvalue problems (50) for the r eigenpairs $(\lambda_i, x^{(i)})$, $i = 1 \dots r$, and modify accordingly the sub-structuring matrix. For instance, if $r = 2$, S is given by

$$S \triangleq \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_2^{(1)} & x_2^{(2)} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & x_N^{(1)} & x_N^{(2)} \end{bmatrix}, \qquad (72)$$

yielding a reduced model of order rN. It is then clear that, to achieve a substantial order reduction, it is necessary to limit both the number of sub-structures and the number of modes to be preserved in the reduced order model.

6 Conclusions

In this note we reviewed some model reduction techniques for generalized RL networks. Such techniques share the property that the variables and parameters of the reduced model maintain the physical meaning of the full-order model. This fact can be useful for both the analysis of the dynamical properties of the system and controller synthesis. The interplay between SMA-based model reduction, state aggregation and coherency has been discussed, showing how the three approaches can be related to one another.

As a final remark, we point out that the reduction schemes described above proved to be useful in the control of tokamak devices. SMA-based model truncation was succesfully employed in the model reduction problem for the ITER machine [20]. The analysis of the participation factors/ratios allowed to provide some clues on how the different structures contribute to forming the unstable mode and other modes that are particularly relevent for control design [9, 10, 14]. In particular, the properties of the reduced model seen in terms of aggregation have been exploited in [14] in the design of an LQG controller, whose tuning has been simplified by the preserved physical meaning of the state variables. Finally, the sub-structuring approach has been applied in [8] to derive very low-order models of the ITER tokamak that were successfully employed to design vertical stabilizing controllers.

References

1. Green, M. and D.J.N. Limebeer (1995). *Linear Robust Control.* Prentice-Hall. Englewood Cliffs, NJ.

2. Anderson, B. and J.B. Moore (1989). *Optimal Control. Linear Quadratic Methods.* Prentice-Hall. Englewood Cliffs, NJ.
3. Davison, E. (1966). *IEEE Trans. Aut. Contr.* AC-25, 93–101.
4. Verghese, G., I. Pérez-Arriaga and F.C. Schweppe (1982), *Circuits, Systems, Signal Proc.*, I, 433–445.
5. Pérez-Arriaga, I., G. C. Verghese, F.L. Pagola, J.L. Sancha and F.C. Schweppe (1990), *Automatica*, 26(2), 215–231.
6. Aoki, M. (1969), *IEEE Trans. Aut. Contr.*, AC-13, 246–253.
7. Kokotovic, P., B. Avramovic, J.H. Chow and J.R. Winkelman (1982), *Automatica*, 18(1), 47–56.
8. Beghi A. and A. Portone (2002). Model reduction by sub-structuring. In: *Proc. of the 10th Mediterranean Conference on Control and Automation*, MED 2002, Published on CD-ROM, paper nr. 331.
9. Beghi, A., D. Ciscato and A. Portone (1997). In *Proc. of the 36th IEEE Conf. on Decision and Control.* Vol. 4. San Diego. pp. 3691–3696.
10. Beghi, A., D. Ciscato, M. Cavinato and G. Marchiori (1998). In *B. Beaumont, P. Libeyre, B. de Gentile and G. Tonon (Eds.), Proc. of the 20th Symposium on Fusion Technology.* Vol. 1. Marseille, France. pp. 507–510.
11. Albanese, R., G. Ambrosino, E. Coccorese, F. Morabito, A. Pironti, G. Rubinacci and S. Scala (1996), *Fusion Technology*, 30, 167–183.
12. Albanese, R. and F. Villone (1998), *Nuclear Fusion*, 38, 723–738.
13. Kokotovic, P., R.E. OMalley and P. Sannuti (1976), *Automatica*, 12, 123–132.
14. Beghi, A. (2001), *IEEE Trans. Contr. Syst. Tech*, 9(4), 574–589.
15. Gantmacher, F. R. (1960). *The theory of matrices* - Vol I and II. Chelsea Publishing Company. New York.
16. Siret, J., G. Michailesco and P. Bertrand (1977), *Int. J. Control*, 26(1), 121–128.
17. Lamba, S. and S. Vittal Rao (1974), *IEEE Trans. Aut. Contr.*, AC-19, 448–450.
18. Avramovic, B., P.V. Kokotovic, J.R. Winkelman and J.H. Chow (1980), *Automatica*, 16, 637–648.
19. Horn, R.A. and C.R. Johnson (1985). *Matrix analysis.* Cambridge: Cambridge University Press, 1985.
20. ITER EDA Final Design Report (1998). Poloidal Field Control. Design Description Document, WBS 4.7, Naka, Japan.

Robust Design of Standard Controllers Under Plant Parameters Uncertainty

Franco Blanchini[1], Stefano Miani[2], and Umberto Viaro[2]

[1] Dipartimento di Matematica e Informatica, Università di Udine, via delle Scienze, 208, 33100 Udine, Italy
blanchini@dimi.uniud.it

[2] Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica, Università di Udine, via delle Scienze, 208, 33100 Udine, Italy
miani.stefano@uniud.it viaro@uniud.it

Summary. This note extends some results presented in [3] concerning the characterization of standard three–parameter controllers satisfying given H_∞ closed–loop specifications. In particular, it is shown that: (i) given the value of a specific parameter, the region where the other two parameters ensure the satisfaction of the considered H_∞ constraint is formed by the union of (a bounded number of) disjoint convex sets even when the plant parameters are not fixed but belong to given intervals, and (ii) the determination of these sets entails the repeated solution of an optimization problem. The procedure is applied to find the parameters of a PID controller for an uncertain electric drive so as to ensure that the H_∞ norm of the sensitivity function is less than a prescribed value.

Keywords. H_∞ control, parametric uncertainty, PID controllers, lead/lag controllers, parameter space, sensitivity.

1 Introduction and motivation

PID and lead/lag controllers have continued to be of interest to control engineers because of their widespread industrial use. Recently, the attention of researchers has mainly concentrated on robustness issues of PID control (cf., e.g., [9] and bibliography therein), with particular regard to the possibility of ensuring an acceptable system behaviour even in the presence of large plant uncertainties or perturbations. Antonio Lepschy considered such problems in several papers that span over many years (cf., e.g., [1] ÷ [6]); these authors had the privilege of working with Him on this subject.

Phase and gain margins have traditionally been important measures of system robustness since if either is small, the feedback control system is close to instability. However, both of these margins can be large and yet the Nyquist diagram of the loop transfer function can pass close to the critical point $-1 + \jmath 0$. A better measure of stability robustness is provided by the distance from the critical point to the nearest point on the Nyquist plot of the loop transfer function. As is well known, this distance is the reciprocal of the H_∞ norm of the sensitivity function that coincides with the reference–to–error transfer function. On the other hand, the H_∞ norm provides a meaningful measure of robustness with respect to other closed–loop functions too, such as the complementary sensitivity function (equal to the reference–to–output transfer function) or a weighted sum of the sensitivity and complementary sensitivity functions [7].

In the following sections, we refer to a general form of closed–loop function and consider the problem of determining the parameters of a standard controller for the usual one–degree-of–freedom unity–feedback system in such a way that the H_∞ norm of the chosen closed–loop function satisfies a given bound. The same problem has been tackled in [8], [10] using a generalization of the Hermite–Biehler theorem and in [3] using simple properties of Möbius transformations. However, in these papers the parameters of the controlled plant are assumed to be fixed. Here, instead, we only assume that the plant parameters lie inside certain intervals (as is the case in [6] where the simple PI control of a time–delayed first–order process is considered), and look for the controller parameters that ensure the satisfaction of the constraint on the H_∞ norm for all of the admissible combinations of plant parameters. In particular, according to geometric arguments similar to those developed in [3], we provide a characterization of the admissible controller parameter regions and suggest a procedure to find their boundaries. The method is finally applied to the synthesis of the PID controller for a *dc*–motor drive with bounded sensitivity.

2 Notation and problem statement

Let us denote the *strictly proper* rational transfer function of the controlled plant by $P(s; q)$, where $q \in Q \subset \mathbf{R}^m$ is the vector of uncertain numerator and denominator coefficients whose values belong to given intervals, and assume that the controller transfer function is

$$C(s) = \frac{x + ys + zs^2}{dx + s + dzs^2} \tag{1}$$

where x, y, z are real parameters to be determined according to certain H_∞ specifications, and d is a real parameter fixed *a priori*, e.g., to obtain the desired steady–state precision.

When $d = 0$, (1) takes the form of a PID controller and x, y, z correspond, respectively, to the integral, proportional and derivative gain. When $d = 1$, (1) takes the form of the classic unit–gain lead/lag controller.

Most of the closed–loop transfer functions of the considered unity–feedback system can be expressed as

$$F(s;q) = \frac{A(s)C(s)P(s;q) + B(s)}{1 + C(s)P(s;q)} \tag{2}$$

where $A(s)$ and $B(s)$ are suitable rational functions. Function (2) may also represent the weighted sum of the sensitivity function $S(s;q) = 1/[1 + C(s)P(s;q)]$ and complementary sensitivity function $T(s;q) = C(s)P(s;q)/[1+C(s)P(s;q)]$. For simplicity, it is assumed that $A(s)$ and $B(s)$ are *proper and stable*.

The problem we consider can be stated as follows.

Problem 1. Find x, y, z in such a way that

$$\sup_{\omega \geq 0} |F(\jmath\omega; q)| \leq \gamma \ \ , \ \forall q \in Q \tag{3}$$

where γ is an assigned positive real number.

Observe that assuming γ constant entails no restriction because any weighting function can be absorbed into $A(s)$ and $B(s)$.

3 Geometric interpretation of condition (3)

For $s = \jmath\omega$, (1) can be written as

$$C(\jmath\omega) = \frac{\lambda + \jmath\omega y}{d\lambda + \jmath\omega} \tag{4}$$

where

$$\lambda := x - \omega^2 z \,. \tag{5}$$

Therefore, parameter y influences only the imaginary part of the numerator of (4), and parameters x and z influence only the real parts of both the numerator and the denominator where they appear in the linear combination (5).

Taking account of (4), the harmonic response of the closed–loop function (2) can be expressed as

$$F(\jmath\omega, \lambda, y, q) = \frac{w_1(\jmath\omega, y, q)\lambda + w_2(\jmath\omega, y, q)}{w_3(\jmath\omega, y, q)\lambda + w_4(\jmath\omega, y, q)} \tag{6}$$

where $w_i(\jmath\omega, y, q)$, $i = 1 \div 4$, are analytic functions of their arguments. For fixed ω, y and q, and $w_1 w_3 \neq w_2 w_4$, (6) defines a Möbius transformation from λ to F so that, as λ varies over the extended real axis $\mathcal{R}_e = [-\infty, +\infty]$, F describes a circle $\Phi_{\omega,y,q}$.

Given ω, y and q, the values of λ satisfying the constraint

$$|F(\jmath\omega, \lambda, y, q)| \leq \gamma \tag{7}$$

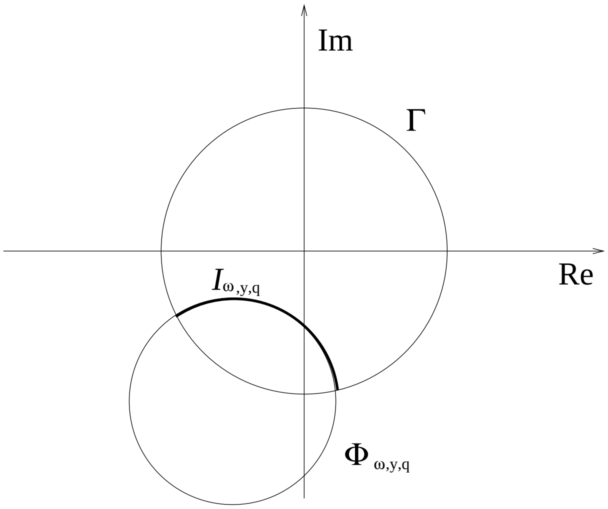


Fig. 1. Given ω, y and q, the values of λ in (5) satisfying (7) correspond to the intersection $\mathcal{I}_{\omega,y,q}$ (bold line) of the circle $\varPhi_{\omega,y,q}$ described by (6) as λ varies over $\mathcal{R}_e$ with the disk $\varGamma$ centred at the origin.

correspond to the intersection $\mathcal{I}_{\omega,y,q}$ of the aforementioned circle $\varPhi_{\omega,y,q}$ with the disk $\varGamma$ of radius γ centred at the origin, as shown in Figure 1.

Of course, circle $\varPhi_{\omega,y,q}$ may be external to disk $\varGamma$, in which case $\mathcal{I}_{\omega,y,q}$ is empty and no value of λ satisfies (7) (from the practical point of view, the case in which $\varPhi_{\omega,y,q}$ is tangent to the boundary of $\varGamma$ from the outside and, thus, $\mathcal{I}_{\omega,y,q}$ reduces to one point only, may be assimilated to this case). If, instead, circle $\varPhi_{\omega,y,q}$ is internal to $\varGamma$ or tangent to its boundary from the inside, (7) is satisfied for all values of λ.

Assuming that $\varPhi_{\omega,y,q}$ crosses (twice) the boundary of $\varGamma$ and denoting by $\lambda_i(\omega,y,q)$ and $\lambda_s(\omega,y,q) > \lambda_i(\omega,y,q)$ the values taken by λ at the two intersections, that is, at the extremes of arc $\mathcal{I}_{\omega,y,q}$, two situations may occur: (i) the values of λ corresponding to the points of $\mathcal{I}_{\omega,y,q}$ belong to the interval

$$\varLambda(\omega,y,q) := [\lambda_i(\omega,y,q), \lambda_s(\omega,y,q)] \tag{8}$$

which happens if $|F(\jmath\omega,\pm\infty,y,q)| = \left|\dfrac{w_1(\jmath\omega,y,q)}{w_3(\jmath\omega,y,q)}\right| > \gamma$;

(ii) the values of λ corresponding to the points of $\mathcal{I}_{\omega,y,q}$ belong to (the closure of) the complement of $\varLambda(\omega,y,q)$, that is, to

$$\varLambda^c(\omega,y,q) = [-\infty, \lambda_i(\omega,y,q)] \bigcup [\lambda_s(\omega,y,q), +\infty] \tag{9}$$

which happens if $|F(\jmath\omega,\pm\infty,y,q)| = \left|\dfrac{w_1(\jmath\omega,y,q)}{w_3(\jmath\omega,y,q)}\right| < \gamma$.

In case (i), the values of parameters z, x which are admissible at frequency ω are those *inside* the stripe $\mathcal{S}_{\omega,y,q} := \{(z,x)\,|\lambda_i(\omega,y,q) \leq x - \omega^2 z \leq \lambda_s(\omega,y,q)\}$ of the (z,x)–plane included between the straight lines

$$x = \omega^2 z + \lambda_i(\omega,y,q) \tag{10}$$

$$x = \omega^2 z + \lambda_s(\omega,y,q)\,. \tag{11}$$

In case (ii), the admissible parameter values are those *outside* $\mathcal{S}_{\omega,y,q}$. The two situations are illustrated in Figure 2.

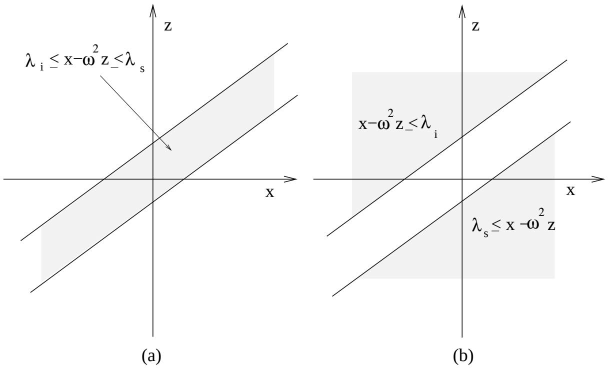

z
$\lambda_i \leq x-\omega^2 z \leq \lambda_s$
x
(a)
z
$x-\omega^2 z \leq \lambda_i$
x
$\lambda_s \leq x - \omega^2 z$
(b)

Fig. 2. (a) Case in which the admissible values of x and z are those inside the stripe $\mathcal{S}_{\omega,y,q} := \{(x,z)\,|\lambda_i(\omega,y,q) \leq x - \omega^2 z \leq \lambda_s(\omega,y,q)\}$. (b) Case in which the admissible values of x and z are those outside $\mathcal{S}_{\omega,y,q}$.

The locus described by (6) as λ varies over $\mathcal{R}_e$ is a circle for all values of q, but the its centre and diameter change, in general, with q as shown in Figure 3. The values $\lambda_i(\omega,y,q)$ and $\lambda_s(\omega,y,q)$ taken by λ at the extremes of the intersection arc $\mathcal{I}_{\omega,y,q}$, if any, change too, in general.

Note that $\lambda_i(\omega,y,q)$ and $\lambda_s(\omega,y,q)$ can easily be computed as the real roots, if any, of the second–degree polynomial equation obtained, for given ω, y, q, from the equation $|F(\jmath\omega,\lambda,y,q)| = \gamma$ with $F(\jmath\omega,\lambda,y,q)$ as in (6).

4 Admissible parameter region

To find the admissible region of the controller parameters x and z, it is necessary to determine first the range of the admissible values of λ in (5) for all q, given the values of ω and y.

If $|F(\jmath\omega,\pm\infty,y,q)| \geq \gamma$ and $\mathcal{I}_{\omega,y,q}$ is nonempty, $\forall q \in Q$ (if $\mathcal{I}_{\omega,y,q}$ is empty for at least one value of q, no value of λ satisfies (7) and, thus, (3)), the problem is that of finding the set

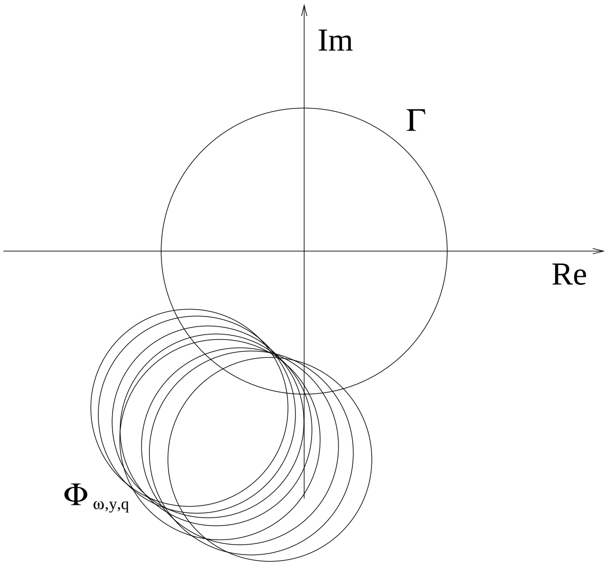


Fig. 3. Intersections $\mathcal{I}_{\omega,y,q}$ of $\Phi_{\omega,y,q}$ with Γ for various values of q with ω and y fixed.

$$\overline{\Lambda}(\omega, y) = \bigcap_q \Lambda(\omega, y, q) \tag{12}$$

which is the intersection of the intervals (8) associated with all of the values of parameter q. When $\overline{\Lambda}(\omega, y)$ is nonempty, (12) consists of a single interval, because the extremes $\lambda_i(\omega, y, q)$ and $\lambda_s(\omega, y, q)$ of $\Lambda(\omega, y, q)$ are continuous functions of q. In this case, the lower and upper extremes of $\overline{\Lambda}(\omega, y)$ are given, respectively, by

$$\overline{\lambda}_i(\omega, y) = \max_q \lambda_i(\omega, y, q) \tag{13}$$

$$\overline{\lambda}_s(\omega, y) = \min_q \lambda_s(\omega, y, q)\,. \tag{14}$$

If $|F(\jmath\omega, \pm\infty, y, q)| \leq \gamma$, $\forall q \in Q$, and $\mathcal{I}_{\omega,y,q}$ is nonempty for at least one value of q (if $\mathcal{I}_{\omega,y,q}$ is empty for all values of q, then all values of λ satisfy (7)), the problem is that of finding the right extreme $\overline{\lambda}_i(\omega, y)$ of the half–line

$$[-\infty, \overline{\lambda}_i(\omega, y)] = \bigcap_q [-\infty, \lambda_i(\omega, y, q)] \tag{15}$$

and the left extreme $\overline{\lambda}_s(\omega, y)$ of the half–line

$$[\overline{\lambda}_s(\omega, y), +\infty] = \bigcap_q [\lambda_s(\omega, y, q), +\infty] \tag{16}$$

where the intersections (15) and (16) are limited to the values of q such that $\mathcal{I}_{\omega,y,q}$ is nonempty. In this case

$$\overline{\lambda}_i(\omega, y) = \min_q \lambda_i(\omega, y, q) \tag{17}$$

$$\overline{\lambda}_s(\omega, y) = \max_q \lambda_s(\omega, y, q) \tag{18}$$

and the set of the admissible values of λ is

$$\Lambda^c(\omega, y) = [-\infty, \overline{\lambda}_i(\omega, y)] \bigcup [\overline{\lambda}_s(\omega, y), +\infty] \,. \tag{19}$$

The problem of finding the admissible values of λ is a little more complicated when $|F(\jmath\omega, \pm\infty, y, q)| \geq \gamma$ for some values of q and $|F(\jmath\omega, \pm\infty, y, q)| \leq \gamma$ for some others, that is, when the point $F(\jmath\omega, \pm\infty, y, q) = \dfrac{w_1(\jmath\omega, y, q)}{w_3(\jmath\omega, y, q)}$ crosses the boundary of Γ as q varies. In this case, it is necessary to consider separately the sets $Q_o := \{q : |F(\jmath\omega, \pm\infty, y, q)| \geq \gamma\}$ and $Q_i := \{q : |F(\jmath\omega, \pm\infty, y, q)| \leq \gamma\}$. The overall set of admissible λ–values is then the intersection, if any, of the two sets corresponding to Q_o and Q_i, respectively. For simplicity, in the following we exclude this rare situation. Therefore, the pairs (x, z) of controller parameters that satisfy constraint (7), $\forall q \in Q$, belong either to the set

$$\mathcal{S}_{\omega,y} := \{(x, z) \mid \overline{\lambda}_i(\omega, y) \leq x - \omega^2 z \leq \overline{\lambda}_s(\omega, y)\} \tag{20}$$

consisting of a stripe of the (x, z)–plane, or to the set

$$\mathcal{S}^c_{\omega,y} := \{(x, z) \,|\, x - \omega^2 z \leq \overline{\lambda}_i(\omega, y) \text{ or } x - \omega^2 z \geq \overline{\lambda}_s(\omega, y)\} \tag{21}$$

consisting of the half–plane below the straight line $x = \omega^2 z + \overline{\lambda}_i(\omega, y)$ and the half–plane above the straight line $x = \omega^2 z + \overline{\lambda}_s(\omega, y)$.

The above analysis must be carried out for all values of ω (in practice, for a suitable number of frequency samples). As a result, the positive semi–axis $\Omega_+ := \{\,\omega \,|\, \omega \geq 0\}$ can be subdivided into consecutive intervals $J_k := \{\,\omega \,|\, \omega_k \leq \omega \leq \omega_{k+1}\}$, each characterized by one of the following situations:
(i) the admissible controller parameters belong to (20), $\forall \omega \in J_k$,
(ii) the admissible controller parameters belong to (21), $\forall \omega \in J_k$,
(iii) all parameters pairs (x, z) are admissible, $\forall \omega \in J_k$ (circle $\Phi_{\omega,y,q}$ is internal to disk $\Gamma, \forall q$),
(iv) no (x, z)–pair satisfies (7) $\forall \omega \in J_k$ (circle $\Phi_{\omega,y,q}$ is external to disk Γ, $\forall q$).

Situation (iii) does not impose any restriction and can be disregarded. Instead, if at least one interval J_k exists where situation (iv) holds, then constraint (7) is too tight for the considered value of y. Therefore, we assume that situation (iv) does not occur and limit the analysis to the intervals J_k where either situation (i) or situation (ii). In the first case, the admissible controller parameter region is the intersection of all the stripes (20) corresponding to every $\omega \in J_k$, which is a single convex set as shown in Figure 4a. In the second case, the admissible controller parameter region is the intersection of all the

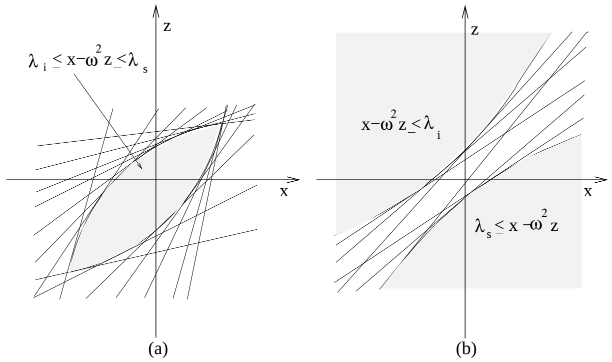


Fig. 4. (a) Intersection of all the stripes (20) corresponding to every $\omega \in J_k$: the admissible region is the shaded convex set. (b) Intersection of all the pairs of half–planes (21) corresponding to every $\omega \in J_k$: the admissible region consists of the pair of shaded disjoint convex sets.

pairs of half–planes (21) corresponding to every $\omega \in J_k$, which consists of two disjoint convex sets as shown in Figure 4b.

According to arguments similar to those used in [3], it can be proved that the set of the admissible parameter pairs (x, z) corresponding to a given value of y is the union of disjoint convex sets, called *convex components*, whose maximal number is bounded (and usually small). Unfortunately, even if the points of a convex component ensure the satisfaction of constraint (7), they may correspond to an unstable system. However, for $y > 0$, if $P(0; q) \neq 0$ and the high–frequency gain of $P(s; q)$ is positive, all of the points inside a convex component in the first quadrant of the (x, z)–plane (the region of practical interest) give rise to the same closed–loop pole distribution. It is therefore enough to check the stability of the feedback system at a unique point of every convex component.

To determine the entire feasible region of the three–dimensional controller parameter space, it is necessary to sweep over all the positive values of y corresponding to stable behaviour. In this regard, the bounds on y provided in [8] and [11] for PID controllers can be exploited to limit the range of the feasible values of y. Graphic techniques can be employed to visualize the feasible three–dimensional region, as shown in Figure 5.

An example concerning the bounded–sensitivity PID control of a *dc*–motor drive is worked out in the next section.

5 PID control with bounded sensitivity

The (uncertain) transfer function from the armature voltage to the shaft angular position of a *dc*–motor with independent excitation can be approximated by

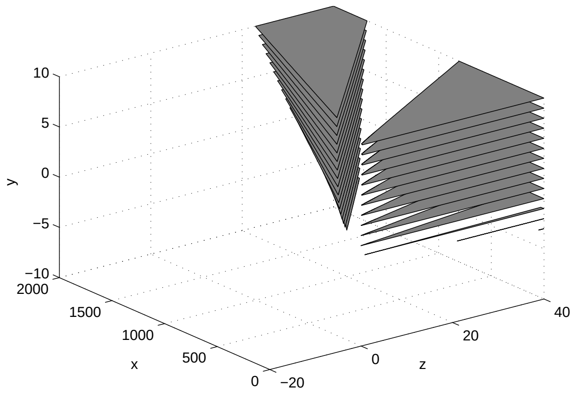


Fig. 5. A sample representation of the feasible region in the three–dimensional controller parameter space.

$$P(s;q) = \frac{K}{s(s+p)} \tag{22}$$

where $q = \{K, p\}$ and

$$K \in [K_{min}, K_{max}] \ \ , \ p \in [p_{min}, p_{max}] \tag{23}$$

with $K_{max} > K_{min} > 0$ and $p_{max} > p_{min} > 0$.

This section deals with the feedback control of the shaft angular position by means of a PID controller with transfer function

$$C(s) = \frac{x + ys + zs^2}{s} \tag{24}$$

($d = 0$ in (1)) in such a way that the H_∞ norm of the sensitivity function $S(s;q) = 1/[1 + C(s)P(s;q)]$ does not exceed an upper bound γ (see (3) with $F(s;q) = S(s;q)$).

Taking (5), (22) and (24) into account, (6) becomes

$$F(\jmath\omega;q) = S(\jmath\omega;q) = \frac{-\omega^2(p + \jmath\omega)}{\lambda - \omega^2(p + \jmath\omega) + \jmath\omega K y}\,. \tag{25}$$

Therefore, $|F(\jmath\omega, \pm\infty, y, q)| = 0$, $\forall q$, which means that circle $\Phi_{\omega,y,q}$ is never external to Γ. It follows that the positive semi–axis Ω_+ can be subdivided into ω–intervals J_k of these two kinds only: (i) intervals where all (x, z)–pairs satisfy (7), and (ii) intervals, called *active intervals*, where the admissible pairs belong to (21). Essentially, the solution of the aforementioned problem entails the determination according to (17) and (18) of $\overline{\lambda}_i(\omega, y)$ and, respectively, $\overline{\lambda}_s(\omega, y)$ for a number of angular frequencies belonging to the active intervals.

The values of $\lambda_i(\omega, y, q)$ and $\lambda_s(\omega, y, q)$ in (17) and (18) correspond to the real roots, if any, of the second–degree polynomial equation obtained, for given ω, y, q, from the equation $|S(\jmath\omega, \lambda, y, q)| = \gamma$ with $S(\jmath\omega, \lambda, y, q)$ as in (25). Their expressions are

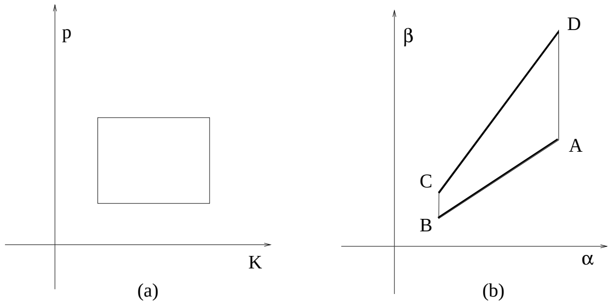
$$\lambda_i(\omega, y, q) = \frac{1}{K}\left\{\omega^2 p - \sqrt{\frac{\omega^4(p^2+\omega^2)}{\gamma^2} - \omega^2(Ky-\omega^2)^2}\right\} \tag{26}$$

$$\lambda_s(\omega, y, q) = \frac{1}{K}\left\{\omega^2 p + \sqrt{\frac{\omega^4(p^2+\omega^2)}{\gamma^2} - \omega^2(Ky-\omega^2)^2}\right\}. \tag{27}$$

To find the minimum $\overline{\lambda}_i(\omega, y)$ of (26) and the maximum $\overline{\lambda}_s(\omega, y)$ of (27) with respect to the plant parameters, it is useful to resort to the transformation

$$\alpha := \frac{1}{K} > 0\,,\ \beta := \frac{p}{K} = p\alpha > 0 \tag{28}$$

by which the rectangular uncertainty set of Figure 6a is transformed into the trapezoidal set of Figure 6b.



(a) (b)

Fig. 6. (a) Rectangular set of allowable values of the original parameter vector $q := (K, p)$. (b) Trapezoidal set of allowable values of the transformed parameter vector $q_t := (\alpha, \beta)$.

In this way, (26) can be rewritten as

$$\lambda_i(\omega, y, q) = \hat{\lambda}_i(\omega, y, q_t) := \omega^2\beta - \sqrt{\frac{\omega^4}{\gamma^2}(\beta^2+\omega^2\alpha^2) - \omega^2(y-\omega^2\alpha^2)^2} \tag{29}$$

where $q_t := (\alpha, \beta)$. Since (29) decreases with β, its minimum occurs when β takes the smallest value compatible with α, that is, $\beta = p_{min}\alpha$ which lies on the AB side of the trapezium represented in Figure 6b. Therefore, $\overline{\lambda}_i(\omega, y)$ can be found by minimizing

$$\omega^2 p_{min}\alpha - \sqrt{\omega^6\alpha^4 + \left[\frac{\omega^4}{\gamma^2}(p_{min}^2+\omega^2) + 2\omega^4 y\right]\alpha^2 - \omega^2 y^2} \tag{30}$$

with respect to α only.

Similarly, the maximum $\overline{\lambda}_s(\omega, y)$ of (27) occurs for $\beta = p_{max}\alpha$ which lies on the CD side of the trapezium represented in Figure 6b, and can be found by maximizing

$$\omega^2 p_{max}\alpha + \sqrt{\omega^6\alpha^4 + \left[\frac{\omega^4}{\gamma^2}(p_{max}^2 + \omega^2) + 2\omega^4 y\right]\alpha^2 - \omega^2 y^2} \qquad (31)$$

with respect to α. Note that (31) increases with α so that its maximum occurs at D.

6 Conclusions

By pursuing the analysis in [3], a characterization of the entire parameter region for the controller (1) ensuring the satisfaction of the H_∞ bound (3) for all possible values of the uncertain plant parameters, has been provided. In particular, it turns out that the controller parameter region is formed by the union of a finite number of disjoint convex sets. The design task (Problem 1 in Section 2) can be accomplished by repeatedly solving the optimization problems (13)–(14) or (17)–(18). The method has been applied in Section 5 to the bounded–sensitivity control of a *dc*–motor drive with an uncertain gain and pole.

References

1. Lepschy A, Viaro U (1985) *J Franklin Inst*, 319:559–567
2. Ferrante A, Krajewski W, Lepschy A, Viaro U (2002) *IEEE Trans Automat Contr*, 47:2117–2121
3. Blanchini F, Lepschy A, Miani S, Viaro U (2004) *IEEE Trans Automat Contr* 49:736–740
4. Krajewski W, Lepschy A, Viaro U (2004) *IEEE Trans Contr Sys Technol* 12:973–983
5. Blanchini F, Lepschy A, Miani S, Viaro U (2005) in *Kuljanic E, Ed, Advanced manufacturing systems and technology. Springer, Wien* 482:297–305
6. Krajewski W, Lepschy A, Miani S, Viaro U (2005) *J Franklin Inst* , 342:674–687
7. Doyle JC, Francis BA, Tannenbaum AR (1992) *Feedback control theory.* MacMillan, New York
8. Ho MT (2003) *Automatica*, 39:1069–1075
9. Bianchini G, Falugi P, Tesi A, Vicino A (2007) *IEEE Trans Automat Contr*, 52:514–520
10. Datta A, Ho MT, Bhattacharyya SP (2000) *Structure and Synthesis of PID Controllers.* Springer, London
11. Söylemez MT, Munro N, Baki H (2003) *Automatica*, 39:121–126

State and parameter estimation approach to monitoring AGR nuclear core

Claudio Bonivento[1], Michael J. Grimble[2], Leonardo Giovanini[3], Mattia Monari[4], and Andrea Paoli[5]

[1] CASY-DEIS, University of Bologna, Italy
claudio.bonivento@unibo.it
[2] ICC, Strathclyde University of Glasgow, United Kingdom
m.grimble@eee.strath.ac.uk
[3] ICC, Strathclyde University of Glasgow, United Kingdom
leonardo.giovanini@eee.strath.ac.uk
[4] former visiting student at ICC, Strathclyde University of Glasgow, United Kingdom
mattia.monari@aliceposta.it
[5] CASY-DEIS, University of Bologna, Italy andrea.paoli@unibo.it

This work is dedicated to the memory of professor and friend Antonio (Toni) Lepschy, University of Padua, who has played a key role in the Italian control community from dawn to our days with his deep culture, scientific influence and great humanity.

1 Introduction

This work concerns with the problem of monitoring an Advanced Gas-cooled Nuclear Reactor (AGR) core. This plant (figure 1) makes use of the heat given by the nuclear efficient reaction to produce electricity by means of steam turbines. These are driven by steam, which is heated, from the AGR gas using a heat exchanger. One of the advantages of a gas cooled reactor is the high temperature that the gas can achieve so that when it is used in conjunction with the heat exchanger and steamed turbine the thermal efficiency is very high.

In the United Kingdom the advanced gas-cooled reactor (AGR) nuclear power stations are approaching the end of their predicted operational live. The reactor core is composed of a hundreds of hollow graphite bricks (that acts as neutron moderator), and the graphite ages because of neutron irradiation and radiolytic oxidation causing distortion and potentially cracking of the bricks since it is impossible to repair or replace the graphite bricks the graphite core is one of the main components that determinate the operational life of

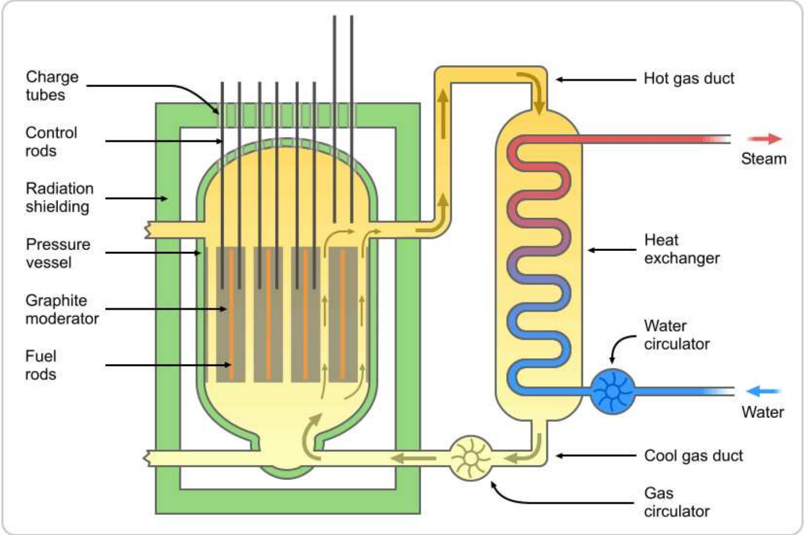


Fig. 1. Schematic diagram of an Advance Gas Cooled Reactor.

a nuclear station. In other terms the major factor that dictates the life of a nuclear power station is the condition of the graphite reactor core, which distort over time with prolonged exposure to heat and radiation.

Currently, it has been proposed to extend the operational lifetime of the nuclear plants if the distortions of the reactor cores are not as severe as initially predicted, and if it is possible to prove that the reactors are still safe to operate. From this, it is clear how important is to keep under monitoring the integrity of the plant and especially of the core; this is actually made possible by a routine performed during planned station outage. These outages occur roughly every three years and result in a large volume of detailed information collected by a system called *Channel Bore Monitoring Unit* (CBMU). This data consists of accurate measures of the channel bore diameter and tilt angles; this information is used to provide an overall assessment of the health of the core.

To perform a more accurate monitoring of the core over their predicted operational life, the estimation of its state should be more frequent; on the other side it is important that the reactor is not offline frequently or for long periods. On the other hand data is also gathered during core refueling operations. Nuclear fission is used to generate heat in order to produce steam to generate electricity from a turbine and, in order to sustain a constant power output, the uranium dioxide fuel needs to be periodically replaced. This process, called reactor refueling, take place with a weekly rate. An important source of information during the refueling phase is the fuel grab load trace data, that consists in collecting information on the position of the uranium bar inserted in the core and information on the force produced by the inter-

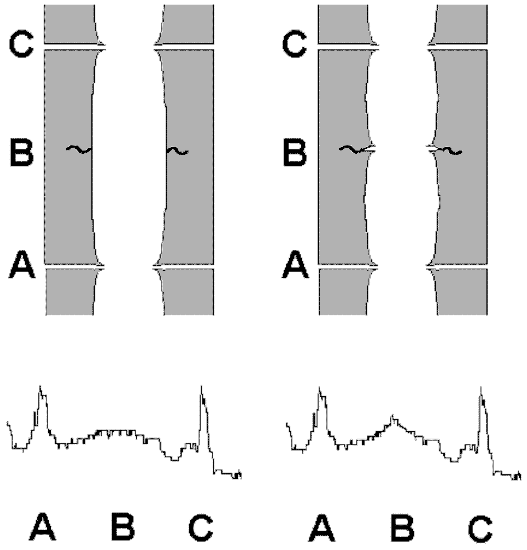


Fig. 2. The upper two diagram show a cut-away of a graphite brick with points A and C illustrating the interface between two bricks. The lower two diagrams show the load force applied on brushes.

action between the wall of the fuel channel and the fuel assembly supporting brushes. Although not originally intended for core condition monitoring purposes, the fuel grab load trace data contains a contribution from frictional interface between the fueling channel wall and the fuel assembly. Since interfaces between adjacent brick layers result in changes in the bore diameter of the channel, as the brushes supporting the fuel rods pass through these features, there is an equivalent change in the friction forces between the walls and the brushes, which correspond to an apparent change of the load force on the fuel assembly. This change in load manifest itself as peaks within the refueling load trace. Figure 2 shows traces of data recorded during a refueling operation. The reader can observe how peaks in the applied force correspond to brick layer interfaces (points A and C); moreover damages over the graphite bricks (e.g. point B) reflect on smaller peaks in the force applied on brushes (see also [1], [2] and [3]).

In [4], CBMU data was compared with load trace data coming from different refueling event and it has been shown how a load trace and a CMBU trace can furnish the same information: as depicted in figure 3 it is possible to recover from the grab force data information on the friction force applied between the supporting brushes and the fuel assembly, which can be used to monitor possible cracks of graphite bricks and hence the health of the nuclear core.

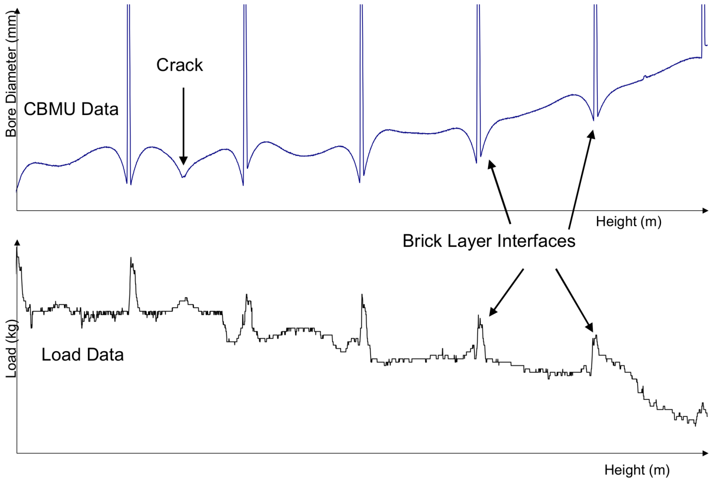


Fig. 3. Comparison of a channel bore monitoring unit trace of channel bore diameter and a refuelling load trace from the same channel.

The purpose of this work is to present a monitoring system based on analytical redundancy and directional residual generation using measurements obtained during the refueling process. In short this problem consists of building an unknown input observer with the role to estimate the friction force produced by the interaction between the wall of the fuel channel and the fuel assembly supporting brushes. This let to estimate the shape of the graphite bricks that comprise the core and, therefore, monitor any distortion of them.

The theoretical machinery exploited in this work is the Kalman filter theory (e.g. [5],[6]), which is used to estimate the information above mentioned. In a different nuclear context, in particular in safeguards problems, a similar approach has been used in [7]. In this paper we will discuss the model of the system used for estimation purposes and the application of a discrete-time Kalman filter to estimate the friction force from the fuel grab load signal stored during the refueling process. Since the initial condition of the system are not known, and considering the fact that the estimation process is performed off-line, a smoothing algorithm based on Kalman filter is introduced to improve the estimate. This is important as a matter of fact that, even if the grab load data is a time signal, as shown in figure 3, it should be considered as parametrized in the height dimension of the fueling channel wall. Hence a perfect estimate both for $t = 0$ and $t = N$ is necessary.

Moreover it will be presented how to deal with the quantization of the filtered data that introduce a noise in data streams (see e.g. [8], [9]). Finally some experimental results will be presented.
More details about this approach, can be found in [10].

2 Refueling model

Each refueling phase provides two data traces, one obtained by lowering the fueling assembly into the nuclear core, and the other one by raising the fuel assembly out from there. The fuel grab load trace data is obtained during the refueling by load cells positioned on the refueling machine which directly measure the force applied by the fuel assembly. This force depends on several factors, among which the most significant are in the following described

a) *The weight of the fuel assembly*: this term depends on the fuel rod mass which changes due to the nuclear reactions in the core. During the extraction process it can be determined once the fuel assembly is out of the reactor.
b) *The frictional forces*: is the quantity that we want to estimate, is caused by the interaction with the stabilizing brushes on the fuel channel wall. These brushes are set directly on the wall and the magnitude of the frictional component depends on the shape of the wall: any distortion in the channel geometry will reflects in friction force changes.
c) *The buoyancy force*: is caused by the gas that, circulating in the fuel chamber, makes the fuel assembly appear lighter. This force is unknown and changes its effect on the fuel assembly with the position of the uranium bar into the channel. But keeping under consideration only a small part of the fuel channel, as a brick, the effect of the buoyancy force can be taken into account as an addictive noise.

During refueling process, the fuel assembly is governed by the interaction of forces that simultaneously act on the fuel assembly:

$$m\mathbf{a} = \sum \mathbf{F} \; ; \tag{1}$$

where m is the fuel assembly mass and $\mathbf{a}$ is its acceleration.

The forces acting on the fuel assembly are, together with the grab load force F_l applied by the supporting brushes on the assembly, its weight mg (where g is the gravitational acceleration), the brushes friction forces F_f and the aerodynamic force F_a due to the gas flow in the fuel chamber:

$$m\mathbf{a} = \mathbf{F_l} + m\mathbf{g} + \mathbf{F_a} + \mathbf{F_f} \; . \tag{2}$$

Rewriting equation (2) expliciting the velocity and the acceleration of the assembly, we obtain:

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1/m \\ 0 \end{bmatrix} F_f + \begin{bmatrix} -F_l/m - g \\ 0 \end{bmatrix} + w \\ y &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + v \; , \end{aligned} \tag{3}$$

where x_1 is the position of the fuel assembly, x_2 is its speed, w is system noise and v is measurements noise. The sign of the friction force is positive because we consider just the reactor discharge, where the friction opposes the movement of the assembly, and therefore narrowing of the channel will result in an increase in a apparent load of the fuel assembly.

In order to rewrite equation (3) in a machine-computable form, we consider its discrete-time approximation, calculating the derivatives of the position and velocity as:

$$\begin{aligned} \dot{x}_1 &= \frac{x_1(t+1) - x_1(t)}{\Delta t} \\ \dot{x}_2 &= \frac{x_2(t+1) - x_2(t)}{\Delta t} \ , \end{aligned} \tag{4}$$

where t is discrete time and Δt is the sample period. This approximation leads to the following discrete time model of the system:

$$\begin{aligned} \begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ \Delta t & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \Delta t \begin{bmatrix} -1/m \\ 0 \end{bmatrix} F_f(t) + \Delta t \begin{bmatrix} -F_l/m - g \\ 0 \end{bmatrix} + w(t) \\ y(t) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + v(t) \ . \end{aligned} \tag{5}$$

In the following sections it will be to presented the use of a discrete-time Kalman filter on system (5) to estimate of the amplitude of friction force F_f.

3 Using Kalman filter and smoother to estimate the core condition

Aim of this section is to present an estimation procedure that, starting from model (5) and having available the set of measures described in Section 1 (i.e. the position of the fuel assembly along the channel and the grab force applied on it), is able to estimate the friction force F_f. In order to estimate the friction force F_f applied along the fueling channel wall, we will first consider it as an unknown input for system (5) and, using an adapted version of Kalman filter for systems with unknown inputs (see [11]), we will estimate the system state. Having the state estimation it is possible to evaluate the friction force term F_f using the first equation in (5). In order to improve the estimation for small time instants (i.e. for the initial position of the fuel assembly), having a first estimation of the unknown input F_f, it is possible to use a Kalman smoother (see Appendix A) to process the system in the reverse way, find an estimation of the state and, consequently, of the friction force at time $t = N; N-1; \ldots 1; 0$. Finally, in order to find an optimal estimation of the system state, and hence an optimal estimation of the friction force, the system will be processed using a forward known input Kalman filter. Roughly speaking running the Kalman filter forward in time we estimate the state of the system, while running it backward in time we make a correction of the previous estimate of the friction force thanks to additional information of the system gathered during the first forward estimation.

Recalling system (5), our aim is to write it in the form

$$\begin{aligned} \begin{bmatrix} z_d(t+1) \\ z_f(t+1) \end{bmatrix} &= \begin{bmatrix} F_1 & F_2 \\ F_3 & F_4 \end{bmatrix} \begin{bmatrix} z_d(t) \\ z_f(t) \end{bmatrix} + \begin{bmatrix} \bar{D} \\ 0 \end{bmatrix} d(t) + \begin{bmatrix} \bar{G}_1 \\ \bar{G}_2 \end{bmatrix} + w \\ y(t) &= \begin{bmatrix} \bar{C}_1 & \bar{C}_2 \end{bmatrix} \begin{bmatrix} z_d(t) \\ z_f(t) \end{bmatrix} + v(t) \ . \end{aligned} \tag{6}$$

where the disturbance $d(t)$ acts as the friction force F_f. Defining the non-singular real matrix U

$$U = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \ , \tag{7}$$

it is possible to find the relation between system (5) and (6):

$$\begin{bmatrix} F_1 & F_2 \\ F_3 & F_4 \end{bmatrix} = U \begin{bmatrix} 1 & 0 \\ \Delta t & 1 \end{bmatrix} U^{-1} := \bar{A} \tag{8}$$

$$\begin{bmatrix} \bar{D} \\ 0 \end{bmatrix} = U \begin{bmatrix} -\Delta t/m \\ 0 \end{bmatrix} := \bar{B} \tag{9}$$

$$\begin{bmatrix} \bar{C}_1 & \bar{C}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} U^{-1} := \bar{C} \tag{10}$$

$$\begin{bmatrix} z_d(t) \\ z_f(t) \end{bmatrix} = Ux(t) := \begin{bmatrix} \bar{x}_1(t) \\ \bar{x}_2(t) \end{bmatrix} \ . \tag{11}$$

Note that the term

$$\Delta t \begin{bmatrix} -F_l/m - g \\ 0 \end{bmatrix} \tag{12}$$

in first equation of system (5) is not present in the correspondent equation of system (6); this matrix, referred as E, will be consider as a known input of the system (5). Following this reasoning, the system can be rewritten in the form

$$\begin{aligned} \begin{bmatrix} \bar{x}_1(t+1) \\ \bar{x}_2(t+1) \end{bmatrix} &= \bar{A} \begin{bmatrix} \bar{x}_1(t+1) \\ \bar{x}_2(t+1) \end{bmatrix} + \bar{B}F_f(t) + \bar{E} + w \\ y(t) &= \bar{C} \begin{bmatrix} \bar{x}_1(t+1) \\ \bar{x}_2(t+1) \end{bmatrix} + v(t) \ . \end{aligned} \tag{13}$$

where

$$\bar{E} = UE = \Delta t \begin{bmatrix} -F_l - g \\ 0 \end{bmatrix} \tag{14}$$

In order to estimate the system state in presence of an unknown input, its effect on the system must be isolated; to this aim it is possible to define a non-singular real matrix V, such that

$$
\begin{aligned}
Vy(t) &:= \begin{bmatrix} \bar{y}_1(t) \\ \bar{y}_2(t) \end{bmatrix} = \bar{y}(t) \\
V\bar{C} &= \begin{bmatrix} \bar{C}_{11} & \bar{C}_{12} \\ 0 & \bar{C}_{22} \end{bmatrix} \\
Vv(t) &= \begin{bmatrix} \bar{v}_1(t) \\ \bar{v}_2(t) \end{bmatrix}
\end{aligned}
\tag{15}
$$

in this way we have transformed the second equation of (13) in the following form:

$$
\begin{aligned}
\bar{y}_1(t) &= \bar{C}_{11}\bar{x}_1(t) + \bar{C}_{12}\bar{x}_2(t) + \bar{v}_1(t) \\
\bar{y}_2(t) &= \bar{C}_{22}\bar{x}_2(t) + \bar{v}_2(t)\ ;
\end{aligned}
\tag{16}
$$

where $\bar{C}_{11}$ is a matrix with rank l in order to preserve system observability.

Now it is possible to rewrite the first equation in (16) as

$$
\bar{x}_1(t) = \bar{C}_{11}^{-1}\left[\bar{y}_1(t) - \bar{C}_{12}\bar{x}_2(t) - \bar{v}_1(t)\right] \tag{17}
$$

and substituting this into the first equation of (13) it is possible to find that

$$
\begin{aligned}
\bar{x}_2(t+1) &= \tilde{A}\bar{x}_2(t) + \tilde{B}\bar{y}_1(t) + \bar{E}_2 + \tilde{G}\tilde{w}(t) \\
\bar{y}_2(t) &= \bar{C}_{22}\bar{x}_2(t) + \bar{v}_2(t)\ ,
\end{aligned}
\tag{18}
$$

where

$$
\begin{aligned}
\tilde{A} &= \left[\bar{A}_{22} - \bar{A}_{21}\bar{C}_{11}^{-1}\bar{C}_{12}\right] \\
\tilde{B} &= \bar{A}_{21}\bar{C}_{11}^{-1} \\
\tilde{G} &= \left[\bar{G}_2 - \bar{A}_{21}\bar{C}_{11}^{-1}\right] \\
\tilde{w}(t) &= \left[\, w(t)\ \bar{v}_1(t) \,\right]^{\mathrm{T}}\ .
\end{aligned}
\tag{19}
$$

Now system (18) is exactly the same as (89) in Appendix A, hence we can find the estimate of the state applying the known input Kalman filter according to the following procedure.

First iteration: unknown input Kalman Filter

State estimation a priori:

$$
\hat{x}_2(t+1) = \tilde{A}\hat{x}_2(t\,|\,t) + \tilde{B}\hat{y}_1(t) \tag{20}
$$

Error covariance a priori:

$$
P_2(t+1) = \tilde{A}P_2(t\,|\,t)\tilde{A}^{\mathrm{T}} + Q_2 \tag{21}
$$

Kalman gain matrix:

$$
K(t+1) = P_2(t+1)\bar{C}_{22}^{\mathrm{T}}\left[\bar{C}_{22}P_2(t+1)\bar{C}_{22}^{\mathrm{T}} + R_2\right]^{-1} \tag{22}
$$

State estimation a posteriori:

$$\hat{x}_2(t+1\,|\,t+1) = \hat{x}_2(t+1) + K(t+1)\left[\bar{y}_2(t+1) - \bar{C}_{22}\hat{x}_2(t+1)\right] \tag{23}$$

Error covariance a posteriori:

$$\begin{aligned} P_2(t+1\,|\,t+1) &= P_2(t+1) - P_2(t+1)\bar{C}_{22}^{\mathrm{T}}\left[\bar{C}_{22}P_2(t+1)\bar{C}_{22}^{\mathrm{T}} + R_2\right]^{-1} \\ &\quad \bar{C}_{22}P_2(t+1) \\ &= P_2(t+1) - K(t+1)\bar{C}_{2}2P_2(t+1) \end{aligned} \tag{24}$$

Initial conditions

$$\hat{x}_2(0) = 0 \qquad P_2(0) = 1e7\,. \tag{25}$$

Having the estimate $\hat{x}_2(t)$ it is possible to compute $\hat{x}_1(t)$ from equation (17) as

$$\hat{x}_1(t) = \bar{C}_{11}^{-1}\left[\bar{y}_1(t) - \bar{C}_{12}\hat{x}_2(t\,|\,t)\right] \tag{26}$$

with conditional covariance

$$P_1(t) = \bar{C}_{11}^{-1}\bar{C}_{12}P_2(t\,|\,t)\bar{C}_{12}^{\mathrm{T}}\bar{C}_{12}^{\mathrm{T}^{-1}} + \bar{C}_{11}^{-1}R_1(t)\bar{C}_{11}^{-1^{\mathrm{T}}}\,, \tag{27}$$

where $R_1(t)$ is the covariance matrix of the noise term $v_1(t)$.

From the estimates $\hat{x}_1(t)$ and $\hat{x}_2(t)$ it is possible to compute the friction force $F_f(t)$ using the first equation of (13):

$$\hat{F}_f(t) = \bar{B}^{-1}\left[\hat{x}_1(t+1) - \bar{A}_{11}\hat{x}_1(t) - \bar{A}_{12}\hat{x}_1(t\,|\,t) - \bar{E}_1\right]\,. \tag{28}$$

Moreover the estimate state $x(t)$ of system (5) and its error covariant matrix can be computed as:

$$\begin{aligned} \hat{x}(t) &= U^{-1}\begin{bmatrix} \hat{x}_1(t) \\ \hat{x}_2(t) \end{bmatrix} \\ P(t) &= U^{-1}\begin{bmatrix} P_1(t\,|\,t) & L(t) \\ L^{\mathrm{T}}(t) & P_2(t\,|\,t) \end{bmatrix} \end{aligned} \tag{29}$$

where

$$L(t) = -\bar{C}_{11}^{-1}\bar{C}_{12}P_2(t\,|\,t)\,. \tag{30}$$

The Kalman filter based algorithm just presented is able to estimate the state of the system even if a disturbance (represented in our case by the friction force determined by the brushes) is acting on it. From this estimate it is possible to compute the magnitude of the friction force F_f simply using (13). It is important to note that the statistic property of the friction force at the instant time $t = 0$ are not known, and therefore the state estimation in $t = 0$ is not appropriate. Remember that this fact reflects in a wrong estimation of the friction force applied on the fuel assembly around its initial position.

The idea to deal with this problem is to use the estimation of the friction force to improve the state estimates just by gathering information in the reverse way. Thus, applying the backward Kalman filter on system (5), here rewritten as

$$\begin{aligned} \bar{x}(t+1) &= A\bar{x}(t) + BF_f(t) + E(t) + w(t) \\ y(t) &= C\bar{x}(t) + v(t) \ ; \end{aligned} \tag{31}$$

it is possible to obtain the optimal estimate of the friction force at time $t = 0$. The backward Markovian model considering now the friction force as a known input is

$$\begin{aligned} \bar{x}_b(t) &= A^{-1}\bar{x}_b(t+1) - A^{-1}B\hat{F}_f(t) - \bar{A}^{-1}E(t) + w(t) \\ y(t) &= C\bar{x}_b(t) + v(t) \ ; \end{aligned} \tag{32}$$

applying the Kalman smoothing algorithm to system (32) it is possible to estimate its state from $t = N$ up to $t = 0$ using the following procedure.

Second iteration: known input Kalman Smoother

State estimation a priori:

$$\hat{x}_b(t-1\,|\,t) = A^{-1}\hat{x}_b(t\,|\,t) - A^{-1}BF_f(t-1) - A^{-1}E(t) \tag{33}$$

Error covariance a priori:

$$P_b(t-1) = A^{-1}P_b(t\,|\,t){A^{-1}}^{\mathrm{T}} + A^{-1}Q(t){A^{-1}}^{\mathrm{T}} \tag{34}$$

Kalman gain matrix:

$$K_b(t-1) = P_b(t-1)C^{\mathrm{T}}\left[CP_b(t-1)C^{\mathrm{T}} + R(t-1)\right]^{-1} \tag{35}$$

State estimation a posteriori:

$$\hat{x}_b(t-1\,|\,t-1) = \hat{x}_b(t-1) + K_b(t-1)\left[y(t-1) - C\hat{x}_b(t-1)\right] \tag{36}$$

Error covariance a posteriori:

$$P_b(t-1\,|\,t-1) = P_b(t-1) - K_b(t-1)CP_b(t-1) \tag{37}$$

Initial conditions:

$$\hat{x}_b(N) = \hat{x}(N\,|\,N) \qquad P_b(N) = P(N\,|\,N) \tag{38}$$

Applying this algorithm, a new state estimate for $t = N$ through $t = 0$ has been computed, and, consequently, the estimate of the friction force from time $t = N$ up to $t = 0$ has been obtained using the second equation of (31).

Running forward in time and backward in time the Kalman filter algorithm, we have obtained an estimate of the static property of the disturbance that acts on the system, which was not known; with this additional information, it is possible to estimate the state of the system in a proper way using a standard forward in time Kalman filter for systems with known inputs.

Third iteration: known input Kalman Filter

State estimation a priori:

$$\hat{x}(t+1) = A\hat{x}(t\,|\,t) + BF_f(t+1) + E(t) \tag{39}$$

Error covariance a priori:

$$P(t+1) = AP(t\,|\,t)A^{\mathrm{T}} + Q(t) \tag{40}$$

Kalman gain matrix:

$$K(t+1) = P(t+1)C^{\mathrm{T}}\left[CP(t+1)C^{\mathrm{T}} + R(t+1)\right]^{-1} \tag{41}$$

State estimation a posteriori:

$$\hat{x}(t+1\,|\,t+1) = \hat{x}(t+1) + K(t+1)\left[y(t+1) - C\hat{x}(t+1)\right] \tag{42}$$

Error covariance a posteriori:

$$P(t+1\,|\,t+1) = P(t+1) - K(t+1)CP(t+1) \tag{43}$$

Initial conditions:

$$\hat{x}(0) = \hat{x}_b(0) \qquad P(0) = P_b(0) \tag{44}$$

Finally the estimation of the friction force can be computed as

$$\hat{F}_f(t) = B^{-1}\left[\hat{x}(t+1) - A\hat{x}(t) - E(t)\right] \ . \tag{45}$$

4 Simulation results of the proposed estimation scheme

The three steps algorithm just explained has been applied to real data stored during refueling operations. In figure 4 and 5 measurements of the grab force F_l and of the fuel assembly position x_2 gathered during the refueling process are shown.

In figure 6 and figure 7 is depicted the estimation of the friction force after the first step of the algorithm, i.e. after having applied the unknown input Kalman filter. It is possible to observe that the estimated friction force has the same trend of the grab force, but its shape is not exactly the same. This

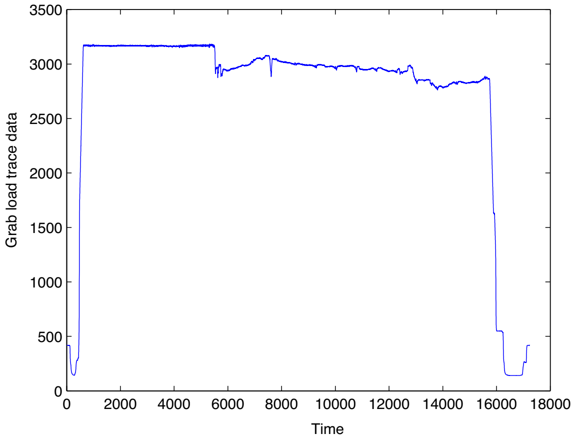


Fig. 4. Grab load trace data gathered during refueling.

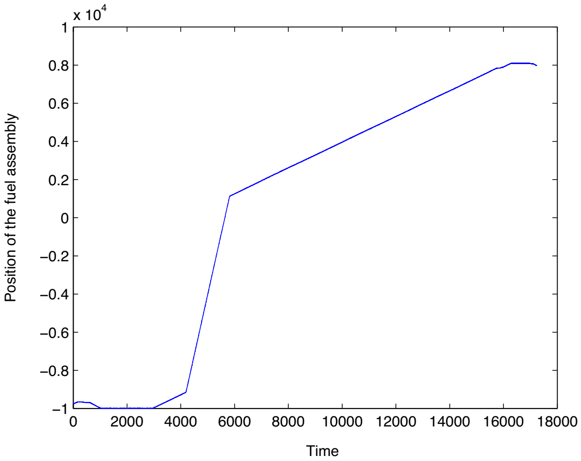


Fig. 5. Position of the fuel assembly during refueling.

is due to the fact that statistic property of the disturbance F_f are not known for $t = 0$.

In figure 8 and figure 9 is presented the estimation of the friction force after having applied the Kalman smoothing algorithm. The result obtained is absolutely better (figure 7).

Remembering that the smoother algorithm has the role to propagate the estimation of the friction force from time $t = N$, to time $t = 0$, a better result can be obtained by processing the system once more by a forward Kalman filter, where now the statistic property of the friction force for $t = 0$ are known, because they are given by the combined use of the first forward Kalman filter

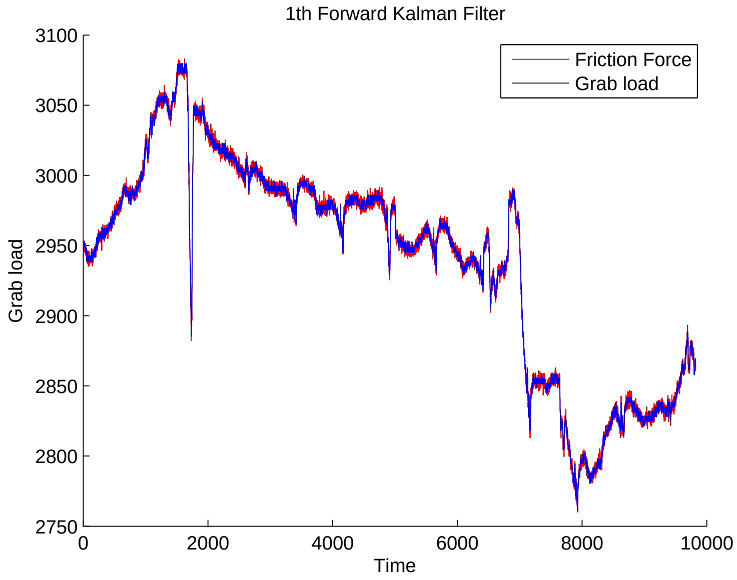


Fig. 6. First iteration estimation.

and the smoother. The results of this third step are shown in figure 10 and figure 11. In this case the trend and the shape of the estimate friction force are exactly the same as the grab load, and this demonstrates that it is possible to obtain an optimal estimation of the acting disturbance without an a priori knowledge on it.

In figure 11 it is possible to observe that still some small errors in the estimate are present; these imperfections are due to the approximate model of the system used to estimate the friction force. For example the model does not consider the noise introduced by the quantization of data, moreover both the mass m of the fuel assembly and the value of the buoyancy force F_a are approximated and considered constant.

In the following section a procedure to deal with the noise introduced by the quantization of data is presented and final simulation results are discussed.

5 Dealing with quantization

Aim of this section is to give some guidelines on how to face the problem of state estimation using quantized measurements; this is necessary since grab load data are quantized and the quantization introduces a noise that affects the estimate. In the following some necessary condition for the maximum likelihood estimate (see [12]) when the observations have been quantized will be given and a Quantization Regression (QR) algorithm (still based on Kalman

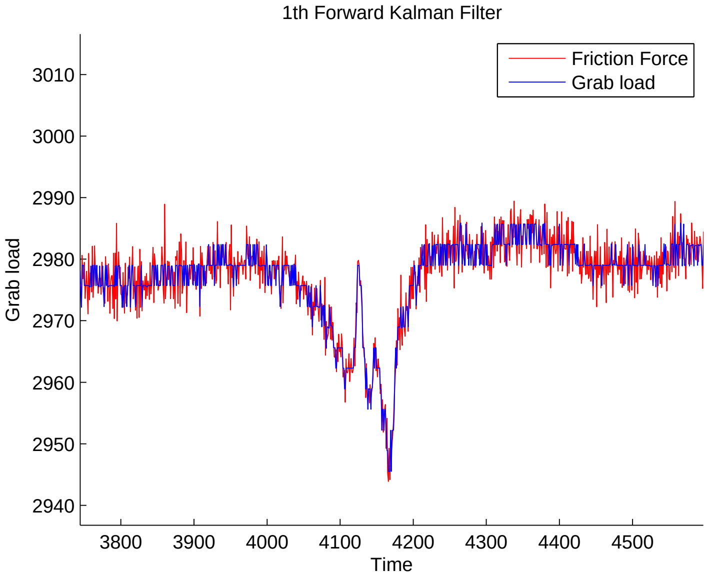


Fig. 7. First iteration estimation (zoom in).

filter) which generates an estimate of an autoregressive time series from quantized measurements will be described.

As reported in [13], the effect of a uniform quantization can be modeled as an additive noise that is uniformly distributed, uncorelated with the input signal, and has white spectrum. Consider the following model with quantized measurements:

$$\begin{aligned} x_{t+1} &= f(x_t, w_t) \\ z_t &= h(x_t) + e_t \\ y_t &= Q_m(z_t) \end{aligned} \tag{46}$$

where $Q_m(\cdot)$ is the quantization function. The problem of optimally estimate the state of (46) is a problem of nonlinear non-Gaussian filtering; as explained in [5] such a problem has a Bayesian solution given by

$$p(x_{t+1} \mid Y_t) = \int_{R^n} p(x_{t+1}|x_t) dx_t$$
$$p(x_t \mid Y_t) = \frac{p(y_t|x_t)p(x_t|Y_{t-1})}{p(y_t|Y_{t-1})} \; .$$

This problem is in general not analytically solvable, but there exists two different approach to deal with it:

a) use an *extended Kalman filter* (EKF) that is a sub-optimal filter for an approximate linear Gaussian model designed using the assumption that the quantization introduce an additive uniform noise;

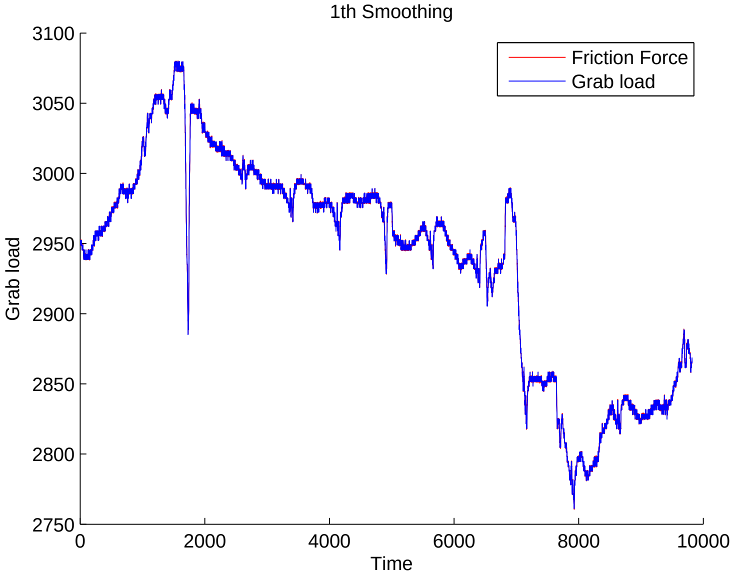


Fig. 8. Second iteration estimation.

b) use a numerical approach to find a miximum-likelihood estimates of parameters, approximating in this way the optimal solution to the Bayesian filtering problem.

Regarding the first approach, it can be easily introduced considering the following linear Gaussian model with quantized observations:

$$\begin{aligned} x_{t+1} &= F_t x_t + G_t w_t \qquad && \mathrm{Cov}(w_t) = Q_t \\ z_t &= H_t x_t + e_t && \mathrm{Var}(e_t) = \sigma^2 \\ y_t &= Q_m(z_t) && \end{aligned} \tag{47}$$

where y_t represents the quantized measurements. Using the assumption that the quantization introduce an additive uniform noise, the optimal filter is given by Kalman filter by increasing the measurement covariance R_t by term equal to $q^2/12$, i.e.

$$R_t = \left(\sigma_t^2 + \frac{q^2}{12}\right) I \ , \tag{48}$$

where q is the quantization box size and I is a suitably dimensioned identity matrix. From (48) it turns out that the measurements covariance matrix R_t is increased of a quantity that depends on how small the quantization box size is and hence on the variance of the quantization noise (cf. [14]). Using (48) it is therefore possible to tune the filter to obtain the best estimation for the problem.

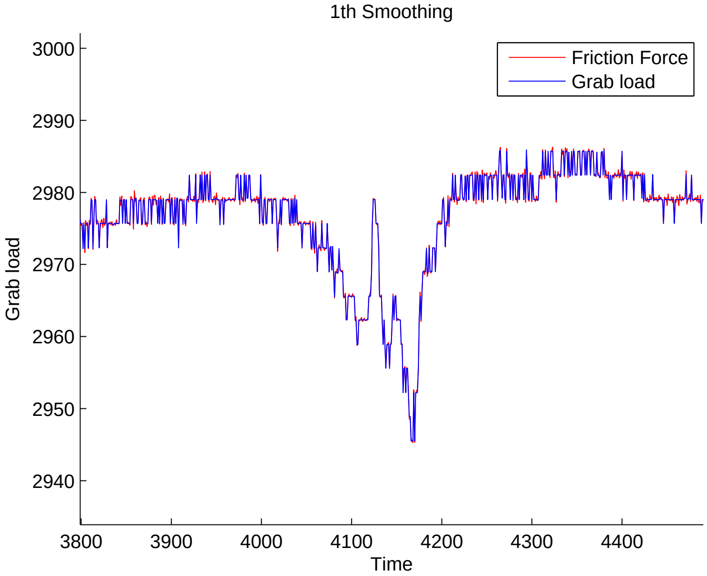


Fig. 9. Second iteration estimation (zoom in).

In the following a slightly different Kalman filter obtained by the Bayesian equation as shown in [15] will be introduced; considering this filter, necessary conditions for the maximum-likelihood estimate of parameters when the observations are quantized will be formulated.

Consider the following linear measurement equation

$$z = \mathrm{H}x + v \tag{49}$$

where x is the vector to be estimated, z is the measurement vector, and v is the measurement noise. Recall that, with non-quantized measurements, the maximum-likelihood estimate (cf [12]) is the value of x that maximizes the likelihood function $L(z;x)$:

$$\hat{x} = \arg[\max_x L(z,x)] = \arg[\max_x p(z:x)] \ , \tag{50}$$

where the notation $p(z:x)$ means the probability-density function of z with x as a parameter of the distribution.

When the measurements are quantized numerical values of z are not available and the knowledge of the measurements is reflected in the inequalities

$$\left\{a^i \leq z^i < b^i\right\} \ , \tag{51}$$

where a^i and b^i are the lower and upper bounds of the quantum interval in which the i-th component of z is known to lie. Considering this fact, the

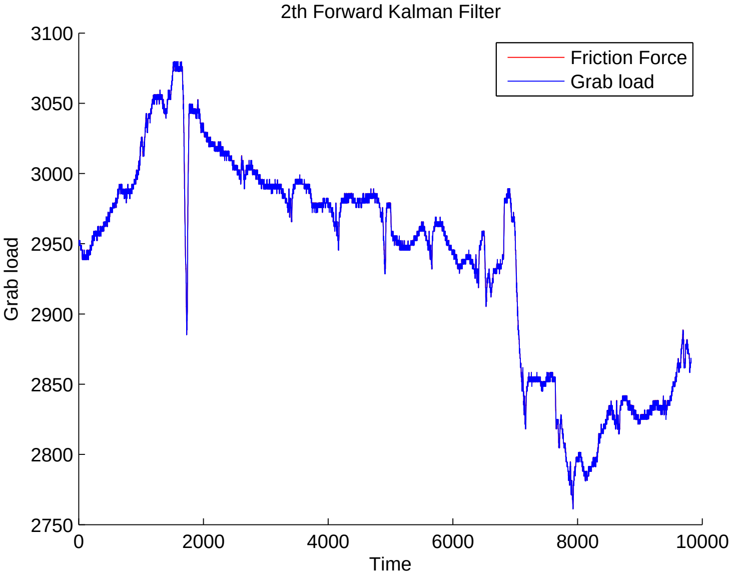


Fig. 10. Third iteration estimation.

likelihood function to be used is the probability that the measurements fall in the hypercube defined by equation (51):

$$L(a^i, b^i, x) = \prod_i P\left[a^i - (Hx)^i \leq v^i < b^i - (Hx)^i\right] \ . \tag{52}$$

Hence the maximum-likelihood estimate of x with quantized measurements is

$$\hat{x} = \arg\left\{\max_x \prod_i P\left[a^i - (Hx)^i \leq v^i < b^i - (Hx)^i\right]\right\}. \tag{53}$$

Denoting with P_i the term $P\left[a^i - (Hx)^i \leq v^i < b^i - (Hx)^i\right]$, such that

$$P_i = \int_{a^i - (\mathrm{Hx})^i}^{b^i - (Hx)^i} p_{v^i}(u)du \ ,$$

the necessary condition for maximum likelihood estimate is the following:

$$\begin{aligned} &\frac{1}{L(a^i, b^i, x)}\left(\frac{\partial L(a^i, b^i, x)}{\partial x}\right) = \sum_i \frac{\partial P_i / \partial x}{P_i} = \\ &= \sum_i \frac{p_{v^i}(b^i - (Hx)^i) - p_v^i(a^i - (Hx)^i)}{P_i} h^i = 0 \ , \end{aligned} \tag{54}$$

where the row vector h^i is the i-th row of H. Hence the problem can be formulated as following. Given

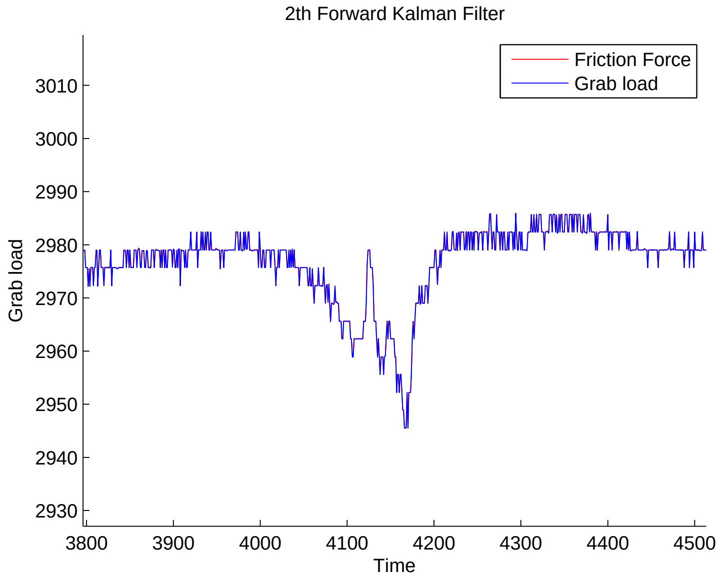


Fig. 11. Third iteration estimation (zoom in).

i) the measurement equation $z = h(x, v)$,

ii) the joint probability density function of parameter and noise vectors $p_{x,v}(\xi, v)$,

iii) the constraint $z \in A$, where A is some hypercube for quantized measurements,

the estimation problem with quantized measurements consists in :

A) finding the conditional mean of $f(x)$ given a measurement z: $E\,[f(x)\,|\,z]$,

B) averaging this function of z considering the constraint $z \in A$.

Assume that the state vector and measurements variables satisfy the relationships

$$\begin{array}{rlrl}
x_{i+1} &= \varPhi_i x_i + w_i & & \\
z_i &= H_i x_i + v_i & & \\
E(x_0) &= \bar{x}_0 & \mathrm{cov}(x_0) &= P_0 \\
E(w_i) &= 0 & E(w_i w_j^T) &= Q_i \delta_{ij} \\
E(v_i) &= 0 & E(v_i v_j^T) &= R_i \delta_{ij} \\
E(w_i v_j^T) &= 0 & E(w_i x_0^T) &= E(v_i x_0^T) = 0
\end{array} \tag{55}$$

where x_i is the system state vector at time t_i, $\varPhi_i$ is the system transition matrix from time t_i to t_{i+1}, w_i is a realization of the process noise at t_i, z_i is the measurement vector at time t_i, H_i is measurements matrix at time t_i and v_i is a realization of the observation noise at time t_i. Each of the m components of the normally distributed vector z has zero mean and lies in a interval whose limits are $\{a^i\}$ and $\{b^i\}$, $a^i \leq z^i < b^i$, $(i = 1, 2, 1 \ldots, m)$. Let (γ^i) $(i = 1, 2, 1 \ldots, m)$ be the m components of the geometric center vector γ of the region A:

$$\gamma^i = \frac{1}{2}(b^i + a^i) \; , \tag{56}$$

and let (α^i) $(i = 1, 2, 1 \ldots, m)$ be the m components of the quantum interval half-widths vector α:

$$\alpha^i = \frac{1}{2}(b^i - a^i) \; . \tag{57}$$

It is possible to show that, expanding the probability-density function in power series in an interval containing γ and neglecting terms higher than the fourth order, the mean and covariance of z conditioned on $z \in A$ are given by

$$E(z \,|\, z \in A) \approx \gamma - \mathit{\Lambda}\mathit{\Gamma}^{-1}\gamma \tag{58}$$

$$\mathrm{cov}(z \,|\, z \in A) \approx \mathit{\Lambda} = \left\{ \frac{(\alpha^i)^2}{3}\delta_{ij} \right\} \tag{59}$$

where $\mathit{\Gamma} = E(zz^T)$ and δ_{ij} is the Kronecker delta. In this case the minimum-variance linear estimate x^* and its covariance E^* are given by

$$x^* = \bar{x} + K^*(\gamma - H\bar{x}) \tag{60}$$

$$E^* = M - MH^T(\mathit{\Gamma} + \mathit{\Lambda})^{-1}HM \tag{61}$$

where

$$K^* = MH^T(\mathit{\Gamma} + \mathit{\Lambda})^{-1} \tag{62}$$

$$\mathit{\Gamma} = \mathrm{cov}(z) = HMH^T + R \; . \tag{63}$$

This problem can therefore be solved recursively with a modified Kalman filter, leading to the following result. Assuming the conditional distribution of the state just before the i-th measurements being $N(\hat{x}_{i\,|\,i+1}, M_i)$, then the Gaussian fit alghorithm for a linear system with quantized system is the following:

$$\hat{x}_{i\,|\,i} = \hat{x}_{i\,|\,i-1} + K_i[E(z_i \,|\, z_i \in A_i) - H_i\hat{x}_{i\,|\,i-1}] \tag{64}$$

$$K_i = M_i H_i^{\mathrm{T}}(H_i M_i H_i^{\mathrm{T}} + R_i)^{-1} \tag{65}$$

$$P_i = M_i - M_i H_i^{\mathrm{T}}(H_i M_i H_i^{\mathrm{T}} + R_i)^{-1} H_i M_i \tag{66}$$

$$E_i = P_i + K_i \mathrm{cov}(z_i \,|\, z_i \in A_i) K_i^{\mathrm{T}} \tag{67}$$

$$\hat{x}_{i+1\,|\,i} = \mathit{\Phi}_i \hat{x}_{i\,|\,i} \tag{68}$$

$$M_{i+1} = \mathit{\Phi}_i E_i \mathit{\Phi}_i^T + Q_i \; , \tag{69}$$

where $\hat{x}_{i\,|\,i}$ is the conditional mean of x_i for quantized measurements up to and including t_i, $\hat{x}_{i\,|\,i-1}$ is the conditional mean of x_i for quantized measurements up to and including t_{i-1}, A_i is the quantum region in which z_i falls, M_i is the conditional covariance of x_i for quantized measurements up to and including t_{i-1}, K_i is the Kalman filter gain matrix at t_i, P_i is the conditional covariance of estimate, and E_i is the conditional covariance of x_i for quantized measurements up to and including t_{i-1}.

Note that equations (64) correctly describe the propagation of the first two moments of the conditional distribution under the assumption of Gaussian noises. Let $e = x - \hat{x}$ be the estimation error, and consider the dynamic

$$e_{i+1\,|\,i} = \Phi_i e_{i\,|\,i} - w_i \ .$$

Since $e_{i\,|\,i}$ is not Gaussian, so $e_{i+1\,|\,i}$ is not Gaussian too, but it tends to a Gaussian distribution because of the addition of Gaussian process noise w_i and of the action performed by the state transition matrix Φ_i. Considering this, equations (64) yield a good approximation of the conditional moments.

Concluding it is important to remark that the recursive algorithm described by (64) is very similar to the algorithm describing the Kalman filter, with two important differences:

- the conditional mean of the measurements vector at t_i is used as an input for the filter;
- the conditional covariance equation is being forced by the random variable $\mathrm{cov}(z_i \,|\, z_i \in A_i)$.

Following the theory just presented a set of simulation with the modified algorithm has been performed and the results are presented in figure 12. It is possible to see that now the estimation algorithm leads to a perfect estimate of the friction force.

6 Concluding remarks

In this work we have presented an estimation algorithm based on Kalman filter to monitor the condition of the core of AGR nuclear stations. In particular, using data stored during the core refueling phase, it is possible to estimate the friction force that the fuel rod apply on the supporting brushes which are embedded in the core wall. In this way it is possible to estimate the shape of the graphite bricks that compose the core and therefore the condition of the core itself.

Future works will consists in gathering existing and historical data in a single location and define patterns in order to determinate whether time, location, operating condition have an effect on the trace.

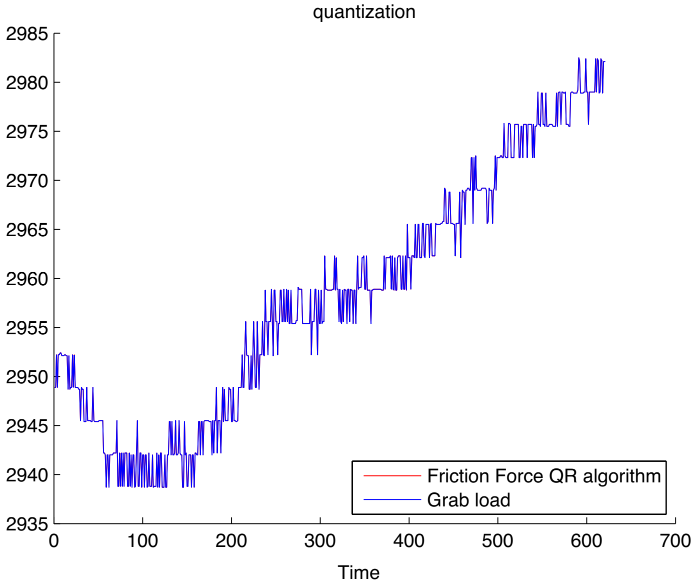


Fig. 12. Estimation of friction force using the modified Kalman filter to deal with quantization (zoom in).

Acknowledgement

Authors wish to thank British Energy Agency and in particular Dr. Jim Reed for making possible a fruitful collaboration and for the permission of using real-plant data.

References

1. Pang Y, Giovanini L, Grimble M J (2006) Condition monitoring of a advanced gas-cooled nuclear reactor core, *Internal report, British Energy*
2. Pang Y, Grimble M J, Ordys A W, Reed J (2007) Condition monitoring of the nuclear graphite core using benchmarking techniques, *Proceedings of the European Control Conference*, Kos, Greece
3. West G M, Jahn G J, McArthur D J, Reed J (2005) Graphite core condition monitoring through intelligent analysis of fuel grab load trace data, *Internal report, British Energy*
4. West G M, McArthur S D J, McDonald J R, Ballantyne A, Reed J, Beynon S (2005) Graphite core brick crack detection through automated load trace analysis, *Internal report, British Energy*
5. Jazwinski A H (1970) Stochastic processes and filtering theory, *Mathematics in science and engineering*, 64
6. Grawal M S, Andrews A P (1993) *Kalman filtering theory and practice*, Prentice Hall, Eanglewood Cliffs, NJ

7. Bonivento C (1983) On line state estimation and indirect measurement in safeguards systems. In: *Argentesi F, Avenhaus R, Franklin M, Shipley J P (eds) Mathematical and statistical methods in nuclear safeguards*, Harwood academic publisher, London, UK
8. Ziskand I, Hertz D (1993) *IEEE Transactions on Signal Processing*, 41:3202–3206
9. Lepschy A, Mian G A, Viaro U (1988) *IEEE Transactions on circuits and systems*, 35:461–466
10. Monari M (2007) Studio di una tecnica di monitoraggio dello stato della camera di combustione di reattori nucleari AGR, Master Thesis, University of Bologna, ITALY
11. Emara-Shabaik H (2003) *Journal of dynamic systems, measurments and control*, 29:482–485
12. Raunch H E, Tung F, Striebel C T (1965) Maximum likelihood estimates of linear dynamical systems, *AIAA Journal*, 3(8)
13. Widrow B, Kollar I, Liu M C (1996)*IEEE Transactions on Instrumentation and Measurments*, 45:482–485
14. Wang M, Thornhill N F, Huangh B (2001) Time series reconstruction from quantized measurments, *Proceedings of the CPCE conference*, Tucson, AZ
15. Curry R (1970) *Estimation and control with quantized measurements,* The MIT Press, Boston, MA

A Kalman filtering, prediction and smoothing

In order to collect symbols and definitions used along the paper, in this Appendix the main formulas of the celebrated Kalman machinery for filtering, prediction and smoothing are briefly reported.

Consider a stochastic system represented by the following model:

$$x_k = \phi_{k-1} x_{k-1} + w_{k-1} \tag{70}$$

$$y_k = H_k x_k + v_k \ . \tag{71}$$

Let $x \in R^n$ (state of the system) and $y \in R^l$ (measurements) be jointly Gaussian random vectors with mean vectors $\mu_x = E\{x\}$ and $\mu_y = E\{y\}$ respectively, and covariance matrixes

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} := \begin{bmatrix} \mathrm{cov}\{x,x\} & \mathrm{cov}\{x,y\} \\ \mathrm{cov}\{y,x\} & \mathrm{cov}\{y,y\} \end{bmatrix} \ . \tag{72}$$

Assume the covariance matrix $\Sigma \in R^{(n+l)\times(n+l)}$ to be positive definite. The measurement and plant noises v_k and w_k are assumed to be zero-mean Gaussian sequences, while the initial value x_0 is considered a Gaussian variate with known mean x_0 and known covariance matrix P_0. According to the previous definitions, the following statements hold:

$$\begin{aligned} E\left\langle w_k \right\rangle &= 0 \\ E\left\langle w_k w_i^T \right\rangle &= \Delta(k-i)Q_k \\ E\left\langle v_k \right\rangle &= 0 \\ E\left\langle v_k v_i^T \right\rangle &= \Delta(k-i)R_k \end{aligned} \tag{73}$$

where $\Delta(k-i)$ stands for the *Kronecker delta function*, and the noise sequences w_k and v_k are assumed to be uncorrelated.

The problem of optimal estimation is to find the minimum variance estimate $\hat{x}(t+m\,|\,t)$ of the state vector $x(t+m)$ based on the observations up to time t of the system (70). This means designing a filter that produce the estimate $\hat{x}(t+m\,|\,t)$ minimizing the performance index

$$J = E\left\{|x(t+m) - \hat{x}(t+m\,|\,t)|^2\right\} \tag{74}$$

We will refer to this problem as *prediction* if $m > 0$, *filtering* if $m = 0$ and *smoothing* if $m < 0$.

Define the estimation error $\tilde{x}(t+m\,|\,t)$ as the difference between the real state value $x(t+m)$ and the estimate $\hat{x}(t+m|t)$:

$$\tilde{x}(t+m\,|\,t) = x(t+m) - \hat{x}(t+m\,|\,t) \tag{75}$$

and let the error covariance matrix be

$$P(t+m\,|\,t) := E\left\{[x(t+m) - \hat{x}(t+m\,|\,t)][x(t+m) - \hat{x}(t+m\,|\,t)]^T\right\}\ . \tag{76}$$

Denoting with $\mathcal{Y}_t$ the linear space generated by the observations, the minimum variance estimation $\hat{x}(t+m\,|\,t)$ is given by the orthogonal projection of $x(t+m)$ onto $\mathcal{Y}_t$

$$\hat{x}(t+m|t) = \hat{E}\left\{x(t+m)\,|\,\mathcal{Y}_t)\right\}\ , \tag{77}$$

i.e. the optimality of $\hat{x}(t+m\,|\,t)$ is obtained when the estimation error $\tilde{x}(t+m\,|\,t)$ is orthogonal to the data space:

$$\tilde{x}(t+m|t) = x(t+m) - \hat{x}(t+m|t) \perp \mathcal{Y}_t\ ; \tag{78}$$

moreover this estimate is unbiased, which means that

$$E\left\{\tilde{x}(t+m|t)\right\} = 0 \qquad \text{for} \quad t = 0, 1, ... \tag{79}$$

Consider now a multivariable Gaussian Markov discrete-time linear system

$$\begin{aligned} x(t+1) &= A(t)x(t) + w(t) \\ y(t) &= C(t)x(t) + v(t) \end{aligned} \tag{80}$$

where $x \in \mathbb{R}^n$ is the state vector, $y \in \mathbb{R}^p$ is the observation vector, $w \in \mathbb{R}^n$ is the plant noise vector, and $v \in \mathbb{R}^p$ is the observation noise vector. Let

$A(t) \in \mathbb{R}^{n\times n}$, $C(t) \in \mathbb{R}^{p\times n}$ be deterministic function of the time t and $w(t)$ and $v(t)$ zero mean Gaussian white noise vectors with covariance matrixes

$$E\left\{\begin{bmatrix} w(t) \\ v(t) \end{bmatrix} \begin{bmatrix} w(t)^T v(t)^T \end{bmatrix}\right\} = \begin{bmatrix} Q(t) & S(t) \\ S^T(t) & R(t) \end{bmatrix} \tag{81}$$

where $Q(t) \in \mathbb{R}^{n\times n}$ is nonnegative defined, and $R(t) \in \mathbb{R}^{pxp}$ is positive defined for all $t = 0, 1, \ldots$. The initial state $x(0)$ is Gaussian with mean $E\{x(0)\} = \mu_x(0)$ and covariance matrix

$$E\left\{[x(0) - \mu_x(0)]\,[x(0) - \mu_x(0)]^T\right\} = \Pi(0)\ ; \tag{82}$$

moreover $x(0)$ is uncorrelated with the noise $w(t)$, $v(t)$, $t = 0, 1, ...$ By using orthogonal projection operators, it is possible to define the following algorithm for the one step ahed Kalman predictor.

State estimation a priori:

$$\hat{x}(t+1) = A(t+1)\hat{x}(t\,|\,t) \tag{83}$$

Error covariance a priori:

$$P(t+1) = A(t)P(t\,|\,t)A^T(t) + Q(t) \tag{84}$$

Kalman gain matrix:

$$K(t+1) = P(t+1)C^T(t+1)\left[C(t+1)P(t+1)C^T(t+1) + R(t+1)\right]^{-1} \tag{85}$$

State estimation a posteriori:

$$\hat{x}(t+1\,|\,t+1) = \hat{x}(t+1) + K(t+1)\left[y(t+1) - C(t+1)\hat{x}(t+1)\right] \tag{86}$$

Error covariance a posterior:

$$P(t+1\,|\,t+1) = P(t+1) - K(t+1)C(t+1)P(t+1) \tag{87}$$

Initial condition:

$$\hat{x}(0) = \mu_x(0) \qquad P(0) = \Pi(0) \tag{88}$$

Consider now a discrete-time stochastic linear system with forcing input

$$x(t+1) = A(t)x(t) + B(t)u(t) + w(t) \tag{89}$$

$$y(t) = C(t)x(t) + v(t) \tag{90}$$

where $u(t) \in \mathbb{R}^m$ is the input vector, and $B(t) \in \mathbb{R}^{n\times m}$ is input distribution matrix. We assume that $u(t)$ is measurable in the sense that $u(t)$ is a function of the outputs.

Exploiting the linearity of the system it is possible to decompose state trajectories in two terms: the free-response $x_w(t)$ and the forced-response $x_u(t)$:

$$x_w(t+1) = A(t)x_w(t) + w(t), \qquad x_w(0) = x(0) \tag{91}$$

$$x_u(t+1) = A(t)x_u(t) + B(t)u(t), \quad x_u(0) = 0 \; ; \tag{92}$$

the solution $x(t+1)$ of (89) is expressed by the superimposition of the effects:

$$x(t+1) = x_w(t+1) + x_u(t+1), \qquad t = 0, 1, \ldots \; .$$

The forced term $x_u(t)$ is known since $u(t)$ is measurable and, defining the state transition matrix $\Phi(t,s)$, it can be computed as

$$x_u(t) = \sum_{k=0}^{t-1} \Phi(t, k+1)B(k)u(k), \qquad t = 0, 1, ... \tag{93}$$

Since $x_u(t)$ is known, the algorithm should compute the estimates of the vector $x_w(t)$ based on the observations and defining the measurements

$$\ell(t) = y(t) - C(t)x_u(t) = C(t)x_w(t) + v(t) \; . \tag{94}$$

Since the system

$$x_w(t+1) = A(t)x_w(t) + w(t) \tag{95}$$

$$\ell(t) = C(t)x_w(t) + v(t) \tag{96}$$

it is completely equivalent to the stochastic system of (80), it is possible to write the Kalman filter algorithm for the stochastic linear dynamic system as previously described, but using the following a priori state estimation equation:

$$\hat{x}(t+1) = A(t+1)\hat{x}(t \,|\, t) + B(t+1)u(t+1) \tag{97}$$

A *smoother* estimates the state of a system at time t using measurements made before and after time t. The accuracy of a smoother is generally better the one obtained by a forward filter, because it use more measurements for its estimate. So the optimum linear smoothing provides an estimate of the past value of the desired quantities. It is possible to represent the problem using a backward Markovian model and therfore to solve the problem using a Kalman filter designed for the backward Markovian model. This filter is called backward Kalman filter and is defined by the following algorithm.

State estimation a priori:

$$\hat{x}_s(t-1 \,|\, t) = A^{-1}(t)\hat{x}_s(t \,|\, t) - A^{-1}D(t-1)u(t-1) \tag{98}$$

Error covariance a priori:

$$P_s(t-1) = A^{-1}(t)P_s(t \,|\, t)A^{-1}(t) + A^{-1}(t)Q(t)A^{-1}(t) \tag{99}$$

Kalman gain matrix:

$$K_s(t-1) = P_s(t-1)C^T(t-1)\left[C(t-1)P_s(t-1)C^T(t-1) + R(t-1)\right]^{-1} \tag{100}$$

State estimation a posterior:

$$\hat{x}_s(t-1\,|\,t-1) = \hat{x}_s(t-1) + K_s(t-1)\,[y(t-1) - C(t-1)\hat{x}_s(t-1)] \tag{101}$$

Error covariance a posterior:

$$P_s(t-1\,|\,t-1) = P_s(t-1) - K_s(t-1)C(t-1)P_s(t-1) \tag{102}$$

Initial condition:

$$\hat{x}_s(N) = \hat{x}(N\,|\,N) \qquad P_s(N) = P(N\,|\,N) \tag{103}$$

Stability Issues in the Disturbance Decoupling Problem for Systems over Rings

Giuseppe Conte and Anna Maria Perdon

Dipartimento di Ingegneria Informatica, Gestionale e dell'Automazione
Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy
gconte@univpm.it, perdon@univpm.it

Summary. The aim of this paper is to review the solvability conditions of the disturbance decoupling problem for systems with coefficients in a ring and to provide new results in case the additional requirement of stability, in a suitable sense, is considered. The problem is approached by making use of geometric methods and tools, which extend those developed for systems with coefficients in a field. It is shown that the results can be interpreted in the framework of time-delay systems, providing new conditions for the solution of the disturbance decoupling problem with stability in that case.

1 Introduction

The problem of decoupling by a suitable feedback the output of a given system from a disturbance input is a classic one in dynamical system and control theory. The solution provided in the framework of linear systems by the geometric approach (see [4], [24]) has been extended to various class of systems, including nonlinear systems, periodic systems, time-delay systems and systems over rings. Here, after recalling a number of known results, we consider the disturbance decoupling problem for system with coefficients in a ring with the additional requirement of stability.
Motivations for our investigation, in addition to the interest in abstract systems over rings, come from the fact that results obtained in the ring framework can be naturally interpreted in the framework of time-delay systems. These have a great importance in several application, where delays due to transportation of materials or to transmission of information cannot be neglected. Here, in Section 2, we introduce systems with coefficients in a ring and, in Section 3, we describe the relation between them and time-delay systems. Then, in Section 4, we state formally the disturbance decoupling problem and we recall the geometric condition for its solution. Computability of those conditions and the practical design of a solution, in the ring framework, are non

trivial issues, but the inherent difficulties can be overcame by characterizing solvability in an alternative way, as recalled in Section 5. From the result of that Section, we derive, in Section 6, a set of new conditions for the solvability of the problem with the additional requirement of stability, in a suitable sense. By interpreting the result in the framework of time-delay systems, one get condition for the solvability of the problem with the additional requirement of stability in the usual sense, which hold in more general situations than those considered in [10].

2 Systems over rings

Let $\mathcal{R}$ denote a commutative ring. By a system with coefficients in $\mathcal{R}$, or a system over $\mathcal{R}$, we mean a linear dynamical system $\Sigma = (A, B, C)$ whose evolution is described by a set of difference equations of the form

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) \\ \quad\; y(t) = Cx(t) \end{cases} \tag{1}$$

where $t \in \mathbb{N}$ is an independents variable, $x(\cdot)$ belongs to the free module $\mathcal{X} = \mathcal{R}^n$, $u(\cdot)$ belongs to the free module $\mathcal{U} = \mathcal{R}^m$, $y(\cdot)$ belongs to the free module $\mathcal{Y} = \mathcal{R}^p$ and A, B, C are matrices of suitable dimensions with entries in $\mathcal{R}$.

By analogy with the classical case of linear, dynamical, discrete-time systems with coefficients in the field of real number, we think of the variables x, u and y as of, respectively, the state, input and output of Σ.

Besides being considered as abstract algebraic objects, systems with coefficient in a ring have been proved to be useful for modeling and studying particular classes of dynamical systems, such as discrete-time systems with integer coefficients, families of parameter dependent systems and time delay systems (see [17], [18] and Section 3).

General results concerning the theory of systems with coefficients in a ring and a number of related control problems can be found in [22],[23], [6], [11] and the references therein. In particular, a geometric theory, similar to the existing one for linear systems over a field (see [4], [24]), has been developed and related concepts, like the notion of controlled invariance and that of conditioned invariance, maintain their relevance in the solution of several design problems in the ring framework (see [11]).

Computation issues in the theory of systems with coefficients in a ring have been considered in [20] and [21], providing tools and methods for a practical application of the geometric approach.

In the following, we will mainly deal with Noetherian rings, that is rings in which non decreasing chains of ideals are stationary (see [3] or [19] for more details), having no zero divisors. Examples of Noetherian rings which are of interest in control theory are the rings of polynomials in one or several variables with real coefficients $\mathbb{R}[\Delta_1, ..., \Delta_k]$, $k \geq 1$. For $k = 1$, $\mathbb{R}[\Delta]$ is, in addition, a principal ideal domain (p.i.d), that is a ring in which any ideal has a single generator.

3 Time delay systems and systems over rings

As a motivation to study systems over rings, let us recall the relationship between them and time delay systems. In general, a *time-invariant, linear, time delay system with non commensurable delays* $h_1, \ldots, h_k$, $h_i \in \mathbb{R}^+$, for $i = 1, \ldots, k$, is described by equations of the the following form

$$\Sigma_d \;=\; \begin{cases} \dot{x}(t) = \sum_{i=1}^{k} \sum_{j=0}^{a} \; A_{ij} x(t - jh_i) + \sum_{i=1}^{k} \sum_{j=0}^{b} \; B_{ij} u(t - jh_i) \\ y(t) = \sum_{i=1}^{k} \sum_{j=0}^{c} \; C_i x(t - jh_i) \end{cases} \tag{2}$$

where A_{ij}, B_{ij}, and C_{ij} are matrices of suitable dimensions with entries in $\mathbb{R}$. In the last years, a great research effort has been devoted to the development of analysis and synthesis techniques for this kind of systems, mainly extending tools and methods from the framework of linear systems with coefficients in a field (see e.g. [5] and [13]). A large part of the difficulties in dealing with systems of the form (2) is due to the fact that their state space has infinite dimension. In order to circumvent this, it is useful to associate to a time delay system a suitable system with coefficients in a ring, as described in the following.

For any delay h_j, let us introduce the delay operator δ_j defined, for any time function $f(t)$, by $\delta_j f(t) := f(t - h_j)$. Accordingly, we can re-write the system (2) as

$$\Sigma_d \;=\; \begin{cases} \dot{x}(t) = \sum_{i=1}^{a} \sum_{j=0}^{k} \; A_{ij} \delta_j^i x(t) + \sum_{i=1}^{b} \sum_{j=0}^{k} B_{ij} \delta_j^i u(t) \\ y(t) = \sum_{i=1}^{c} \sum_{j=0}^{k} \; C_{ij} \delta_j^i x(t) \end{cases}$$

Now, by formally replacing the delay operators δ_j by the algebraic unknowns Δ_j, it is possible to associate to Σ_d the discrete-time system Σ over the ring $\mathcal{R} = \mathbb{R}[\Delta_1, ..., \Delta_k]$ defined by the equations

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) \\ \quad\;\; y(t) = Cx(t) \end{cases}$$

where the matrices A, B, C are defined as

$$A := \sum_{i=1}^{a} \sum_{j=0}^{k} \; A_{ij} \Delta_j^i, \quad B := \sum_{i=1}^{b} \sum_{j=0}^{k} \; B_{ij} \Delta_j^i, \quad C := \sum_{i=1}^{c} \sum_{j=0}^{k} \; C_{ij} \Delta_j^i$$

Actually, the time delay system Σ_d and the associated system Σ over $\mathbb{R}[\Delta_1, ..., \Delta_k]$ are quite different objects from a dynamical point of view, but they share the structural properties that depend on the defining matrices.

Therefore, control problems concerning the input/output behavior of Σ_d can be naturally formulated in terms of the input/output behavior of Σ and they can possibly be solved in the framework of systems over rings, whose state spaces are finite dimensional modules. In turn, solutions found in that framework often can be interpreted in the original delay-differential framework, to solve the original problem. This is the case, for instance, of the Disturbance Decoupling Problem considered below, as well as many as for others (see [9], [11] and references therein).

4 Statement of the Disturbance Decoupling Problem

Let the system Σ be described over the ring $\mathcal{R}$ by equations of the form

$$\Sigma \;=\; \begin{cases} x(t+1) = Ax(t) + Bu(t) + Dq(t) \\ \quad\;\; y(t) = Cx(t) \end{cases} \tag{3}$$

where $q \in \mathcal{Q} = \mathcal{R}^k$ is a disturbance.

The **Disturbance Decoupling Problem with Measurable Disturbances** (DDPM) for Σ consists in finding an integer $n_a \geq 0$ and a dynamic feedback law of the form

$$\begin{cases} x_a(t+1) = A_1 x(t) + A_2 x_a(t) + G_1 q(t) \\ \quad\;\; u(t) = Fx(t) + Hx_a(t) + G_2 q(t) \end{cases} \tag{4}$$

where $x_a \in \mathcal{X}_a := \mathcal{R}^{n_a}$, A_1, A_2, F, H, G_1 and G_2 are matrices of suitable dimensions with entries in the ring $\mathcal{R}$, such that the output of the compensated system Σ_c does not depend on q.
In case $n_a = 0$, (4) reduces to a static feedback law.

The standard way to solve the DDPM makes use of fundamental notions of the geometric approach, as recalled below.

Definition 1. *[14] Let Σ be a system defined over $\mathcal{R}$ by equations of the form (1). A submodule $\mathcal{V} \subseteq \mathcal{X} = \mathcal{R}^n$ is called*
i) *(A, B)-invariant or controlled invariant if*

$$A\mathcal{V} \subseteq \mathcal{V} + \mathrm{Im}B$$

ii) *$(A+BF)$-invariant or controlled invariant of feedback type, shortly feedback invariant, if there exists a linear map $F : \mathcal{X} \longrightarrow \mathcal{U}$ such that*

$$(A + BF)\mathcal{V} \subseteq \mathcal{V}$$

Any such F is called a friend of $\mathcal{V}$.

The family of controlled invariant submodules contained in a given submodule $\mathcal{K} \subseteq \mathcal{R}^n$ is closed with respect to the sum, therefore, if $\mathcal{R}$ is a Noetherian ring, it has a maximum element denoted by $\mathcal{V}^*(\mathcal{K})$.
The main result concerning the DDPM is stated as follows.

Proposition 1. *[7] Given a system Σ, defined over the Noetherian ring $\mathcal{R}$ by equations of the form (3), the DDPM for Σ is solvable if and only if*

$$Im\ D \subseteq \mathcal{V}^* + ImB \tag{5}$$

where $\mathcal{V}^*$ is the maximum (A, B)-invariant submodule contained in $Ker\ C$.

It has to be remarked that, if $\mathcal{R}$ is a field, *i)* and *ii)* in Definition 1 are equivalent (see [4], [24]). In that case, as well as in all cases in which $\mathcal{V}^*$ is feedback invariant, a static solution to the DDPM, when (5) holds, is simply given by any friend F of $\mathcal{V}^*$, whose feedback action causes $\mathcal{V}^*$ to become invariant with respect to the closed loop dynamics and forces the image of the disturbance to evolve in $Ker\ C$. Obviously, any other controlled invariant $\mathcal{V} \subseteq Ker\ C$ such that $Im\ D \subseteq \mathcal{V} + ImB$ can be alternatively used to define, by any friend, a solution.
If $\mathcal{R}$ is not a field, *i)* and *ii)* are no longer equivalent (easy examples can for instance be constructed over the ring of real polynomials in one indeterminate, see [14]) and, if $\mathcal{V}^*$ is not of feedback type, in order to obtain a solution it is necessary to construct a suitable extension Σ_e of Σ. This is motivated by the following result.

Proposition 2. *[9] Let Σ be a system defined by equations of the form (1) over a Noetherian ring $\mathcal{R}$. Assume that the controlled invariant submodule $\mathcal{V}$ is a direct summand of the state module $\mathcal{X}$, i.e. there exists a submodule $\mathcal{W}$ such that $\mathcal{X} = \mathcal{W} \oplus \mathcal{V}$, then $\mathcal{V}$ is of feedback type.*

Basically, the procedure to find a solution to the DDPM consists in considering the system extension Σ_e given by the equations

$$\Sigma_e = \begin{cases} x_e(t+1) = A_e x_e(t) + B_e u_e(t) + D_e q(t) \\ \qquad\; y(t) = C_e x_e(t) \end{cases}, \tag{6}$$

with

$$x_e = \begin{bmatrix} x(t) \\ x_a(t) \end{bmatrix},\ u_e = \begin{bmatrix} u(t) \\ u_a(t) \end{bmatrix}$$

$$A_e = \begin{bmatrix} A & 0 \\ 0 & 0_{r\times r} \end{bmatrix},\ B_e = \begin{bmatrix} B & 0 \\ 0 & I_{r\times r} \end{bmatrix},\ D_e = \begin{bmatrix} D \\ 0_{r\times r} \end{bmatrix},\ C_e = [C\ 0_{r\times r}]$$

where r is the dimension of $\mathcal{V}^*$. The module $\mathcal{V}_e$ generated by the columns of the matrix $V_e = \begin{bmatrix} V \\ I_{r\times r} \end{bmatrix}$, where V is a matrix whose columns generates $\mathcal{V}^*$, is, by construction, a direct summand of $\mathcal{X}_e = \mathcal{R}^n \bigoplus \mathcal{R}^r$ contained in $Ker\ C_e$ and

$$Im\ D_e \subseteq V_e + Im\ B_e.$$

Henceforth, a solution is represented by a friend of $\mathcal{V}_e$. Remark that the above construction can be carried on starting from any controlled invariant submodule $\mathcal{V} \subseteq Ker\ C$ for which $Im\ D \subseteq \mathcal{V} + Im\ B$.

Remark 1. Proposition 1 provides a complete characterization of the solvability of the DDPM over a Noetherian ring, but its use in practice requires the possibility to compute $\mathcal{V}^*$. If $\mathcal{R}$ is a field this can be done by the so called *Invariant Subspace Algorithm* (ISA)(see [24]). Unfortunately ISA does not work if $\mathcal{R}$ is not a field and, although an alternative algorithm has been given in [2] for Principal Ideal Domains, no algorithm is available for the more general case of Noetherian rings.

5 A computable solution

A different way to construct a solution to the DDPM, which avoids the use of $\mathcal{V}^*$ and is based on computable submodules, has been proposed in [1]. To describe it, let us introduce some other geometric notions.

Definition 2. *Let Σ be a system defined over $\mathcal{R}$ by equations of the form (1). A submodule $\mathcal{S} \subseteq \mathcal{X}$ is called (A, C)-invariant or conditioned invariant if*

$$A(\mathcal{S} \cap \ker C) \subseteq \mathcal{S}$$

The family of all conditioned invariant submodules containing a given submodule $\mathcal{K}$ has a minimal element, denoted by $\mathcal{S}^*(\mathcal{K})$. It can be shown that $\mathcal{S}^*(\mathcal{K})$ coincides with the limit of the following sequence

$$\begin{aligned} \mathcal{S}_0 &= \mathcal{K} \\ \mathcal{S}_{i+1} &= \mathcal{K} + A(\mathcal{S}_i \cap \ker C) \end{aligned} \tag{7}$$

Over a Noetherian ring the a sequence (7) being non decreasing, converges in a finite number of steps, so providing an algorithm to compute practically $\mathcal{S}^*(\mathcal{K})$ (see [21]).

Definition 3. *[8], [9] Let Σ be a system defined over $\mathcal{R}$ by equations of the form (1). A submodule $\mathcal{R} \subseteq \mathcal{X}$ is said to be a pre–controllability submodule if*

i) $\mathcal{R}$ is controlled invariant;
ii) $\mathcal{R} = \mathcal{S}_*(\mathcal{R})$, where $\mathcal{S}_*(\mathcal{R})$ is the minimum element of the family $\mathcal{S}_R$

$$\mathcal{S}_R = \{\mathcal{S} \subseteq \mathcal{X} \text{ such that } \mathcal{S} = \mathcal{R} \cap (A\mathcal{S} + ImB)\}. \tag{8}$$

The family of all pre–controllability submodules contained in a given submodule $\mathcal{K}$ has a maximum element, denoted by $\mathcal{R}^*(\mathcal{K})$.
It has been shown in [1] that $\mathcal{R}^*(\mathcal{K})$ coincides with the limit of the following sequence

$$\begin{cases} \mathcal{R}_0 := \mathcal{S}^*(Im\ B) \cap \mathcal{K} \cap A^{-1}(ImB) \\ \mathcal{R}_k := \mathcal{S}^*(Im\ B) \cap \mathcal{K} \cap A^{-1}(\mathcal{R}_{k-1} + Im\ B) \end{cases} \tag{9}$$

where A^{-1} denotes the inverse image of A and $\mathcal{S}^*(Im\ B)$ the minimal conditioned invariant submodule containing $Im\ B$.
Over a Noetherian ring $\mathcal{R}$ the a sequence (9), being non decreasing, converges in a finite number of steps, so providing an algorithm to compute practically $\mathcal{R}^*(\mathcal{K})$ (see [21]).
Remark that in case of a system Σ defined by equations (3) we can construct pre–controllability submodules using either the input matrix B or the input matrix $[B\ D]$. If necessary, we will make distinction by using a suffix.
We can now give an alternative, computable condition for the solution of DDPM.

Proposition 3. *[1] Given a system Σ, defined over the Noetherian ring $\mathcal{R}$ by equations of the form (3), the DDPM for Σ is solvable if and only if*

$$ImD \subseteq \mathcal{R}^*_{[BD]}(Ker\ C) + Im\ B \tag{10}$$

where $\mathcal{R}^*_{[B\ \ D]}(Ker\ C)$ is the maximum pre-controllability submodule in $Ker\ C$ constructed with respect to the input matrix $[B\ D]$.

The proof of Proposition 3 relies on the fact that condition (10) is equivalent to condition (5) of Proposition 1, namely $Im\ D \subseteq \mathcal{V}^* + Im\ B$, as shown in [1].

6 DDPM with stability

In addition to being useful for stating a computable condition for the solution of the DDPM, pre–controllability subspaces are instrumental in analyzing the structure of the closed loop system in which the disturbance is decoupled.
This fact turns out to be useful in considering the DDPM with the additional requirement of stability. Since a ring cannot, in general, be endowed with a natural metric structure, stability for systems with coefficient in a ring must be dealt with in a formal way as follows (see [16] and [12]).
Given a ring $\mathcal{R}$, a subset $S \subseteq \mathcal{R}[z]$ of polynomials in the indeterminate z with coefficients in $\mathcal{R}$ is said to be an *Hurwitz set* if (i) it is multiplicatively closed, (ii) it contains at least an element of the form $z - \alpha$, with $\alpha \in \mathcal{R}$,(iii) it contains all the monic factors of all its elements.

Definition 4. *Given a system Σ of the form (1) with coefficients in $\mathcal{R}$ and an Hurwitz set S, $\Sigma$is said S-stable if $det(zI - A)$ belongs to S.*

Remark 2. Referring to the relation between time delay systems and systems over rings in the case of a single delay, S-stability in the ring framework corresponds to stability in the time delay framework if the Hurwitz set S is chosen as

$$S = \{p(z, \Delta) \in \mathcal{R}[z] \text{ such that } p(\gamma, e^{-\gamma h}) \neq 0 \\ \text{for all } \gamma \in \mathcal{C} \text{with} \mathbb{R}\gamma \geq 0\}. \tag{11}$$

The **Disturbance Decoupling Problem with Measurable Disturbances and S-stability** (DDPMS) for a system Σ of the form (3) and a given Hurwitz set S consists in finding an integer $n_a \geq 0$ and a dynamic feedback law of the form (4), such that the compensated system Σ_c is S-stable and its output does not depend on q.

To investigate the solvability conditions of the DDPMS let us consider the construction of the solution to the DDPM described in Section 4. There, in order to extend Σ to Σ_e, we start from a controlled invariant submodule $\mathcal{V} \subseteq \mathcal{X}$ of dimension r, for which $Im\ D \subseteq \mathcal{V} + Im\ B$ holds, which can be either $\mathcal{V}^*$, if it can be computed, or $\mathcal{R}^*_{[B\ D]}(Ker\ C)$. In both cases let us denote by $\mathcal{W}$ the maximum pre–controllability submodule $\mathcal{R}^*_{[B]}(\mathcal{V})$. In Σ_e we have the (A_e, B_e)-invariant submodules $\mathcal{W}_e = span \begin{bmatrix} W \\ I \end{bmatrix}$ and $\mathcal{V}_e = span \begin{bmatrix} V \\ I \end{bmatrix}$, where W and V are matrices whose columns span $\mathcal{W}$ and $\mathcal{V}$, respectively.

Assuming that $\mathcal{W}_e$ is a direct summand of $\mathcal{V}_e$, as it happens when $\mathcal{R} = \mathbb{R}[\Delta_1, \ldots, \Delta_k]$, we can write $\mathcal{X}_e = \mathcal{X} \oplus \mathcal{R}^r = \mathcal{W}_e \oplus \mathcal{W}_1 \oplus \mathcal{W}_2$ for some submodules $\mathcal{W}_1$ and $\mathcal{W}_2$, such that $\mathcal{V}_e = \mathcal{W}_e \oplus \mathcal{W}_1$. Choosing a basis of $\mathcal{X}_e$ consisting of the union of basis of $\mathcal{W}_e$, $\mathcal{W}_1$, $\mathcal{W}_2$ and expressing A_e and B_e accordingly, we get

$$A_e = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}; B_e = \begin{bmatrix} B_1 & 0 \\ 0 & 0 \\ 0 & B_3 \end{bmatrix}. \tag{12}$$

The structure of B_e depends on the fact that $Im\ B \bigcap \mathcal{V}$ is contained in $\mathcal{R}^*_{[B]}$. The structure of A_e depends on this and on the fact that $\mathcal{W}_e$ is (A_e, B_e)-invariant. The dynamic matrix A_{ec} of the compensated system, for any friend $F = \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \end{bmatrix}$ of $\mathcal{V}_e$, takes therefore the form

$$A_{ec} = \begin{bmatrix} A_{11} + B_1 F_{11} & A_{12} + B_1 F_{12} & A_{13} + B_1 F_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} + B_3 F_{23} \end{bmatrix} \tag{13}$$

showing that the dynamics of the block A_{22} remains fixed for any choice of the friend F.

We can therefore state the following Proposition.

Proposition 4. *With the above notations, a necessary condition for the solvability of the DDPMS for Σ is that $det(sI - A_{22})$ belongs to S.*

From the structure of A_{ec}, it can be seen that a sufficient condition for the solvability of the DDPMS for Σ depends on the possibility of choosing F_{11} and F_{33} so that $det(sI - (A_{11} + B_1F_{11})$ and $det(sI - (A_{33} + B_3F_{23}))$ are elements of S. Note that the components F_{11} and F_{33} of F can actually be chosen arbitrarily. To this aim it is useful to recall the following definition.

Definition 5. *Let the system Σ be described over the ring $\mathcal{R}$ by equations of the form (1). Σ, or, alternatively, the pair (A, B), is said to be coefficient assignable if, by a suitable choice of the feedback F, the coefficients of $det(sI - (A + BF))$ can be arbitrarily assigned.*

Proposition 5. *With the above notations, a sufficient condition for the solvability of the DDPMS for Σ is that $det(sI - A_{22})$ belongs to S and that the pairs (A_{11}, B_1) and (A_{33}, B_3) are coefficient assignable.*

Coefficient assignability for systems over rings is related to structural properties, like (strong) reachability, in different ways, depending on the properties of the ring of coefficients (see e.g. [6]).

Remark 3. Keeping in mind the content of Section 2, it is clear that the DDPMS has a corresponding version in the time-delay framework. Letting the Hurwitz set S be chosen as in (11), the above results can be applied to the disturbance decoupling problem with the requirement of stability, in the classical sense, in the time-delay framework, getting conditions for its solvability. These hold in a more general situation than that considered in [10], where the slightly restrictive hypothesis $\mathcal{V}^* \bigcap Im\ B$ was assumed.

References

1. Assan J., J.F. Lafay and A.M. Perdon (1999) *Systems and Control Letters*, 37, pp. 153-161
2. Assan J., J.F. Lafay, A.M. Perdon, J.J. Loiseau (1999) *Proceedings 38th IEEE-CDC*, 4, pp. 4216 - 4221, Phoenix, AZ
3. Atiyah M. F. and I. G. Macdonald (1969) Introduction to Commutative Algebra, *Addison-Wesley Series in Mathematics*, Addison-Wesley Publishing Company
4. Basile G. and G. Marro (1992) Controlled and Conditioned Invariants in Linear System Theory, *Prentice Hall*, Englewood Cliffs, New Jersey
5. Boukas E. K. and Z. K. Liu (2002) Deterministic and Stochastic Time-Delay Systems,*Series on Control Engineering*, Birkhauser Boston
6. Brewer J.W., J. W. Bunce and F. S. Van Vleck (1986), *Linear Systems Over Commutative Rings*, Marcel Dekker, New York
7. Conte G. and A. M. Perdon (1995) *SIAM Journal on Control and Optimization* , 33
8. Conte G. and A. M. Perdon (1995) The Decoupling Problem for Systems over a Ring, *Proceedings of 34th IEEE CDC Conference*, New Orleans, Louisiana
9. Conte G. and A.M. Perdon (1998) The Block Decoupling Problem for Systems over a Ring, *IEEE Trans. Automatic Control*, 43

10. Conte G. and A.M. Perdon (1999) Disturbance decoupling with stability for delay differential systems, *Proceedings 38th IEEE-CDC*, Phoenix, AZ
11. Conte G. and A.M. Perdon (2000) Systems over Rings: Geometric Theory and Applications, *Annual Review in Control*, 24
12. Datta K.B. and M.L.J. Hautus (1984) Decoupling of multivariable control systems over unique factorization domains, *SIAM J. on Control and Optimization* , 22, 1
13. Gu K., V. L. Kahritonov and J. Chen (2003) Stability of Time-Delay Systems, *Series on Control Engineering*, Birkhauser Boston
14. Hautus M.L.J. (1982) Controlled invariance in sytems over rings, *Springer Lecture Notes in Control and Information Sciences*, 39
15. Hautus M.L.J. (1984) Disturbance rejection for systems over rings, *Springer Lecture Notes in Control and Information Sciences*, 58 (1984)
16. Hautus M.L.J. and E.D. Sontag (1980) In *Algebraic and Geometric Methods in Linear Systems Theory*, C. Byrnes and C. Martin Eds., Lecture Notes in Applied Mathematics, American Mathematical Society Publications, bf 18, Providence, RI, pp. 99–136
17. Kamen Edward W. (1975) *Mathematical System Theory*, 9, 1 pp. 57–74
18. Kamen Edward W. (1978) Lectures on Algebraic System Theory: Linear Systems Over Rings, *N.A.S.A. Contractor Report n.3016*
19. Lang Serge (1984) Algebra, 2nd edition,*Advanced Book Program*, Addison-Wesley Publishing Company, Menlo Park, California
20. Perdon A.M., G. Guidone-Peroli and M. Caboara (2003) Algorithms for geometric control of systems over rings, *Proceedings 2nd IFAC Conference Control Systems (CSD'03)*, Bratislava, Slovak Republic
21. Perdon A.M., M. Anderlucci and M. Caboara (2006) *International Journal of Control*, 79, 11 pp. 1401–1417
22. Sontag E.D. (1976) Linear systems over commutative rings: A survey, *Ricerche di Automatica*, 7, pp. 1–34
23. Sontag E.D. (1981) *Control science and technology for the progress of society*, 1, pp 325–330, Kyoto
24. Wonham M. (1985) *Linear multivariable control: a geometric approach, 3rd Ed.* (Springer Verlag, 1985).

Some comments on ν-Support Vector Machines

Francesco Dinuzzo and Giuseppe De Nicolao

Dipartimento di Informatica e Sistemistica, Università di Pavia,
francesco.dinuzzo@gmail.com, giuseppe.denicolao@unipv.it

1 Introduction

In the last decade, Support Vector Machines (SVM) [11] have emerged as powerful methods for learning from examples. Indeed, they have been applied successfully to both classification and regression problems. A notable property of obtained solutions is that they depend only on a sparse selection of training examples called support vectors.

However, the performance of SVM heavily depends on the choice of some design parameters like the complexity parameter C in classification and also the width of the dead zone ϵ in regression. More precisely, the complexity parameter controls the degree of regularization that is forced into the solution and corresponds to the inverse of the regularization parameter used in the Tikhonov regularization approach to the solution of ill-posed problems. On the other hand, the ϵ parameter is just the width of the dead zone in the ϵ-insensitive loss function, meaning that residuals smaller than ϵ are not accounted for in the cost function. By the way, this guarantees insensitivity of the solution against the data associated with such residuals.

The search for automated tuning procedures for the design parameters is motivated by the fact that these parameters, especially the complexity one, do not have an intuitive interpretation that could be exploited to get hints on their optimal values. A robust tuning of the design parameters could be obtained by k-fold cross validation that has the further advantage of not relying on strong statistical assumptions. However, k-fold cross validation is computationally expensive as it requires the calculation of k distinct solutions for each candidate vector of design parameters, which renders the search of the optimal parameter vector rather cumbersome. This motivates the interest for alternative tuning methods. In the regression context, it appears natural to extend approaches already available for other kernel methods, such as regularization networks and Gaussian processes. Among these approaches one may mention so-called objective criteria such as GCV, AIC, C_p. All these criteria rely on the degrees of freedom of the estimator, a measure of the sensitivity of the estimate with respect to the training data. Such sensitivity is trivially evaluated for linear algorithms, e.g. regularization networks and Gaussian processes, but, until recently, a thorough analysis was not available for SVM.

Recent results have demonstrated that the approximate degrees of freedom of regression SVM are equal to the number of marginal support vectors, that is the number of residuals whose absolute value equals ϵ [2, 3, 6]. This result opens the way to the use of several established tuning criteria. However, due to the intrinsic nonlinearity of SVM, the number of support marginal vectors constitutes only an approximation of the true degrees of freedom. Moreover, the tuning of classification SVM remains an open problem.

An alternative way to address the tuning problem is to introduce modified SVM whose parameters are more easily interpreted. Along this direction, ν-SVM [8] rely on the parameter ν that is directly related to the number of support vectors. More specifically, it can be shown that ν is a lower bound to the fraction of support vectors and an upper bound to the fraction of outliers[1]. Moreover, under proper assumptions, it can be shown that ν converges with probability 1 to the asymptotic fraction of support vectors as the training set size grows to infinity. This facts are collectively known as the "ν-property" [7]. It is apparent that ν is directly related to the degree of sparsity of the solution so that, differently from parameter C, sensible values of ν may be guessed from the properties of the problem at hand.

Although SVM are often introduced as finite dimensional optimization problems, it is well known that they can be seen as the solution of regularization problems in Reproducing Kernel Hilbert Spaces (RKHS) with convex and non-smooth loss functions. Subdifferential calculus is an analytic tool that is particularly well suited for dealing with non-smooth convex problems. In particular, subdifferential calculus has been recently used to obtain quantitative representation results for kernel machines [9, 1, 2, 4]. The main purpose of the present note is to provide an alternative derivation of the "ν-property", showing that it follows directly from a restatement of a necessary condition for optimality involving subdifferentials.

After some preliminaries (Section 2), the derivation is carried out for regression SVM (Section 3), classification SVM (Section 4), and also novelty detection SVM (Section 5). A generalized formulation of the ν-property that encompasses all the previous cases is given in Section 6. Regression ν-SVM with the Huber loss function are discussed in Section 7, where the solution to an open problem stated in [7] is provided. Finally, in the conclusions (Section 8), we conjecture that better bounds for ν could be obtained if the order of application of the necessary conditions were exchanged.

[1] Here, the term "outlier" is used in a generic sense to indicate support vectors that do not fall at the margin of the used loss function.

2 Preliminaries

First, let us introduce some notation relative to supervised and unsupervised learning problems. In a supervised learning problem, we have two sets X (input set) and Y (output set). A training set $D = \{(x_i, y_i)\}_{i=1}^{\ell}$ of pairs independently extracted from a distribution P over $X \times Y$ is available. The goal is to estimate a decision function

$$g : X \to Y,$$

that predicts the output, given a test input. In a two-class classification problem, the output set can be taken as $Y = \{-1, +1\}$, while, in a regression problem, the set Y is the set of real numbers $\mathbb{R}$.

In an unsupervised learning problem, we only have one set X. The training data are $D = \{x_i\}_{i=1}^{\ell}$, independently extracted from a distribution P over X, and the goal is to extract some relevant information about P.

Let us introduce a similarity measure in X

$$\begin{aligned} K : X \times X &\to \mathbb{R}, \\ (x_1, x_2) &\to K(x_1, x_2), \end{aligned}$$

namely a function that, given two inputs x_1 and x_2, returns a real number that measures their similarity. More specifically, we are interested in similarity measures, called kernels, that corresponds to dot products in some dot product space $\mathcal{H}$. We suppose that the input data are mapped into $\mathcal{H}$ via a nonlinear operator $\mathbf{\Phi}$ (feature map)

$$\begin{aligned} \mathbf{\Phi} : X &\to \mathcal{H}, \\ x &\to \mathbf{x}, \end{aligned}$$

and that K is defined as

$$K(x_1, x_2) = \langle \mathbf{\Phi}(x_1), \mathbf{\Phi}(x_2) \rangle_{\mathcal{H}} = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_{\mathcal{H}} .$$

Support Vector Machines are kernel-based learning algorithms and are able to deal with classification, regression and unsupervised learning problems.

As an example, let us formulate the ν-SVM algorithm for classification, hereafter named ν-SVC. Its aim is to estimate a decision function of the form $\hat{g}(x) = \mathrm{sgn}(\hat{f}(x))$, with

$$\begin{aligned} \hat{f}(x) &= \langle \mathbf{w}, \mathbf{\Phi}(x) \rangle_{\mathcal{H}} + b \\ \mathbf{w} &\in \mathcal{H}, b \in \mathbb{R}. \end{aligned}$$

Vector $\mathbf{w}$ (weights) and scalar b (bias) are obtained by solving the optimization problem

$$\begin{aligned} \min_{\substack{\mathbf{w}\in\mathcal{H},\\ \boldsymbol{\xi}\in\mathbb{R}^\ell\\ \rho,b\in\mathbb{R}}} \quad & \left(\frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{\ell}\sum_{i=1}^{\ell}\xi_i\right), \\ \text{subject to} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle_{\mathcal{H}} + b) \geq \rho - \xi_i, \\ & \xi_i \geq 0, \\ & \rho \geq 0. \end{aligned}$$

In order to solve this problem, it is usual to consider its dual Lagrangian formulation, which is just a quadratic programming problem. It turns out that the weight vector is given by

$$\mathbf{w} = \sum_{i=1}^{\ell} y_i \alpha_i \mathbf{\Phi}(x_i),$$

where α_i are the Lagrange multipliers of the dual formulation. By substituting this expression in the decision function, we have

$$g(x) = \operatorname{sgn}\left(\sum_{i=1}^{\ell} y_i\alpha_i \langle \mathbf{\Phi}(x_i), \mathbf{\Phi}(x)\rangle + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{\ell} y_i\alpha_i K(x_i, x) + b\right).$$

In general, some coefficients α_i will be equal to zero, so that the decision function depends only on the subset of training data associated with nonzero multipliers. These data are called support vectors. If, for some given i, it results $y_i\hat{f}(\mathbf{x}_i) > \rho$, we say that we have a margin error.

A major motivation for the introduction of ν-SVC is that they enjoy the following property.

Proposition 1 (ν-property [7]). *Suppose that we run ν-SVC with some kernel K, with the result $\rho > 0$. Then:*

1. *ν is an upper bound on the fraction of margin errors.*
2. *ν is a lower bound on the fraction of support vectors.*
3. *Suppose that the joint distribution $P(x, y)$ is such that neither $P(x, y = 1)$ nor $P(x, y = -1)$ contains any discrete component. Suppose, moreover, that the kernel K is analytic and non-constant. With probability 1, asymptotically, ν equals both the fraction of support vectors and the fraction of margin errors.*

The original proof of this property relies on the analysis of the dual problem. However, Support Vector Machines can be also formulated as regularization problems in RKHS, see e.g. [5] . In the following sections, assuming that $\mathcal{H}$ is an RKHS of functions $f : X \to Y$, we will show that the ν-property admits a simple and direct proof based on subdifferential calculus. Without loss of generality, we will neglect the bias term b in the formulation of the algorithm.

In this note, we are interested only in the first two points of the ν-property. The third point guarantees that, under the stated assumption on the distribution P and the kernel, the bounds for ν of the first two points become more accurate as the sample size ℓ grows to infinity. An intuitive justification for the third point is that the asymptotic margin tends to a zero measure set in the input space, so that the fraction of data that lie exactly on the margin tends to zero.

3 ν-Support Vector Regression

The ν-SVR (Support Vector Regression) problem consists of finding the pair $(\hat{f}, \hat{\epsilon})$ that satisfies

$$(\hat{f}, \hat{\epsilon}) = \arg \min_{\substack{f \in \mathcal{H} \\ \epsilon \in \mathbb{R}^+}} \left(\ell C \nu \epsilon + C \sum_{i=1}^{\ell} V\big(y_i - f(\mathbf{x}_i), \epsilon\big) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right),$$

where $\nu \in [0,1]$, $C > 0$ and V is the ϵ-insensitive loss function defined by

$$V(y_i - \hat{f}(\mathbf{x}_i), \epsilon) = \begin{cases} 0, & |y_i - \hat{f}(\mathbf{x}_i)| \leq \epsilon \\ |y_i - \hat{f}(\mathbf{x}_i)| - \epsilon, & |y_i - \hat{f}(\mathbf{x}_i)| > \epsilon \end{cases}$$

Note that in standard formulations of SVR, ϵ is a fixed parameter. On the contrary, here we write $V(y_i - \hat{f}(\mathbf{x}_i), \epsilon)$ to underline the dependency of the loss function on ϵ .

Observe that $V(y_i - \hat{f}(\mathbf{x}_i), \epsilon)$ is convex in *both* the arguments $y_i - \hat{f}(\mathbf{x}_i)$ and ϵ. This observation can be used to obtain in a very simple way the so-called ν-property. First, consider the case in which the optimal value of ϵ is different from zero. Then, by minimizing with respect to ϵ, we can write the following necessary condition for optimality:

$$0 \in \partial_\epsilon \left(\ell C \nu \epsilon + C \sum_{i=1}^{\ell} V\big(y_i - f(\mathbf{x}_i), \epsilon\big) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right),$$

where ∂_ϵ denotes the subdifferential with respect to ϵ.

Since the regularization term does not depend on ϵ and, by standard properties of subdifferentials, the subdifferential of the sum of convex functions coincides with the sum of subdifferentials, we can write

$$\nu \in -\frac{1}{\ell} \sum_{i=1}^{\ell} \partial_\epsilon V\big(y_i - f(\mathbf{x}_i), \epsilon\big).$$

It is easy to see that

$$\partial_\epsilon V\big(y_i - f(\mathbf{x}_i), \epsilon\big) = \begin{cases} \{0\}, & |y_i - f(\mathbf{x}_i)| < \epsilon \\ [-1, 0], & |y_i - f(\mathbf{x}_i)| = \epsilon \\ \{-1\}, & |y_i - f(\mathbf{x}_i)| > \epsilon \end{cases}$$

Let

$$I = \{1, 2, \dots, \ell\},$$

and define the index sets I_{in}, I_M, I_{out} associated with data inside the ϵ-tube, on the margin of the ϵ-tube, and outside the ϵ-tube, respectively:

$$\begin{aligned} I_{in} &= \{i \in I : |y_i - \hat{f}(x_i)| < \epsilon\}, \\ I_M &= \{i \in I : |y_i - \hat{f}(x_i)| = \epsilon\}, \\ I_{out} &= \{i \in I : |y_i - \hat{f}(x_i)| > \epsilon.\} \end{aligned}$$

The set I_{SV} that identifies the support vectors is

$$I_{SV} = I_M \cup I_{out}.$$

Then, the ν-property immediately follows observing that

$$\partial_\epsilon V\big(y_i - f(\mathbf{x}_i), \epsilon\big) = \begin{cases} \{0\}, & i \in I_{in} \\ [-1, 0], & i \in I_M \\ \{-1\}, & i \in I_{out} \end{cases}$$

implies

$$\nu \in -\frac{1}{\ell}\left[\sum_{i \in I_M}^{\ell} [-1, 0] + \sum_{i \in I_{out}}^{\ell} (-1)\right] = \left[\frac{\#I_{out}}{\ell}, \frac{\#I_{SV}}{\ell}\right].$$

Note that when the optimal value of ϵ is zero, we still have $\nu \geq \#I_{out}/\ell$. Indeed, let us impose that the right derivative of the functional with respect to ϵ is nonnegative:

$$D_\epsilon^+\left(\ell C \nu \epsilon + C \sum_{i=1}^{\ell} V\big(y_i - f(\mathbf{x}_i), \epsilon\big) + \frac{1}{2}\|f\|_{\mathcal{H}}^2\right) \geq 0,$$

Then,

$$\nu \geq -\frac{1}{\ell} \sum_{i \in I_{out}}^{\ell} (-1) - \sum_{i \in I_M \cup I_{in}}^{\ell} 0 = \frac{\#I_{out}}{\ell}.$$

4 ν-Support Vector Classification

The ν-SVC algorithm amounts to solving the variational problem

$$(\hat{f}, \hat{\rho}) = \arg \min_{\substack{f \in \mathcal{H} \\ \rho \in \mathbb{R}^+}} \left(-\ell\nu\rho + \sum_{i=1}^{\ell} V\big(y_i f(\mathbf{x}_i), \rho\big) + \frac{1}{2}\|f\|_{\mathcal{H}}^2 \right),$$

where $\nu \in [0, 1]$, and V is the hinge loss function with adaptive margin defined by

$$V(y_i \hat{f}(\mathbf{x}_i), \rho) = \Big(\rho - y_i \hat{f}(\mathbf{x}_i)\Big)_+ .$$

It can be easily seen seen that $V(y_i \hat{f}(\mathbf{x}_i), \rho)$ is convex with respect to both its arguments. Suppose that the optimal value of ρ is strictly positive. Then, by minimizing with respect to ρ, we have

$$0 \in \partial_\rho \left(-\ell\nu\rho + \sum_{i=1}^{\ell} V\big(y_i f(\mathbf{x}_i), \rho\big) + \frac{1}{2}\|f\|_{\mathcal{H}}^2 \right),$$

that is

$$\nu \in \frac{1}{\ell} \sum_{i=1}^{\ell} \partial_\rho V\big(y_i f(\mathbf{x}_i), \rho\big).$$

Define the index sets

$$\begin{aligned} I_{in} &= \{i \in I : y_i \hat{f}(\mathbf{x}_i) < \rho\} \\ I_M &= \{i \in I : y_i \hat{f}(\mathbf{x}_i) = \rho\} \\ I_{out} &= \{i \in I : y_i \hat{f}(\mathbf{x}_i) > \rho\} \\ I_{SV} &= I_M \cup I_{out}. \end{aligned}$$

The subdifferential of the loss function is equal to

$$\partial_\rho V\big(y_i f(\mathbf{x}_i), \rho\big) = \begin{cases} \{0\}, & i \in I_{in} \\ [0, 1], & i \in I_M \\ \{1\}, & i \in I_{out} \end{cases}$$

Again, it follows that

$$\nu \in \frac{1}{\ell} \left[\sum_{i \in I_M}^{\ell} [0, 1] + \sum_{i \in I_{out}}^{\ell} 1 \right] = \left[\frac{\# I_{out}}{\ell}, \frac{\# I_{SV}}{\ell} \right].$$

5 Algoritms for novelty detection

We now consider two algorithms for novelty detection: the One-Class SVM algorithm [7] and the Soft Margin Ball algorithm with ν-parametrization [7, 10]. In the novelty detection problem, it is assumed that all the training data belong to the same class and the goal is to estimate a function that decides if a new test point belongs to the same class of the training data or is a novel object.

The One-Class SVM algorithm is very similar to ν-SVC and can be formulated as follows:

$$(\hat{f}, \hat{\rho}) = \arg \min_{\substack{f \in \mathcal{H} \\ \rho \in \mathbb{R}}} \left(-\ell \nu \rho + \sum_{i=1}^{\ell} \left(\rho - f(\mathbf{x}_i)\right)_+ + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right).$$

Note that there are two subtle differences with respect to ν-SVC. First, the value of ρ is not constrained to be nonnegative (this is useful when $\nu = 0$). Second, there are no labels y_i. As a matter of fact, when $\nu > 0$, the One-Class SVM algorithm coincides with ν-SVC where all the labels y_i are set to $+1$. The derivation of the ν-property by means of subdifferential arguments is analogous to that for ν-SVC and is therefore omitted.

The Soft Margin Ball algorithm has to do with enclosing the images of the input data produced by the feature map

$$\mathbf{\Phi} : X \to \mathcal{H}, \qquad \mathbf{\Phi}(x) = K(x, \cdot),$$

in a small ball of radius R in the space $\mathcal{H}$, while allowing some outliers to be out of the ball. The associated variational problem is

$$(\hat{f}, \hat{R}) = \arg \min_{\substack{f \in \mathcal{H} \\ R \in \mathbb{R}^+}} \left(\ell \nu R^2 + \sum_{i=1}^{\ell} \left(\|f(\cdot) - K(x_i, \cdot)\|_{\mathcal{H}}^2 - R^2 \right)_+ \right)$$

Minimization with respect to R yields

$$0 \in \partial_R \left(\ell \nu R^2 + \sum_{i=1}^{\ell} \left(\|f(\cdot) - K(x_i, \cdot)\|_{\mathcal{H}}^2 - R^2 \right)_+ \right)$$

implying

$$\nu \in \frac{1}{\ell} \sum_{i=1}^{\ell} \partial \left(\|f(\cdot) - K(x_i, \cdot)\|_{\mathcal{H}}^2 - R^2 \right)_+$$

In analogy with the previous derivations, we define the index sets:

$$\begin{aligned} I_{in} &= \{ i \in I : \|f(\cdot) - K(x_i, \cdot)\|_{\mathcal{H}} < R \} \\ I_M &= \{ i \in I : \|f(\cdot) - K(x_i, \cdot)\|_{\mathcal{H}} = R \} \\ I_{out} &= \{ i \in I : \|f(\cdot) - K(x_i, \cdot)\|_{\mathcal{H}} > R \} \end{aligned}$$

Then, once again, the ν property follows:

$$\nu \in \frac{1}{\ell} \left[\sum_{i \in I_M}^{\ell} [0, 1] + \sum_{i \in I_{out}}^{\ell} 1 \right] = \left[\frac{\#I_{out}}{\ell}, \frac{\#I_{SV}}{\ell} \right]$$

6 Generalized ν-property

By comparing all the previous cases, we can conclude that the ν-property is a simple consequence of the structure of the subdifferential for the hinge loss and is associated with two general forms of variational problem:

$$(\hat{f}, \hat{\rho}) = \arg \min_{\substack{f \in \mathcal{H} \\ \rho \in \mathbb{R}^+}} \left(-\ell \nu \rho + \sum_{i=1}^{\ell} (\rho - \varPhi_i(f))_+ + \varOmega(f) \right),$$

$$(\hat{f}, \hat{\epsilon}) = \arg \min_{\substack{f \in \mathcal{H} \\ \epsilon \in \mathbb{R}^+}} \left(\ell \nu \epsilon + \sum_{i=1}^{\ell} (\varPhi_i(f) - \epsilon)_+ + \varOmega(f) \right),$$

where $\varPhi_i$ $(i = 1, \ldots, \ell)$ and $\varOmega$ are generic nonlinear functionals from $\mathcal{H}$ into $\mathbb{R}$.

All the specific examples in the previous sections can be seen as particular cases of the above forms in which $\varOmega(f) = \frac{1}{2}\|f\|^2_{\mathcal{H}}$:

- The ν-SVR algorithm corresponds to the second form with
$$\varPhi_i(f) = |f(x_i) - y_i|,$$
- The ν-SVC algorithm corresponds to the first form with
$$\varPhi_i(f) = f(x_i)y_i,$$
- The One Class SVM algorithm corresponds to the first form with
$$\varPhi_i(f) = f(x_i),$$
- The Soft Margin Ball algorithm corresponds to the second form with
$$\varPhi_i(f) = \|f(\cdot) - K(x_i, \cdot)\|^2_{\mathcal{H}}, \qquad \epsilon = R^2$$

In general, by defining the index sets

$$\begin{aligned} I_{in} &= \{i \in I : \varPhi_i(f) < \rho\} \\ I_M &= \{i \in I : \varPhi_i(f) = \rho\} \\ I_{out} &= \{i \in I : \varPhi_i(f) > \rho\} \\ I_{SV} &= I_M \cup I_{out}, \end{aligned}$$

we always have

$$\nu \geq \frac{\# I_{out}}{\ell},$$

and, for $\rho \neq 0$ ($\epsilon \neq 0$), we also have

$$\nu \in \left[\frac{\# I_{out}}{\ell}, \frac{\# I_{SV}}{\ell} \right]. \tag{1}$$

7 ν-SVR with Huber loss function

This section can be viewed as the solution of the open problem stated in exercise 9.21 of [7]. More precisely, we generalize ν-SVR to the Huber loss function, which is quadratic within a tube and linear outside (differently from the usual ϵ-insensitive one which is zero inside the ϵ-tube and linear outside). Moreover, we highlight the relationship between ν and the breakdown point σ of the Huber loss function.

Consider the problem

$$\left(\hat{f}, \hat{\sigma}\right) = \arg \min_{\substack{f \in \mathcal{H} \\ \sigma \in \mathbb{R}^+}} \left(\frac{\nu \ell \sigma}{2} + \sum_{i=1}^{\ell} V\big(y_i - f(\mathbf{x}_i), \sigma\big) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right)$$

where V is the Huber loss function defined by

$$V(y_i - \hat{f}(\mathbf{x}_i), \sigma) = \begin{cases} \frac{1}{2\sigma}(y_i - \hat{f}(\mathbf{x}_i))^2, & |y_i - \hat{f}(\mathbf{x}_i)| \leq \sigma \\ |y_i - \hat{f}(\mathbf{x}_i)| - \frac{\sigma}{2}, & |y_i - \hat{f}(\mathbf{x}_i)| > \sigma \end{cases}$$

Once again, the loss function is convex with respect to both its arguments. Moreover, $V\big(y_i - f(\mathbf{x}_i), \cdot\big)$ is also differentiable, so that it is not necessary to employ subdifferentials. It is sufficient to impose

$$0 = \frac{d}{d\sigma} \left(\frac{\nu \ell \sigma}{2} + \sum_{i=1}^{\ell} V\big(y_i - f(\mathbf{x}_i), \sigma\big) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right),$$

that is

$$\nu = -\frac{2}{\ell} \sum_{i=1}^{\ell} \frac{d}{d\sigma} V\big(y_i - f(\mathbf{x}_i), \sigma\big).$$

Now, we have

$$\frac{d}{d\sigma} V\big(y_i - f(\mathbf{x}_i), \sigma\big) = \begin{cases} \frac{-1}{2\sigma^2}(y_i - \hat{f}(\mathbf{x}_i))^2, & |y_i - \hat{f}(\mathbf{x}_i)| \leq \sigma \\ -\frac{1}{2}, & |y_i - \hat{f}(\mathbf{x}_i)| > \sigma \end{cases}$$

In analogy with the previous section, define the index sets

$$\begin{aligned} I_Q &= \{ i \in I : |y_i - \hat{f}(\mathbf{x}_i)| \leq \sigma \} \\ I_L &= \{ i \in I : |y_i - \hat{f}(\mathbf{x}_i)| > \sigma \} \end{aligned}$$

associated with the data whose residuals falls in the quadratic trait or in the linear trait of the Huber loss functions, respectively. We have

$$\nu = \frac{1}{\hat{\sigma}^2 \ell} \sum_{i \in I_Q} (y_i - \hat{f}(\mathbf{x}_i))^2 + \frac{\# I_L}{\ell},$$

$$\hat{\sigma} = \frac{1}{\sqrt{\nu}} \sqrt{\frac{\sum_{i \in I_Q} (y_i - \hat{f}(\mathbf{x}_i))^2}{\nu - \frac{\# I_L}{\ell}}}.$$

This last expression is interesting because it can be seen as an estimate of the noise standard deviation, parameterized by ν. Moreover, we see that it must be

$$\nu > \frac{\# I_L}{\ell}$$

implying that, also in this case, ν is an upper bound for the fraction of outliers.

8 Conclusions

In this note, we analyzed ν-SVM algorithms within the framework of Regularization Theory in Reproducing Kernel Hilbert Spaces. We considered the following algorithms: ν-SVC, ν-SVR, One Class SVM, Soft Margin Ball, and ν-Huber SVR. By exploiting the convexity of the objective functional, we showed that the so-called "ν-property" is a simple consequence of an optimality condition involving the subdifferential. We also showed that the ν-property can be generalized to a large class of kernel methods.

An interesting question is whether the inclusion (1) can be improved. In order to obtain (1), we just imposed the optimality condition with respect to ρ (or ϵ). Consider, for simplicity, the case of ν-SVC and suppose that we fix ρ to some feasible value. Now, we exchange the order of application of necessary conditions by first minimizing with respect to f. Then, the resulting $\hat{f}$ would also depend on ρ so that

$$\hat{\rho} = \arg \min_{\rho \in \mathbb{R}^+} \left(-\ell \nu \rho + \sum_{i=1}^{\ell} V\big(y_i \hat{f}(\mathbf{x}_i; \rho), \rho\big) + \frac{1}{2} \|\hat{f}(\cdot; \rho)\|_{\mathcal{H}}^2 \right).$$

We conjecture that this expression could be the starting point for deriving better bounds for ν.

Acknowledgements

This research has been supported by the Italian Ministry of University and Research through the FIRB Project "Learning theory and application".

References

1. E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. (2004). *Journal of Machine Learning Research*, 5:1363-1390.

2. F. Dinuzzo. (2007). Regularized Kernel Methods. Tesi di Laurea Specialistica, Università di Pavia, Italy.
3. F. Dinuzzo, M. Neve, U. P. Gianazza, and G. De Nicolao. (2007). On the representer theorem and equivalent degrees of freedom of SVR. Submitted.
4. F. Dinuzzo and G. De Nicolao. (2007). Unconstrained representation for regularized kernel methods. Submitted.
5. T. Evgeniou, M. Pontil, and T. Poggio. (2000). *Advances in Computational Mathematics*,13:1-150.
6. L. Gunter and J. Zhu. (2005). In *Y. Weiss, B. Schölkopf, and J. Platt, editors, Advances in Neural Information Processing Systems 18*, 483-490. MIT Press, Cambridge, MA.
7. B. Schölkopf, A. J. Smola. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. (Adaptive Computation and Machine Learning).* MIT Press.
8. B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. (2000). *Neural Computation*, 12:1207-1245.
9. I. Steinwart. (2003). *Journal of Machine Learning Research*, 4:1071-1105.
10. D.M.J. Tax and R.P.W. Duin. (1999) In *M. Verleysen editor, Proceedings ESANN*, 251-256, Brussels, D Facto.
11. V. Vapnik. (1995). *The Nature of Statistical Learning Theory.* Springer, New York, NY, USA.

The Iterative–Interpolation Approach to L_2 Model Reduction

Augusto Ferrante[1], Wiesław Krajewski[2], and Umberto Viaro[3]

[1] Dipartimento di Ingegneria dell'Informazione, Università di Padova, via Gradenigo 6/B, 35131 Padova, Italy
augusto@dei.unipd.it

[2] Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01447 Warsaw, Poland
krajewsk@ibspan.waw.pl

[3] Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica, Università di Udine, via delle Scienze 208, I-33100 Udine, Italy
viaro@uniud.it

To the memory of our friend Toni

Summary. This paper is concerned with the construction of reduced–order models of high–dimensional linear systems in such a way that the L_2 norm of the impulse–response error is minimized. Two convergent algorithms that draw on previous procedures presented by the same authors, are suggested: one refers to s–domain representations, the other to time–domain state–space representations. The algorithms are based on an iterative scheme that, at any step, satisfies certain interpolation constraints deriving from the optimality conditions. To make the algorithms suitable to the reduction of very large–scale systems, resort is made to Krylov subspaces and Arnoldi's method. The performance of the reduction algorithms is tested on two benchmark examples.

Keywords. Linear systems, model reduction, output–error minimization, L_2 norm, Krylov subspaces, Arnoldi's algorithm.

1 Introduction

Linear time–invariant models are often used to describe phenomena in physical and economic contexts because many tools are available for their study.

However, when the system complexity is high or its size big, the number of state variables may be too large for simulation and control purposes, and it is mandatory to approximate the original high–order model by means of a lower–order one.

Many approaches to model reduction have been proposed over the past decades. Among the most popular ones, we can mention the Padé-like methods leading to the retention of some Markov parameters and time moments [1], the aggregation techniques retaining some modes of the original system [2], the Hankel–norm optimization methods [3], the techniques based on principal component analysis and balanced realizations [4], and the methods aiming at the retention of suitable first–order information indices (Markov parameters or time moments) and second–order information indices (impulse–response powers) [5], [6].

Recently, the techniques based on moment matching have been tackled with the aid of numerically robust and reliable procedures such as the Arnoldi, Lanczos or rational Krylov methods (see, for example, [7], [8] and [9]) thus allowing the reduction of very large–scale systems. On the other hand, the techniques based on the minimization of an error norm still seem to be more appropriate to model reduction which is essentially an optimization problem.

This paper deals with the minimization of the L_2 norm of the output error, defined as the difference between the impulse responses of the original and reduced–order model. The L_2 norm has an appealing physical interpretation (power) which explains its wide–spread use (see, for example, [10], [11], [12] concerning the time domain, and [13], [14], [15] concerning the frequency domain).

Finding the L_2–optimal reduced–order model of a complex system is a hard task that often requires the solution of an ill–conditioned mathematical programming problem. Many of the available methods are therefore difficult to implement and rather inefficient; in particular, gradient techniques are not always satisfactory. Among the non–gradient approaches, the iterative–interpolation algorithm first suggested in [16] and further developed in [14] and [17] with reference to the *frequency domain* seems to be one of the most efficient procedures. This algorithm, however, is not always convergent. To overcome this difficulty, a variant of the method characterized by steps of shortened length has been proposed in [18]. The iterative–interpolation approach has recently been extended to *time–domain state–space* representations [19]. In this paper we pursue the ideas in [18] and [19] by suggesting two convergent algorithms in frequency and time domain, respectively

The next part of the paper is organized as follows. Section 2 formulates the L_2–optimal model–reduction problem and recalls the iterative interpolation method. Section 3 shows how convergence can be guaranteed and proposes two algorithms for models in either input–output or state–space form. These algorithms are applied in Section 4 to two benchmark examples that show the efficiency of the proposed techniques.

2 Problem statement and iterative–interpolation method

Let the state–space equations of the original full–order linear time–invariant *asymptotically–stable* system be

$$\dot{x}(t) = A\,x(t) + b\,u(t)\,, \tag{1}$$

$$y(t) = c\,x(t)\,, \tag{2}$$

where $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}$, $u(t) \in \mathbb{R}$, and n is the dimension of a minimal realization. The input–output behavior of system (1)–(2) is characterized by the transfer function

$$f(s) = \frac{n(s)}{d(s)} = c\,(s\,I - A)^{-1}\,b\,, \tag{3}$$

where deg $d(s) = n$ and deg $n(s) \leq n - 1$. Similarly, assume that the state–space equations of the reduced–order model are:

$$\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{b}u(t)\,, \tag{4}$$

$$\tilde{y}(t) = \tilde{c}\tilde{x}(t)\,, \tag{5}$$

where $\tilde{x}(t) \in \mathbb{R}^{\tilde{n}}$ with $\tilde{n} < n$, $\tilde{y}(t) \in \mathbb{R}$, $u(t) \in \mathbb{R}$. The related transfer function is

$$g(s) = \frac{m(s)}{c(s)} = \tilde{c}\,(s\,I - \tilde{A})^{-1}\,\tilde{b}\,, \tag{6}$$

where deg $c(s) = \tilde{n}$ and deg $m(s) \leq \tilde{n} - 1$.

In frequency domain, the optimal model–reduction problem can be formulated as follows: given an original transfer function $f(s)$ of order n, find a transfer function $g(s)$ of preassigned lower order $\tilde{n} < n$ in such a way that the (squared) H_2 norm of $e(s) := f(s) - g(s)$, that is,

$$\|e\|_2^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} e(j\omega)e^*(j\omega)d\omega\,, \tag{7}$$

is minimized (the star indicates complex conjugate).

The time–domain state–space version of this problem can be formulated as follows: given a state–space model (A, b, c) of minimal dimension n, find a model $(\tilde{A}, \tilde{b}, \tilde{c})$ of dimension $\tilde{n} < n$ so as to minimize the (squared) L_2 norm of the output error $e(t) := y(t) - \tilde{y}(t)$, that is,

$$\|e\|_2^2 = \int_{-\infty}^{\infty} e^2(t)\,d\,t\,, \tag{8}$$

where $y(t)$ and $\tilde{y}(t)$ are the responses of the full–order and reduced–order model to a unit–impulse input $u(t)$.

Assuming for notational simplicity that all of the poles of $g(s)$ are simple (but not necessarily real) and denoting them by $p_1, \ldots, p_{\tilde{n}}$, function $g(s)$ minimizes (7) only if the following *interpolation conditions* are satisfied [10]:

$$\left.\begin{aligned} f(-p_k) - g(-p_k) &= 0\,, \\ f'(-p_k) - g'(-p_k) &= 0\,, \end{aligned}\right\} \qquad k = 1, 2, \ldots, \tilde{n}\,. \tag{9}$$

These conditions can be expressed more compactly in terms of the numerator and denominator polynomials in (3) and (6) as

$$n(s)\,c(s) - m(s)\,d(s) = q(s)\,c^2(-s)\,, \tag{10}$$

where $q(s)$ is a polynomial whose degree is, at most, $n - \tilde{n} - 1$.

Equating the coefficients of the equal powers of s on both sides of the polynomial identity (10), we obtain $n + \tilde{n} - 1$ equations in the $2\,\tilde{n}$ unknown coefficients of $m(s)$ and $c(s)$ as well as in the $n - \tilde{n} - 1$ unknown coefficients of the “auxiliary” polynomial $q(s)$. These equations can be solved by means of an iterative procedure based on the recurrence relation

$$n(s)\,c^{(h+1)}(s) - m^{(h+1)}\,(s)d(s) \;= q^{(h+1)}(s)\,[c^{(h)}(-s)]^2\,, \tag{11}$$

which determines polynomials $c^{(h+1)}(s)$ and $m^{(h+1)}$ from the polynomial $c^{(h)}(s)$ previously computed.

The resulting algorithm can be presented as follows:

Algorithm 1

1. Make a guess of the initial reduced–order denominator polynomial $c^{(0)}(s)$ and set $h = 0$.
2. Given $c^{(h)}(s)$, evaluate $c^{(h+1)}(s)$, $m^{(h+1)}(s)$ and $q^{(h+1)}(s)$ on the basis of (11).
3. If the stopping criterion is satisfied, form the reduced–order denominator and numerator as $c(s) = c^{(h+1)}(s)$ and $m(s) = m^{(h+1)}(s)$, respectively. Otherwise, set $h = h + 1$ and return to phase 2.

Phase 2 requires the solution of a set of linear equations. The corresponding system matrix is sparse with a structure that guarantees the efficiency of the procedure whose details are illustrated in [14]. Equation (11) can be given an interesting interpretation; precisely, it states that function $g^{(h+1)}(s)$ must interpolate the original transfer function $f(s)$, with intersection number 2, at the opposites of the poles $p_1^{(h)}, \ldots, p_{\tilde{n}}^{(h)}$ of $g^{(h)}(s)$, that is,

$$\left.\begin{aligned} f(-p_k^{(h)}) - g^{(h+1)}(-p_k^{(h)}) &= 0\,, \\ f'(-p_k^{(h)}) - (g^{(h+1)})'(-p_k^{(h)}) &= 0\,, \end{aligned}\right\} \qquad k = 1, 2, \ldots, \tilde{n}\,. \tag{12}$$

In other words, the numerator of the current error function $e^{(h+1)}(s) := f(s) - g^{(h+1)}(s)$ must contain the factor $[c^{(h)}(-s)]^2$. This is the reason why the above procedure has been called *iterative–interpolation* algorithm.

The idea of iterative interpolation has been recently extended to state–space representations such as (4) and (5) in [19] where the interpolation problem arising at each iteration is solved by means of Krylov–subspace methods. To this purpose, denoting by $p_k^{(h)}$ the $\tilde{n}$ eigenvalues $\lambda_k(\tilde{A}^{(h)})$, $k = 1, 2, \ldots, \tilde{n}$, of the reduced–order system matrix $\tilde{A}^{(h)}$ computed in the preceding iteration, two matrices V and Z are formed in the current iteration such that

$$Im(V) = span\{(-p_1^{(h)} I - A)^{-1} b, \ldots, (-p_{\tilde{n}}^{(h)} I - A)^{-1} b)\} \ , \qquad (13)$$

$$Im(Z) = span\{(-p_1^{(h)} I - A^T)^{-1} c^T, \ldots, (-p_{\tilde{n}}^{(h)} I - A^T)^{-1} c^T)\} \qquad (14)$$

and $Z^T V = I$. In this way the reduced–order model (4) and (5) with

$$\tilde{A} = \tilde{A}^{(h+1)} := Z^T A\, V \ , \ \tilde{b} = \tilde{b}^{(h+1)} := Z^T b \ , \ \tilde{c} = \tilde{c}^{(h+1)} := c\, V \qquad (15)$$

satisfies the following interpolation conditions analogous to (12) [9], [20]:

$$\left.\begin{array}{l} c(-p_k^{(h)} I - A)^{-1} b = \tilde{c}^{(h+1)}(-p_k^{(h)} I - \tilde{A}^{(h+1)})^{-1} \tilde{b}^{(h+1)} \\ c(-p_k^{(h)} I - A)^{-2} b = \tilde{c}^{(h+1)}(-p_k^{(h)} I - \tilde{A}^{(h+1)})^{-2} \tilde{b}^{(h+1)} \end{array}\right\} \quad k = 1, 2, \ldots, \tilde{n} \, . \qquad (16)$$

The corresponding algorithm can be presented as follows:

Algorithm 2

1. Make a guess of the eigenvalues $p_1^{(0)}, \ldots, p_{\tilde{n}}^{(0)}$ of the initial reduced–order system matrix $\tilde{A}^{(0)}$ (often, the $\tilde{n}$ dominant eigenvalues of A represent a good starting point) and set $h = 0$.
2. Choose V and Z such that
   $Im(V) = span\,\{(-p_1^{(h)} I - A)^{-1}\, b, \ldots, (-p_{\tilde{n}}^{(h)} I - A)^{-1}\, b\}$,
   $Im(Z) = span\,\{(-p_1^{(h)} I - A^T)^{-1}\, c^T, \ldots, (-p_{\tilde{n}}^{(h)} I - A^T)^{-1}\, c^T\}$,
   $Z = Z\,(Z^T\, V)^{-T}$.
3. Compute $\tilde{A}^{(h+1)} = Z^T\, A\, V$.
4. If the stopping criterion is satisfied, form the reduced–order model matrices as $\tilde{A} = \tilde{A}^{(h+1)}$, $\tilde{b} = Z^T\, b$ and $\tilde{c} = c\, V$. Otherwise, let $p_k^{(h+1)} = \lambda_k(\tilde{A}^{(h+1)})$, $k = 1, \ldots, \tilde{n}$, set $h = h + 1$ and return to phase 2.

To find V and Z, the rational Arnoldi method can conveniently be applied [19], which makes the above algorithm suitable to the reduction of very large–scale systems.

Algorithms 1 and 2 have proven to be efficient [14], [19] but are not always convergent. The next section shows how these algorithm can be modified to ensure convergence.

3 Convergent variants of the iterative–interpolation algorithms

Denoting by $\mathbf{c}^{(i)}$ the vector formed from the coefficients of $s^{\tilde{n}-1}$, $s^{\tilde{n}-2}$, $\ldots$, s^0, in every *monic* polynomial $c^{(i)}(s)$ generated by (11), recursion (11) can be reformulated as

$$\mathbf{c}^{(h+1)} = \Phi(\mathbf{c}^{(h)})\,, \tag{17}$$

where $\Phi : \mathbb{R}^{\tilde{n}} \to \mathbb{R}^{\tilde{n}}$ is a continuously differentiable function. Therefore, the reduced–order model minimizing (7) corresponds to a fixed point $\hat{\mathbf{c}}$ of Φ:

$$\hat{\mathbf{c}} = \Phi(\hat{\mathbf{c}})\,. \tag{18}$$

It has been shown in [14] and [18] that:
(i) the eigenvalues of the Jacobian $\frac{\partial \Phi}{\partial \mathbf{c}}$ at every fixed point $\hat{\mathbf{c}}$ are real,
(ii) at a fixed point $\hat{\mathbf{c}}$ corresponding to a saddle point of the objective function at least one eigenvalue of the Jacobian is greater than 1, and
(iii) at a fixed point corresponding to a minimum of the objective function every eigenvalue of the Jacobian is less than 1 but not necessarily greater than -1, which explains why the iterative–interpolation algorithm (17) is not always convergent.

To avoid this difficulty, resort can be made to Newton–like fixed–point algorithms [21], [22] which are, however, computationally demanding. The approximation problem can be solved more efficiently by replacing Φ in (17) by another continuously differentiable function with the following properties:
(i) it has the same fixed points $\hat{\mathbf{c}}$ as Φ,
(ii) all the eigenvalues of its Jacobian at every $\hat{\mathbf{c}}$ have magnitude less than 1, and
(iii) it can be obtained *easily* from Φ with no *a priori* information about $\hat{\mathbf{c}}$.

These properties are exhibited, for a suitably small value of the real parameter α, by the function $\Phi_\alpha : \mathbb{R}^{\tilde{n}} \to \mathbb{R}^{\tilde{n}}$ obtained from Φ according to

$$\Phi_\alpha(\mathbf{c}) = \alpha\Phi(\mathbf{c}) + (1-\alpha)\mathbf{c}\,, \tag{19}$$

Therefore, the iterative procedure (17) can be replaced by

$$\bar{\mathbf{c}}^{(h+1)} = \Phi_\alpha(\bar{\mathbf{c}}^{(h)})\,, \tag{20}$$

Starting from $\bar{\mathbf{c}}^{(h)} = \mathbf{c}^{(h)}$, recursion (20) determines vector $\bar{\mathbf{c}}^{(h+1)}$ as a linear combination of $\mathbf{c}^{(h)}$ and $\mathbf{c}^{(h+1)} = \Phi(\mathbf{c}^{(h)})$ according to the combination coefficients $1-\alpha$ and α, respectively.
Denoting by λ_m the smallest eigenvalue of $\frac{\partial \Phi}{\partial \mathbf{c}}(\hat{\mathbf{c}})$, if

$$0 < \alpha < \alpha_m = \frac{2}{1-\lambda_m}\,, \tag{21}$$

then all the eigenvalues of $\frac{\partial \Phi_\alpha}{\partial \mathbf{c}}(\hat{\mathbf{c}})$ have magnitude less than 1 [18], and the sequence of vectors $\{\bar{\mathbf{c}}^{(i)}\}$ generated by (20) converges to $\hat{\mathbf{c}}$ from a suitable neighbourhood of $\hat{\mathbf{c}}$.

From the above arguments the following algorithm is obtained:

Algorithm 3

1. Select $\alpha \in (0,1)$, make a guess of the initial coefficient vector $\bar{\mathbf{c}}^{(0)}$, and set $h = 0$.
2. Let $\mathbf{c}^{(h)} = \bar{\mathbf{c}}^{(h)}$ and compute $\mathbf{c}^{(h+1)} = \Phi(\mathbf{c}^{(h)})$ on the basis of (11) as in Algorithm 1.
3. Form $\bar{\mathbf{c}}^{(h+1)} = \alpha\, \mathbf{c}^{(h+1)} + (1-\alpha)\, \bar{\mathbf{c}}^{(h)}$.
4. If the stopping criterion is satisfied, form the reduced–order denominator and numerator as $c(s) = c^{(h+1)}(s)$ and $m(s) = m^{(h+1)}(s)$, respectively. Otherwise, set $h = h+1$ and return to phase 2.

Therefore the steps made by Algorithm 3 at each iteration are α times shorter than the steps made by Algorithm 1.

A similar approach can be followed to ensure the convergence of a variant of the time–domain state–space iterative–interpolation procedure. It suffices to introduce a few additional operations into Algorithm 2 for: (i) computing the characteristic polynomial $c^{(h+1)}(s)$ associated with $\tilde{A}^{(h+1)}$, (ii) shortening the step made at every iteration by a factor α, and (iii) evaluating the roots $\bar{p}_1, \ldots, \bar{p}_{\tilde{n}}$ of the polynomial $\bar{c}^{(h+1)}(s)$ formed from $\bar{\mathbf{c}}^{(h+1)}$.

The resulting algorithm is outlined next:

Algorithm 4

1. Select $\alpha \in (0,1)$, and form the coefficient vector $\bar{\mathbf{c}}^{(0)}$ by choosing the roots $\bar{p}_1^{(0)}, \ldots, \bar{p}_{\tilde{n}}^{(0)}$ of the characteristic polynomial $\bar{c}^{(0)}(s)$ associated with the initial reduced–order system matrix $\tilde{A}^{(0)}$. Set $h = 0$.
2. Choose V and Z such that
   $Im(V) = span\,\{(-\bar{p}_1^{(h)} I - A)^{-1}\, b, \ldots, (-\bar{p}_{\tilde{n}}^{(h)} I - A)^{-1}\, b\}$,
   $Im(Z) = span\,\{(-\bar{p}_1^{(h)} I - A^T)^{-1}\, c^T, \ldots, (-\bar{p}_{\tilde{n}}^{(h)} I - A^T)^{-1}\, c^T\}$,
   $Z = Z\,(Z^T V)^{-T}$.
3. Compute $\tilde{A}^{(h+1)} = Z^T A\, V$.
4. If the stopping criterion is satisfied, form the reduced–order model matrices as $\tilde{A} = \tilde{A}^{(h+1)}$, $\tilde{b} = Z^T b$ and $\tilde{c} = c\,V$. Otherwise:
   (i) find the characteristic polynomial $c^{(h+1)}(s)$ associated with $\tilde{A}^{(h+1)}$ and form the related coefficient vector $\mathbf{c}^{(h+1)}$,
   (ii) compute $\bar{\mathbf{c}}^{(h+1)} = \alpha\, \mathbf{c}^{(h+1)} + (1-\alpha)\, \bar{\mathbf{c}}^{(h)}$ and form the related polynomial $\bar{c}^{(h+1)}(s)$,
   (iii) evaluate the roots $\bar{p}_k^{(h+1)}$, $k = 1, \ldots, \tilde{n}$, of $\bar{c}^{(h+1)}(s)$, and
   (iv) set $h = h+1$ and return to phase 2.

Shortening the steps by α ensures the convergence of Algorithms 3 and 4 also when Algorithms 1 and 2 do not converge. However, if the latter algorithms converge, which is not known *a priori*, their modified versions may require a larger number of iterations to find a (locally) optimal solution, as shown by the examples in the next section.

4 Examples

In the following, the modified iterative–interpolation approach is applied to two meaningful examples. The purpose of this section is threefold, that is: (i) to show that the modified algorithms can find an optimal solution also when the original iterative–interpolation algorithms fail, (ii) to see how the convergence of the modified algorithms depends on α, and (iii) to compare the speed of the modified algorithms with that of the original ones when the latter converge.

A Model of the ACES structure

The state–space model describing the dynamic relation between a torque actuator and an approximately collocated torsional rate sensor for the ACES (Active Control Technique Evaluation for Spacecraft) structure at NASA Marshall Space Flight Center has dimension 17 [23]. The nonzero entries of the matrix A are:

$$\begin{array}{lll}
a_{1,1} = -0.031978272 & a_{1,2} = -78.54 & a_{1,17} = 0.0097138566 \\
a_{2,1} = 78.54 & a_{2,2} = -0.031978272 & a_{2,17} = -0.0060463517 \\
a_{3,3} = -5.152212 & a_{3,4} = -51.457677 & a_{3,17} = -0.021760771 \\
a_{4,3} = 51.457677 & a_{4,4} = -5.152212 & a_{4,17} = 0.0054538246 \\
a_{5,5} = -0.1351159 & a_{5,6} = -15.417859 & a_{5,17} = -0.02179972 \\
a_{6,5} = 15.417859 & a_{6,6} = -0.1351159 & a_{6,17} = -0.015063913 \\
a_{7,7} = -0.42811443 & a_{7,8} = -14.698408 & a_{7,17} = 0.01042631 \\
a_{8,7} = 14.698408 & a_{8,8} = -0.42811443 & a_{8,17} = 0.0088479697 \\
a_{9,9} = -0.064896745 & a_{9,10} = -12.077045 & a_{9,17} = -0.030531575 \\
a_{10,9} = 12.077045 & a_{10,10} = -0.064896745 & a_{10,17} = -0.030260987 \\
a_{11,11} = -0.048520356 & a_{11,12} = -8.9654448 & a_{11,17} = -0.016843335 \\
a_{12,11} = 8.9654448 & a_{12,12} = -0.048520356 & a_{12,17} = -0.011449591 \\
a_{13,13} = -0.036781718 & a_{13,14} = -4.9057426 & a_{13,17} = -0.1248007 \\
a_{14,13} = 4.9057426 & a_{14,14} = -0.036781718 & a_{14,17} = 0.0005136047 \\
a_{15,15} = -0.025112482 & a_{15,16} = -3.8432892 & a_{15,17} = -0.035415526 \\
a_{16,15} = 3.8432892 & a_{16,16} = -0.025112482 & a_{16,17} = -0.028115589 \\
 & & a_{17,17} = -92.399784\,,
\end{array}$$

The vectors b and c are given by:

$$b = \begin{bmatrix} 1.8631111 \\ -1.1413786 \\ -1.2105758 \\ 0.31424169 \\ 0.31424169 \\ -0.211128913 \\ 0.19552894 \\ -0.037391511 \\ -0.01049736 \\ -0.011486242 \\ -0.029376402 \\ 0.0082391613 \\ -0.012609562 \\ -0.0022040505 \\ -0.030853234 \\ 0.0011671662 \\ 0 \end{bmatrix}, \quad c^T = \begin{bmatrix} -0.0097138566 \\ 0.0060463517 \\ 0.021760771 \\ -0.0054538246 \\ 0.02179972 \\ 0.015063913 \\ -0.01042631 \\ -0.0088479697 \\ 0.030531575 \\ 0.030260987 \\ 0.016843335 \\ 0.011449591 \\ 0.1248007 \\ -0.0005136047 \\ 0.035415526 \\ 0.028115589 \\ 184.79957 \end{bmatrix}.$$

Matrix A has one real eigenvalue at -92.399784 and eight pairs of complex conjugate eigenvalues at $-0.03198 \pm \jmath 78.54$, $-5.152 \pm \jmath 51.457$, $-0.135 \pm \jmath 15.418$, $-0.428 \pm \jmath 14.698$, $-0.065 \pm \jmath 12.077$, $-0.0485 \pm \jmath 8.965$, $-0.037 \pm \jmath 4.906$, $-0.025 \pm \jmath 3.843$.

As is often the case in L_2 model reduction, the objective functional has different local minima. Starting from a 6th–order system matrix $\tilde{A}^{(0)}$ with eigenvalues at -1, -2, -3, -4, -5 and -6, Algorithm 2 finds a locally optimal 6th–order model $(\tilde{A}, \tilde{b}, \tilde{c})$ in 11 iterations. The resulting matrix $\tilde{A}$ has three pairs of complex conjugate eigenvalues at $-0.0320 \pm \jmath 78.5402$, $-0.0247 \pm \jmath 3.8429$ and $-0.0368 \pm \jmath 4.9042$; the square norm of the related output error is $\|e\|_2^2 = 0.0094$. If, instead, the eigenvalues of the initial reduced–order matrix $\tilde{A}^{(0)}$ are -10, -20, -30, -40, -50 and -60, Algorithm 2 finds in 10 iterations a different locally optimal 6th–order solution $(\tilde{A}, \tilde{b}, \tilde{c})$ characterized by a square norm of the output error equal to $\|e\|_2^2 = 0.0065$. In this case matrix $\tilde{A}$ has three pairs of complex conjugate eigenvalues at $-0.0320 \pm \jmath 78.5400$, $-0.1061 \pm \jmath 15.4379$ and $-5.3888 \pm \jmath 52.3342$. Similar results are obtained by applying Algorithm 1 to models in transfer–function form.

Finding a reduced–order model of dimension 7 is harder. If the eigenvalues of the 7th–order $\tilde{A}^{(0)}$ are -10, -20, -30, -40, -50, -60 and -70, Algorithm 2 finds a locally optimal 7th–order solution $(\tilde{A}, \tilde{b}, \tilde{c})$ in 163 iterations. The resulting matrix $\tilde{A}$ has three pairs of complex conjugate eigenvalues at $-0.0320 \pm \jmath 78.5400$, $-0.1091 \pm \jmath 15.4289$, $-4.8710 \pm \jmath 51.2334$ and one real eigenvalue at -11.1231. The square norm of the related output error is $\|e\|_2^2 = 0.0063$.

Algorithm 2, however, can not find a solution if the eigenvalues of the initial 7th–order matrix $\tilde{A}^{(0)}$ are -1, -2, -3, -4, -5, -6 and -7. Instead, starting from this initial guess Algorithm 4 finds three different locally optimal 7th–

order models $(\tilde{A}, \tilde{b}, \tilde{c})$ depending on the value of α. Precisely, for $\alpha = 0.8$ and $\alpha = 0.7$ the resulting optimal matrix $\tilde{A}$ has three pairs of complex conjugate eigenvalues at $-0.0319 \pm \jmath 78.5401$, $-0.0230 \pm \jmath 3.8441$, $-0.0330 \pm \jmath 4.9037$ and one real eigenvalue at -76.2562; this solution is reached in 61 iterations for $\alpha = 0.8$ and 63 for $\alpha = 0.7$. The value of the square error norm for this first solution is 0.0086. For $\alpha = 0.6$, $\alpha = 0.5$ and $\alpha = 0.4$ the locally optimal matrix $\tilde{A}$ has three pairs of complex conjugate eigenvalues at $-0.0320 \pm \jmath 78.5402$, $-0.0247 \pm \jmath 3.8429$, $-0.0368 \pm \jmath 4.9043$ and one real eigenvalue at -0.2860; this solution is reached in 29 iterations for $\alpha = 0.6$, 39 for $\alpha = 0.5$ and 50 for $\alpha = 0.4$. The value of the square error norm for this second solution is 0.0094. For $\alpha = 0.3$, $\alpha = 0.2$ and $\alpha = 0.1$ the locally optimal matrix $\tilde{A}$ has three pairs of complex conjugate eigenvalues at $-0.8.2380 \pm \jmath 47.1055$, $-0.0340 \pm \jmath 4.9038$, $-0.0234 \pm \jmath 3.8437$ and one real eigenvalue at -58.6214; this solution is reached in 43 iterations for $\alpha = 0.3$, 60 for $\alpha = 0.2$ and 114 for $\alpha = 0.1$. The value of the square error norm for this third solution is 0.0701.

B Model of a CD player

The full–order state–space model describing the dynamic relation between the lens actuator and the radial–arm position of a portable CD player is characterized by 120 state variables [24]. Since all of the poles of this system are complex (sixty complex conjugate pairs), it is difficult to find an optimal approximant of odd order. For example, starting from the (arbitrary) initial guess $p_1^0 = -0.5$, $p_2^0 = -1$, $p_3^0 = -2$, $p_4^0 = -3$, $p_5^0 = -4$, $p_6^0 = -5$, $p_7^0 = -6$, $p_8^0 = -7$, $p_9^0 = -8$, Algorithm 2 can not find a 9th–order model. Instead, Algorithm 4 with $\alpha = 0.1$ finds an optimal solution of order 9 in 55 iterations starting from the same initial guess. The square norm of the related output error is $\|e\|_2^2 = 30.2335$. Note, by comparison, that the 9th–order model obtained according to the popular TBR (Truncated Balanced Realization) method [4] is characterized by $\|e\|_2^2 = 35.149$. With $\alpha = 0.5$ Algorithm 4 arrives at the optimal solution in only 16 iterations.

Algorithm 2 finds an optimal 10th–order approximant in 7 iterations starting from $p_1^0 = -0.5$, $p_2^0 = -1$, $p_3^0 = -2$, $p_4^0 = -3$, $p_5^0 = -4$, $p_6^0 = -5$, $p_7^0 = -6$, $p_8^0 = -7$, $p_9^0 = -8$, $p_{10}^0 = -10$. The corresponding square error norm is $\|e\|_2^2 = 21.04$. As is expected, Algorithm 4 requires a larger number of iterations to arrive at the same solution. Table 1 shows how this number depends on α.

Table 1. Number of iterations vs. α for the optimal 10th–order approximant

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
number of iterations	48	30	22	22	21	17	15	12	9	7

5 Conclusions

The iterative–interpolation algorithms for L_2–optimal model reduction have proven to be very efficient compared to alternative procedures with the same objective. In Section 3, two algorithms that ensure convergence to an optimal solution have been presented . Essentially, this result is obtained by shortening the steps made at every iteration by the original algorithms outlined in Section 2. Two benchmark examples have been worked out in Section 4 to show how the algorithms perform and how the speed of convergence depends on the step size.

References

1. Bultheel A, Van Barel M (1986) *J Comp Appl Math*, 14:401–438
2. Hickin J, Sinha NK (1980) *IEEE Trans Automat Contr*, 25:1121–1127
3. Glover K (1984) *Int J Control*, 39:1115–1193
4. Moore BC (1981) *IEEE Trans Automat Contr*, 26:17–32
5. de Villemagne C, Skelton RE (1987) *Int J Control*, 46: 2141–2169
6. Krajewski W, Viaro U (2007) *Numerical Algorithms*, 44:83–98
7. Gallivan K, Grimme EJ, Van Dooren P (1996) *Numerical Algorithms*, 12:33–63
8. Datta BN (2003) *Future Generation Computer Systems*, 9:1253–1263
9. Antoulas AC (2005) *Approximation of Large–Scale Dynamical Systems*. SIAM Book Series: Advances in Design and Control, Philadelphia
10. Meier L, Luenberger DG (1967) *IEEE Trans Automat Contr*, 12:585–588
11. Wilson DA (1974) *Int J Control*, 20:57–64
12. Hyland DC, Bernstein DS (1985) *IEEE Trans Automat Contr*, 30:1201–1211
13. Spanos JT, Milman MH, Mingori DL (1992) *Automatica*, 28:897–909
14. Krajewski W, Lepschy A, Redivo-Zaglia M, Viaro U (1995) *Numerical Algorithms*, 9:355–377
15. Fulcheri P, Olivi M (1998) *SIAM J Control Optim*, 36:2103–2127
16. Lepschy A, Mian GA, Pinato G, Viaro U (1991) *Proc 30th Conf. Decision and Control*. Brighton England, 1991, pp. 2321–2324.
17. Ferrante A, Krajewski W, Lepschy A, Viaro U (1998) In: *Banka S, Domek S, Emirsajlow Z (eds) Proc Int Symp Math Models in Automation and Robotics.* Miedzyzdroje, Poland
18. Ferrante A, Krajewski W, Lepschy A, Viaro U (1999) *Automatica*, 35:75–79
19. Gugercin S, Antoulas AC, Beattie CA (2006) A rational Krylov iteration for optimal H_2 model reduction. In: *Proc 17th Int Symp Math Theory Networks and Systems.* Kyoto
20. Grimme EJ (1997) Krylov projection methods for model reduction. PhD Thesis, ECE Department, University of Illinois, Urbana–Champaign
21. Ferrante A, Lepschy A, Viaro U (2001) *It. J. Pure Appl Math*, 9:179–186
22. Ferrante A, Lepschy A, Viaro U (2001) *It. J. Pure Appl Math*, 10:47–54
23. Zigic D, Watson LT, Collins EG, Bernstein DS (1992) *Int J Control*, 56:173–191
24. Chahlaoui Y, Van Dooren P (2002) A collection of benchmark examples for model reduction of linear time–invariant dynamical systems. SLICOT Working Note

The Frisch Scheme in Algebraic and Dynamic Identification Problems

Roberto Guidorzi, Roberto Diversi and Umberto Soverini

Dipartimento di Elettronica, Informatica e Sistemistica, Università di Bologna
Viale del Risorgimento 2, 40136 Bologna, Italy
rguidorzi@deis.unibo.it, rdiversi@deis.unibo.it,
usoverini@deis.unibo.it

1 Introduction

The search for connections between observations ("laws of nature") is at the basis of the development of scientific knowledge and can be traced back at least some thousand years as shown, for instance, by the large amount of clay tablets concerning the so called astronomical diaries compiled by the Mesopotamian astronomers. This search for knowledge is characterized by two basic steps, the necessity of performing a quantification of observations, i.e. the transformation of observations into numerical entities and the subsequent extraction of relations between the obtained values, to be used for interpretation, prediction, control or other purposes. If we observe n different variables

$$x_1,\ x_2,\ \ldots,\ x_n \tag{1}$$

and denote with

$$x_{1i},\ x_{2i},\ \ldots,\ x_{ni} \tag{2}$$

the values that these variables assume at the i-th observation, the search for a law describing the behavior of the process that has generated the observations is the search for a mathematical relation

$$f(x_1, x_2, \ldots, x_n) = 0 \tag{3}$$

satisfied by every set of observations, i.e. such that, for every i,

$$f(x_{1i}, x_{2i}, \ldots, x_{ni}) = 0. \tag{4}$$

Even assuming that the considered process is actually governed by a law of the type (3), the observations will never satisfy, in all practical situations, relation (4) because of the errors that will be inevitably introduced during the quantification step (e.g. noise in analog systems, finite number of possible values and noise in digital environments etc.). The deduction of the law behind the observations is thus a problem that does not admit any solution because the observations will not satisfy, in general, any relation. All procedures leading to the extraction of abstract relations from real data rely, in fact, on modified observations. This modification process should be carried out on the basis of the exact knowledge of the nature of the errors since it affects, in a very substantial way, the final result.

As a simple example to illustrate this point, let us assume the existence of a linear relation, described by the scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$, linking the variables (1):

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0. \tag{5}$$

The same relation can be described also, when $\alpha_i \neq 0$, by means of the equivalent, asymmetric relation

$$x_i = \beta_1\, x_1 + \cdots + \beta_{i-1}\, x_{i-1} + \beta_{i+1}\, x_{i+1} + \cdots + \beta_n\, x_n \tag{6}$$

where $\beta_j = -\alpha_j/\alpha_i$. Let us assume also that only one of the observations, x_{ki} is affected by zero–mean additive errors. By denoting with $\hat{x}_{ji}$ the true values, with $\tilde{x}_{ji}$ the observation errors and with x_{ji} the actual observations, these quantities will be linked by the relations

$$x_{ji} = \hat{x}_{ji} \quad \text{for } j \neq k \tag{7}$$

$$x_{ki} = \hat{x}_{ki} + \tilde{x}_{ki}. \tag{8}$$

This is the well known context of the least squares and we can easily find the optimal and asymptotically unbiased solution (Gauss–Markov theorem) by means of the least squares algorithm. Note, however, that this can be done if and only if we actually know *which* variable is affected by observation errors since, from a geometrical point of view, we perform an orthogonal projection of the vector of the N noisy observations $[x_{k1}\, x_{k2}\, \ldots\, x_{kN}]^T$ on the (hyper)plane defined by the vectors of noiseless observations $[x_{j1}\, x_{j2}\, \ldots\, x_{jN}]^T$ $(j \neq k)$ and *substitute* $[x_{k1}\, x_{k2}\, \ldots\, x_{kN}]^T$ with its orthogonal projection. If we remain in this context (a single observation is affected by errors) but with no *a priori* information about which of the variables is noisy, the problem cannot be solved. In fact it is impossible to select the correct solution among the n possible ones that can be obtained by considering as affected by errors in turn x_1, x_2, …, x_n. Moreover, if the errors are not restricted to a single variable, none of these n solutions will be, even asymptotically, correct.

This elementary example outlines two basic points:

1) The extraction of models from data affected by errors requires, as intermediate step, the deduction of new data from the available ones. The model will then be deduced from these new data, not from the original ones.

2) In absence of precise information on the errors it is possible to formulate different assumptions, each leading, in general, to extract different sets of data from the observations and, consequently, to different models.

Every systematic procedure to deduce a model (or a family of models) from data affected by errors is defined as a "scheme" [1, 2]. Many different schemes have been investigated and described in the literature. None of these schemes can be considered, *per se*, as superior to any other; what changes is simply the set of assumptions. As a consequence, the results that can be obtained by applying different schemes to the same set of data depend essentially on the "distance" between the assumptions behind the scheme and the actual situation; all other claims are related more to faith than to science. A very complete analysis of the assumptions behind different schemes has been proposed by Kalman in [1, 2, 3].

The content of this paper falls inside the Errors-in-Variables (EIV) context that assumes the presence of additive noise on all variables. This is a challenging environment that has seen an increasing amount of research only during the last decades. One of the appealing features of the models that can be deduced in EIV contexts concerns their intrinsic capability of relying on a limited set of *a–priori* assumptions. This feature suggests the use of EIV models in applications like, for instance, diagnosis, where the interest is focused on a realistic description of a process rather than on other aspects, like prediction. For a complete overview on EIV identification see [4] and the references therein.

This paper concerns the scheme proposed by the Nobel prize Ragnar Frisch in 1934 [5] and its application to the problem of deducing linear relations from noisy observations concerning both algebraic processes (this term is used here to denote static processes that can be described by sets of relations linking measures performed at the same time) and dynamic processes (i.e. processes described by difference equations). The Frisch scheme is an interesting compromise between the great generality of the EIV environment and the possibility of performing real applications. Moreover, the Frisch scheme encompasses some other important schemes like, for instance, Least Squares and the Eigenvector Method and plays, consequently, an important role also from a conceptual point of view.

The Frisch scheme does not lead, at least in the algebraic case, to a single solution but to a whole family of solutions compatible with a given set of noisy observations. This fact has often diverted the attention towards simpler schemes leading to a single solution so that the smart environment proposed by Frisch has not received, for many decades, the attention that it deserved.

As will be shown in the following, the analysis of the Frisch scheme leads to two separate loci of solutions, one in the parameter space and another in the space of the noise variances; of course the points of these loci are linked by well defined relations. Some fundamental results [1, 3] describe these maps as well as the shape of the loci in the parameter space under specific conditions. Unfortunately, however, the locus of solutions in the parameter space can be easily defined only when the data are compatible with a single linear relation; in all other cases the performed analyses have evidentiated the extremely complex structure of this locus, that prevents its practical use [6]. On the contrary, the investigation of the properties of the locus of solutions in the noise space has offered a key for a deeper analysis that shows that this locus does never degenerate and enjoys some consistent properties [7, 8].

A problem of great importance in the econometric field consists in determining the maximal number of linear relations compatible with a given set of noisy data. The importance attributed to this problem is due to the fact that its solution is considered as linked to the extraction of the maximal information from the data [9]. The solution of this problem in the context of the Frisch scheme has been possible only by making reference to the properties of the locus of solutions in the noise space [10]; other approaches have led to determine an upper bound for this number [11].

When the data are generated by a linear time–invariant dynamic process and the Frisch context is used for its identification, it is necessary to consider the loci of solutions under the constraints imposed by the time shift properties of dynamic systems [12]. It can be surprising to discover that, in this respect, the dynamic case can be seen as a subcase of the algebraic one and that the shift properties of dynamic systems lead (in general) to a unique solution [13, 14, 15]. Moreover this solution is linked to the solution of the maximal corank problem in the algebraic case.

All previous statements are true when the assumptions behind the Frisch scheme are exactly fullfilled and this can be assumed only in asymptotic conditions. In all practical cases this cannot be achieved not only because real data sets are necessarily limited but also because of a whole series of violations due to non linearity, non stationarity etc. Frisch identification procedures require thus the introduction of suitable criteria [16, 17, 18].

The Frisch scheme in the identification of dynamic processes enjoys some peculiarities that make it particularly suitable for the solution of specific problems like filtering and fault detection and isolation [19, 20, 21].

The purpose of this paper is to outline some results obtained in the analysis of the Frisch scheme in its original algebraic context and, in particular, to discuss the solution of the problem of determining the maximal number of linear relations compatible with a given set of noisy data. Other relevant topics of this analysis concern the properties of the loci of the Frisch solutions in the noise and parameter spaces as well as the possibility of obtaining single solutions. This discussion is carried out in Section 2. The second part of the paper regards the extension of the original algebraic environment to the dynamic one, where it is shown how the Frisch scheme can lead, differently from the algebraic case, to a single solution. This topic and the associated algorithms required for the application to real processes are discussed in Section 3, first in the SISO case and then in the MIMO one. Section 4 shows how mapping some classical problems like the blind identification of FIR transmission channels and the identification of noisy autoregressive models into a dynamical Frisch context can extend the limits of previous approaches. Some concluding remarks are finally reported in Section 5.

2 The Frisch scheme in the algebraic case

A Estimating linear relations from noisy data: statement of the problem

Consider the linear algebraic process (5); by denoting with X the $N \times n$ matrix whose rows contain the N observations (2)

$$X = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1N} & x_{2N} & \dots & x_{nN} \end{bmatrix}, \tag{9}$$

relation (5) can be written in the form

$$X\,A = 0 \tag{10}$$

where

$$A = \big[\, \alpha_1,\, \alpha_2,\, \dots,\, \alpha_n \,\big]^T. \tag{11}$$

In the more general case of q linear relations between the variables, A will be a $(n \times q)$ matrix with columns given by the q sets of coefficients describing the $q = n - \operatorname{rank} X$ independent linear relations linking the data. Relation (10) can be rewritten also by substituting X with

$$\Sigma = \frac{X^T X}{N} \tag{12}$$

i.e., under the assumption of null mean value of the variables, with the sample covariance matrix of the data. In absence of noise and in presence of linear relations Σ will be singular and positive semidefinite

$$\Sigma \geq 0. \tag{13}$$

Every solution, A, with maximal rank, of the equation

$$\Sigma A = 0 \tag{14}$$

is a basis of $\ker \Sigma$. When the data are corrupted by noise, $\operatorname{rank} X = n$, no linear relations are compatible with the observations and Σ is positive definite

$$\Sigma > 0. \tag{15}$$

In situations of this kind, linear relations can be extracted only by modifying X or Σ, i.e. the data.

B Assumptions behind estimation schemes

If no assumptions are introduced, *any* set of noisy data is compatible with *any* solution. The assumptions usually introduced on the errors (noise) to restrict the number of admissible solutions are the following:

1) The noise is additive; every observation is the sum of an unknown exact part $\hat{x}_i$, and of a noise term $\tilde{x}_i$:

$$x_i = \hat{x}_i + \tilde{x}_i \tag{16}$$

2) The mean value of $\hat{x}_i$ and $\tilde{x}_i$ is null:

$$\sum_{t=1}^{N} \hat{x}_{it} = 0, \qquad\qquad \sum_{t=1}^{N} \tilde{x}_{it} = 0. \tag{17}$$

3) The sequences of noise samples are orthogonal to the sequences of noiseless variables:

$$\sum_{t=1}^{N} \tilde{x}_{it}\, \hat{x}_{jt} = 0 \quad \forall\; i, j. \tag{18}$$

Under these assumptions:

$$X = \hat{X} + \tilde{X} \tag{19}$$

$$\hat{X}^T \tilde{X} = 0 \tag{20}$$

$$\Sigma = \hat{\Sigma} + \tilde{\Sigma} \tag{21}$$

$$\Sigma > 0 \tag{22}$$

$$\tilde{\Sigma} \geq 0 \;\text{ or }\; \tilde{\Sigma} > 0 \tag{23}$$

$$\hat{\Sigma} \geq 0 \quad \text{and} \quad \det \hat{\Sigma} = 0. \tag{24}$$

The problem of determining linear relations compatible with noisy data can be formulated as follows:

Problem 1 [1, 2] – Given a sample covariance matrix of noisy observations, Σ, determine positive definite or semidefinite noise covariance matrices $\tilde{\Sigma}$ such that

$$\hat{\Sigma} = \Sigma - \tilde{\Sigma} \geq 0 \quad \text{and} \quad \det \hat{\Sigma} = 0. \tag{25}$$

Any basis of $\ker \hat{\Sigma}$ will describe a set of linear relations compatible with the data and with assumptions (16)–(18).

C The Frisch scheme

This scheme, proposed by Ragnar Frisch in 1934 [5], is based on assumptions (16)–(18) and on the further assumption of mutual independence of the noise sequences

$$\sum_{t=1}^{N} \tilde{x}_{it}\, \tilde{x}_{jt} = 0 \quad \forall i \neq j. \tag{26}$$

As a consequence of (26), the sample covariance matrix of the noise will be diagonal. By introducing the suffix n to denote the dimension of square matrices, we will thus have

$$\tilde{\varSigma}_n = \mathrm{diag}\left[\, \tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2 \,\right] \geq 0 \qquad \text{or} \qquad > 0 \tag{27}$$

where $\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2$ are the sample variances of the noise terms $\tilde{x}_1, \ldots, \tilde{x}_n$. Every positive definite or semidefinite diagonal matrix $\tilde{\varSigma}_n$ such that

$$\hat{\varSigma}_n = \varSigma_n - \tilde{\varSigma}_n \geq 0 \quad \text{and} \quad \det \hat{\varSigma} = 0 \tag{28}$$

is a *solution* of the Frisch scheme. The corresponding point $P = (\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2) \in \mathcal{R}^n$ can be considered as an admissible solution in the noise space while the parameters α_i in (5) or β_i in (6) define the associate solution in the parameter space.

It can be observed that the set of parameters (5) or (6) refers only to the case of $\dim \ker \hat{\varSigma}_n = 1$. A solution in the noise space can, however, be associated also with noise covariance matrices $\tilde{\varSigma}_n$ such that $\dim \ker \hat{\varSigma}_n > 1$ that correspond to multiple independent linear relations between the columns (rows) of $\hat{\varSigma}_n$. The maximal dimension of $\ker \hat{\varSigma}_n$ will be denoted in the following as $\mathrm{Maxcor}_F(\varSigma_n)$ (maximal corank of $\varSigma_n$ under the assumptions of the Frisch scheme) [2].

Properties of the solutions in the noise space

A problem of great importance in the analysis of the properties and in the application of the Frisch scheme concerns the description of the loci of solutions in the noise and parameter spaces. While the locus of solutions in the parameter space has nice properties only in a well defined case (compatibility of the data with a single linear relation under the assumptions of the Frisch scheme, i.e. $\mathrm{Maxcor}_F(\varSigma_n) = 1$), the locus of solutions in the noise space is the same in every situation and is described by the following theorem [12].

Theorem 1 – All admissible solutions in the noise space lie on a convex (hyper)surface $\mathcal{S}(\varSigma_n)$ whose concavity faces the origin and whose intersections with the coordinate axes are the points $(0, \ldots, \tilde{\sigma}_i^2, \ldots, 0)$ corresponding to the n least squares solutions (see Fig. 1).

Definition 1 [7] – The (hyper)surface $\mathcal{S}(\varSigma_n)$ will be called *singularity (hyper)surface* of $\varSigma_n$ because its points define noise covariance matrices $\tilde{\varSigma}_n$ associated with singular matrices $\hat{\varSigma}_n$.

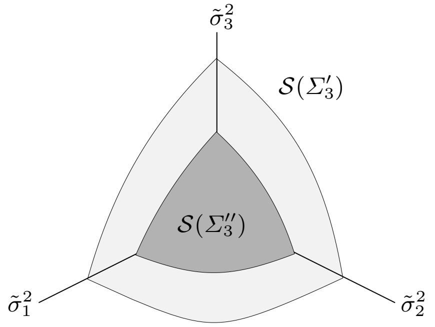


Fig. 1. Loci $\mathcal{S}(\Sigma_3)$ of admissible noise points for $n = 3$ and different amounts of noise

A problem of great relevance concerns the conditions under which a covariance matrix is compatible with more linear relations i.e. the evaluation of $\text{Maxcor}_F(\Sigma_n)$. A fundamental result concerning this problem is the following [2].

Theorem 2 – $\text{Maxcor}_F(\Sigma_n) = 1$ if and only if all entries of Σ_n^{-1} are positive or can be made positive (Frobenius–like according to the definition of Kalman [2]) by changing the sign of some variables.

Under the conditions of Theorem 2, the locus of solutions in the parameter space is described by the following theorem that shows the great relevance of the n least squares solutions (that correspond to the assumption that only one variable is affected by errors):

Theorem 3 – When $\text{Maxcor}_F(\Sigma_n) = 1$, the coefficients $\alpha_1, \ldots, \alpha_n$ of all linear relations compatible with the Frisch scheme lie (by normalizing one of the coefficients to 1) inside the simplex whose vertices are defined by the n least squares solutions (see Fig. 2).

Other important properties of the loci of solutions in the noise and parameter spaces are described by the following theorems.

Theorem 4 – When $\text{Maxcor}_F(\Sigma_n) = 1$ the points of the simplex of solutions in the parameter space are linked by a one–to–one relation (isomorphism) to the points of $\mathcal{S}(\Sigma_n)$.

Theorem 5 [22] – When $\text{Maxcor}_F(\Sigma_n) > 1$, $\mathcal{S}(\Sigma_n)$ is nonuniformly convex.

Theorem 6 [10] – All points of $\mathcal{S}(\Sigma_n)$ where $\text{Cor}\,(\Sigma_n) = k$ $(k > 1)$ are accumulation points for those where $\text{Cor}\,(\Sigma_n) = k - 1$.

Computation of $\mathbf{Maxcor}_{\boldsymbol{F}}(\boldsymbol{\Sigma_n})$

Despite its simple formulation, the problem of determining the maximal number of linear relations compatible with a set of noisy data in the context of the Frisch scheme has remained unsolved for many years. One of the reasons

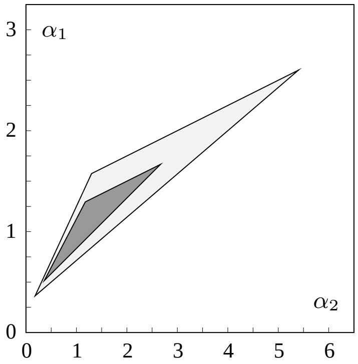


Fig. 2. Loci of admissible parameters for $n = 3$ and different amounts of noise

is probably due to the focus of many researches on the locus of the solutions in the parameter space and to the practical impossibility of describing this locus, when $\mathrm{Maxcor}_F\,(\Sigma_n) > 1$, except than in elementary cases. An upper bound to $\mathrm{Maxcor}_F\,(\Sigma_n)$ has been given in [11]; geometric conditions to evaluate $\mathrm{Maxcor}_F\,(\Sigma_n)$ have, instead, been given in [10] on the basis of the analysis of the properties of the locus of the solutions in the noise space.

Define, to this purpose, the singularity (hyper)surface $\mathcal{S}(\Sigma_{n/r})$ as the locus of the points $(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2) \in \mathcal{R}^r$ such that

$$\Sigma_n - \mathrm{diag}\left[\,\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2, 0, \ldots, 0\,\right] \geq 0 \tag{29}$$

and Σ_r as the sample covariance matrix of the first r variables. Then the following geometric relations hold:

Theorem 7 [23] – $\mathcal{S}(\Sigma_{n/r})$ lies always under or on $\mathcal{S}(\Sigma_r)$ (see Fig. 3).

Theorem 8 [10] – $\mathrm{Maxcor}_F\,(\Sigma_n) \geq q$ if and only if $\mathcal{S}(\Sigma_{n-q+1}) \cap \mathcal{S}(\Sigma_{n/n-q+1}) \neq \{0\}$ for every subset of $n-q+1$ variables, i.e. for every permutation of the data leading to different subgroups in the first $n-q+1$ positions.

Theorem 8 allows the straightforward formulation of an algorithm to compute $\mathrm{Maxcor}_F\,(\Sigma_n)$ by testing whether it is $\geq 2, 3, \ldots$ until the required conditions are no longer satisfied.

Radial parameterization for Frisch singularity hypersurfaces

The existence of common points between different singularity hypersurfaces can be easily and efficiently verified by relying on a radial parameterization of these surfaces. A parameterization of this kind can be used effectively for computing the points of $\mathcal{S}(\Sigma_n)$ and also to perform fast searches on $\mathcal{S}(\Sigma_n)$ to minimize a given cost function. It has been originally introduced in [24] to solve

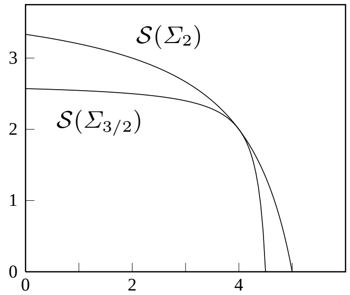


Fig. 3. Common points between $\mathcal{S}(\Sigma_2)$ and $\mathcal{S}(\Sigma_{3/2})$ in a (3×3) covariance matrix with $\mathrm{Maxcor}_F(\Sigma_3) = 2$

the congruence problems associated with the identification of multivariable processes by associating multivariable models to *directions* in the noise space instead than to specific points of singularity hypersurfaces [25]. It is important to note that this parameterization can be used also for the direct computation of the distance between two singularity hypersurfaces along a given direction.

Radial parameterizations will be described here for the algebraic case; their extension to the dynamic case is straightforward.

Let $\xi = (\xi_1, \ldots, \xi_n)$ be a generic point in the first orthant of $\mathcal{R}^n$; the intersection, $P = (\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2)$, between the straight line through the origin and ξ with $\mathcal{S}(\Sigma_n)$ satisfies the conditions

$$\Sigma_n - \tilde{\Sigma}_n \geq 0 \tag{30}$$

and

$$\lambda\, P = \xi \quad \text{with} \quad \lambda > 0. \tag{31}$$

It follows that

$$\det\left(\Sigma_n - \frac{1}{\lambda}\,\tilde{\Sigma}_n^{\xi}\right) = 0 \tag{32}$$

where

$$\tilde{\Sigma}_n^{\xi} = \mathrm{diag}\left[\,\xi_1,\, \ldots, \xi_n\,\right]. \tag{33}$$

Relation (32) is equivalent $(\Sigma_n > 0)$ to

$$\det\left(\lambda\, I - \Sigma_n^{-1}\tilde{\Sigma}_n^{\xi}\,\right) = 0 \tag{34}$$

so that the solution compatible with condition (30) is given by

$$P = \frac{\xi}{\lambda_M} \tag{35}$$

with

$$\lambda_M = \max\,\mathrm{eig}\left(\Sigma_n^{-1}\tilde{\Sigma}_n^{\xi}\,\right). \tag{36}$$

The points of $\mathcal{S}(\Sigma_n)$ associated with straight lines from the origin can thus be obtained by computing Σ_n^{-1} and the intersection between any line and $\mathcal{S}(\Sigma_n)$ by means of (35) and (36).

These results define a parameterization of singularity hypersurfaces that associates their points with the sheaf of lines from the origin in the first orthant.

D Computation of a single solution in the context of the Frisch scheme

The properties of the loci of solutions in the noise and parameter spaces show that, given a (sample) data covariance matrix Σ_n it is impossible to discriminate any solution (i.e. any decomposition of Σ) against any other, unless additional information (e.g. noise variance ratios) is available. It is however possible to take advantage of the differences between data belonging to two finite sequences and estimate, under some conditions, the linear relation actually linking the noiseless data. For this purpose we will introduce some abstract definitions and conditions that will eventually lead to algorithms appliable in real cases.

Complete sets of data for the Frisch scheme

The definitions and properties that follow concern the asymptotic case (infinite sequence of data).

Definition 2 [7] – Two noise–free data covariance matrices of the same linear algebraic process, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are defined as *independent* if

$$\dim\ker \hat{\Sigma}_1 = \dim\ker \hat{\Sigma}_2 = \dim\ker \left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right) = 1. \tag{37}$$

Property 1 – If $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are independent there exists a unique (modulo scaling) vector A satisfying the conditions

$$\hat{\Sigma}_1\ A = \hat{\Sigma}_2\ A = \left(\hat{\Sigma}_1 - \hat{\Sigma}_2\right)\ A = 0. \tag{38}$$

Definition 3 [7] – Two noisy data covariance matrices of the same linear algebraic process, $\Sigma_1 > 0$ and $\Sigma_2 > 0$ are defined as *independent* if

$$\dim\ker \left(\Sigma_1 - \Sigma_2\right) = 1. \tag{39}$$

Theorem 9 [7] – Two independent noisy covariance matrices, Σ_1 and Σ_2, satisfy the following conditions under the Frisch scheme

$$\dim\ker \left(\Sigma_1 - \tilde{\Sigma}\,\right) = \dim\ker \left(\Sigma_2 - \tilde{\Sigma}\,\right) = 1 \tag{40}$$

$$\left(\Sigma_1 - \Sigma_2\right) A = \left(\Sigma_1 - \tilde{\Sigma}\right) A = \left(\Sigma_2 - \tilde{\Sigma}\right) A = 0, \tag{41}$$

where $A = [\alpha_1 \, \alpha_2 \, \dots \, \alpha_n]^T$ defines the process model (5) and $\tilde{\Sigma} \geq 0$ is a diagonal matrix satisfying the conditions

$$\Sigma_1 - \tilde{\Sigma} \geq 0, \quad \det(\Sigma_1 - \tilde{\Sigma}) = 0, \tag{42}$$

$$\Sigma_2 - \tilde{\Sigma} \geq 0, \quad \det(\Sigma_2 - \tilde{\Sigma}) = 0. \tag{43}$$

Theorem 10 [7] – Among all points common to the hypersurfaces of admissible noise points associated with the independent noisy covariance matrices Σ_1 and Σ_2, one and only one point is mapped, according to Σ_1 and Σ_2, into the same point of the parameter space (see Figs. 4 and 5).

Corollary 1 – The Frisch scheme leads to a unique solution determined by every pair of independent noisy data covariance matrices of the process.

Corollary 2 – Two independent noisy data covariance matrices of a process constitute a *complete* set of data for the Frisch scheme.

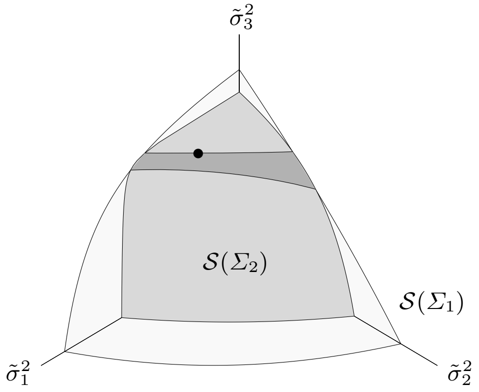


Fig. 4. Admissible noise points ($n = 3$).

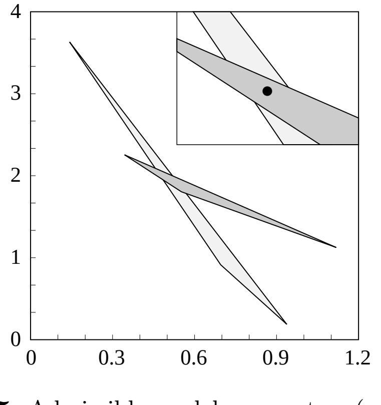


Fig. 5. Admissible model parameters ($n = 3$).

Determination of the Frisch solution from real data

In all practical cases, even when two data sets are available, it is worthless testing whether they meet the independence conditions. Theorem 10, however, allows defining a consistent criterion to search for solutions even when the intersection between $\mathcal{S}(\Sigma_1)$ and $\mathcal{S}(\Sigma_2)$ does not contain any point mapped, by Σ_1 and Σ_2 into the same point of the parameter space.

Criterion 1 [26] – Consider a pair of covariance matrices Σ_1 and Σ_2 and their loci of solutions, $\mathcal{S}(\Sigma_1)$, $\mathcal{S}(\Sigma_2)$ in the noise space. The best approximation of the actual noise variances will be given by the point $P \in \mathcal{S}(\Sigma_1) \cap \mathcal{S}(\Sigma_2)$ that minimizes the Euclidean norm of the distance between the parameter vectors A' and A'' associated to P by Σ_1 and Σ_2.

Remark 1 – Criterion 1 is consistent since the cost function $f(P) = \|A' - A''\|_2$ annihilates when Σ_1 and Σ_2 are independent.

Remark 2 – Once that the minimum of $f(P)$ has been found, two solutions, A' and A'' will be available and their distance is a measure of the reliability of the procedure. Their mean value can be taken as problem solution.

Remark 3 – It can be observed that the outlined procedure can be applied even when the simplexes associated with Σ_1 and Σ_2 do not share common points.

3 The Frisch scheme in the dynamic case

A The SISO case

The extension of the Frisch scheme to the identification of dynamical processes can rely on some properties that, differently from the algebraic case, allow to obtain a single solution also when a single sequence of data is available. To allow a simpler formulation of the problem, the SISO case will be firstly considered while the identification of MIMO systems will be treated only in a second time. Consider a dynamic SISO system of order n described by the input–output model

$$\hat{y}(t+n) = \sum_{k=1}^{n} \alpha_k \, \hat{y}(t+k-1) + \sum_{k=1}^{n+1} \beta_k \, \hat{u}(t+k-1) \qquad (44)$$

where $\hat{u}(t)$ denotes the input at time t and $\hat{y}(t)$ the output. Consider also noisy input/output observations, $u(t)$ and $y(t)$ given by

$$u(t) = \hat{u}(t) + \tilde{u}(t) \qquad (45)$$

$$y(t) = \hat{y}(t) + \tilde{y}(t) \qquad (46)$$

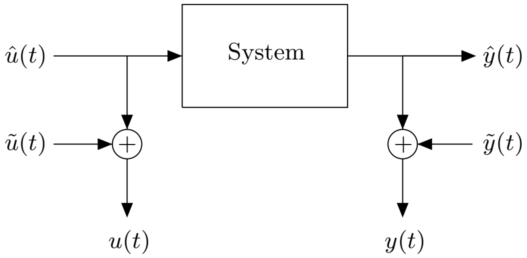


Fig. 6. The dynamic Frisch scheme context

where $\tilde{u}(t)$ and $\tilde{y}(t)$ are white processes with zero mean, mutually uncorrelated and uncorrelated with $\hat{u}(t)$ (see Fig. 6).

Define now the Hankel matrices

$$X_k(y) = \begin{bmatrix} y(1) & \dots & y(k) \\ y(2) & \dots & y(k+1) \\ \vdots & \ddots & \vdots \\ y(N) & \dots & y(k+N-1) \end{bmatrix}, \tag{47}$$

$$X_k(u) = \begin{bmatrix} u(1) & \dots & u(k) \\ u(2) & \dots & u(k+1) \\ \vdots & \ddots & \vdots \\ u(N) & \dots & u(k+N-1) \end{bmatrix}, \tag{48}$$

the matrix of input/output samples

$$X_k = \left[\, X_{k+1}(y)\ X_{k+1}(u) \,\right] \tag{49}$$

and the sample covariance matrices $\varSigma_k$ given by

$$\varSigma_k = \frac{X_k^T X_k}{N} = \begin{bmatrix} \varSigma(yy) & \varSigma(yu) \\ \varSigma(uy) & \varSigma(uu) \end{bmatrix}. \tag{50}$$

Denoting with $\tilde{\sigma}_u^{2*}$ and $\tilde{\sigma}_y^{2*}$ the variances of $\tilde{u}(t)$ and $\tilde{y}(t)$ and with P^* the point

$$P^* = \left(\, \tilde{\sigma}_y^{2*}, \tilde{\sigma}_u^{2*} \,\right), \tag{51}$$

the previous assumptions establish that, when $N \to \infty$

$$\varSigma_k = \hat{\varSigma}_k + \tilde{\varSigma}_k^* \tag{52}$$

where

$$\tilde{\varSigma}_k^* = \operatorname{diag}\left[\tilde{\sigma}_y^{2*} I_{k+1},\, \tilde{\sigma}_u^{2*} I_{k+1}\right]. \tag{53}$$

The identification problem, in the context of the Frisch scheme, consists in determining the order and the parameters of model (44), or of any equivalent state–space model, and the additive noise variances $\tilde{\sigma}_y^{2*}$, $\tilde{\sigma}_u^{2*}$ on the basis of the knowledge of the noisy sequences $u(\cdot)$, $y(\cdot)$ or, equivalently, of the sequence of increasing–dimension matrices Σ_k for $k = 1, 2, \ldots$.

Model (44) implies, for every input sequence persistently exciting of order $n+1$, the nonsingularity of $\hat{\Sigma}_1, \ldots, \hat{\Sigma}_{n-1}$ and the singularity of $\hat{\Sigma}_k$ for $k \geq n$. For any value of k (lower, equal or larger than n), a point $P = (\tilde{\sigma}_y^2, \tilde{\sigma}_u^2)$ belonging to the first orthant of the noise space, defines an admissible solution if and only if

$$\dim \ker\bigl(\Sigma_k - \tilde{\Sigma}_k\bigr) = 1, \tag{54}$$

$$\Sigma_k - \tilde{\Sigma}_k \geq 0 \tag{55}$$

where $\tilde{\Sigma}_k$ is the noise covariance matrix defined by P

$$\tilde{\Sigma}_k = \tilde{\Sigma}_k(P) = \mathrm{diag}\,[\tilde{\sigma}_y^2\, I_{k+1},\, \tilde{\sigma}_u^2\, I_{k+1}]. \tag{56}$$

The corresponding solution in the parameter space, $\theta(P) = [\alpha_1(P), \ldots, \alpha_n(P), -1, \beta_1(P), \ldots, \beta_{n+1}(P)]^T$, is univocally defined by $\ker\,(\Sigma_k - \tilde{\Sigma}_k)$, i.e. by the relation

$$\hat{\Sigma}_k(P)\,\theta(P) = \bigl(\Sigma_k - \tilde{\Sigma}_k(P)\bigr)\,\theta(P) = 0. \tag{57}$$

Theorem 11 [12] – For every $k > 0$ all admissible points define a convex curve $\mathcal{S}(\Sigma_k)$ in the first quadrant of the noise plane $\mathcal{R}^2$ with a concavity facing the origin. The point $P^* = (\tilde{\sigma}_y^{2*}, \tilde{\sigma}_u^{2*})$ associated with the actual noise variances belongs to all curves $\mathcal{S}(\Sigma_k)$ when $k \geq n$ and $\theta(P^*)$ is the true parameter vector, θ^*.

Theorem 12 [12] – If i and j are integers with $j > i$, then $\mathcal{S}(\Sigma_j)$ lies under or on $\mathcal{S}(\Sigma_i)$.

Remark 4 – $\mathcal{S}(\Sigma_k)$ partitions the noise space $\mathcal{R}^2$ into the regions of the points σ^+ associated with positive definite matrices $\hat{\Sigma}_k = \Sigma_k - \tilde{\Sigma}^+$ and of the points σ^n associated with non definite and negative definite matrices $\hat{\Sigma}_k = \Sigma_k - \tilde{\Sigma}^-$. These regions lie under and over $\mathcal{S}(\Sigma_k)$ respectively.

Remark 5 – Note that the dimension of the noise space is always equal to the total number of inputs and outputs (two for the SISO case) i.e. to the number of variables, like in the algebraic case. The dimension of the parameter space depends also on the order of the process.

Remark 6 – Theorem 11 can be considered as a corollary of Theorem 8 since, because of the well–known shift property of dynamical systems, $\mathrm{Maxcor}_F\, \Sigma_k = k - n + 1$ when $k \geq n$.

B Frisch identification of real processes and model selection criteria

The key property described by Theorem 1 holds only when the (asymptotic) properties assumed for the additive noise sequences (mutual orthogonality and orthogonality with the input/output sequences) hold, i.e. when $\tilde{u}(\cdot)$ and $\tilde{y}(\cdot)$ are uncorrelated white sequences with infinite length. In all other cases no common point between different curves can be observed. Similar consequences follow from violations on the linearity and time–invariance assumptions. Moreover, the algorithms that can be developed to estimate a single solution from real data can exhibit robustness and reliability problems that require the development of suitable criteria.

The number of criteria that can be developed is relatively large. Many of them, however, are not endowed with sufficient robustness for real applications. As an example, it is possible to cite the selection criterion based on the minimal radial distance between two adjacent curves or, more generally, hypersurfaces. It is easy to show that such a criterion can select *any* point by properly scaling the data and insensitivity to data scaling is just *one* of the requirements for possible criteria.

Among the criteria that have shown a good robustness level it is possible to mention the following ones.

The shifted relation criterion [16, 17]

This criterion is based on the rank deficiency property of the matrices $\hat{\Sigma}_k(P^*)$ for $k \geq n$ and is based on a cost function that requires the computation of the intersections of a line from the origin with the curve associated with the considered model order, $\mathcal{S}(\Sigma_n)$ and with the subsequent one $\mathcal{S}(\Sigma_{n+1})$.

The covariance–matching criterion [19]

This criterion is based on a cost function that compares the theoretical covariances of the EIV process

$$\begin{aligned}\gamma(t) &= \alpha_1\, y(t) + \alpha_2\, y(t+1) + \cdots - y(t+n) + \beta_1\, u(t) + \cdots + \beta_{n+1}\, u(t+n)\\ &= \alpha_1\, \tilde{y}(t) + \alpha_2\, y(t+1) + \cdots - \tilde{y}(t+n) + \beta_1\, \tilde{u}(t) + \cdots + \beta_{n+1}\, \tilde{u}(t+n)\end{aligned}$$

with the sample values computed from the data. Both values are associated with the considered point of the curve and are (asymptotically) coincident only in the point corresponding to the actual noise variances.

A criterion based on high–order Yule–Walker equations [18]

This criterion, that can also be considered as belonging to the family of instrumental variable methods, is based on the computation of the closing errors of a set of high order Yule–Walker equations. Also in this case the closing error depends on the considered model and is asymptotically null only in the point associated with the true noise variances and process parameters.

The application of the described or of other possible criteria can be performed by minimizing their value on the singularity curve associated with the selected model order. This can be performed by using standard search algorithms and by selecting a suitable stop threshold. The efficiency of practical implementations can take great advantage from parameterizations of the curves that allow to perform the search by computing only a very limited number of points, like the radial parameterization described in [24] that can be easily extended to the dynamic case [25].

C The MIMO case

The extension of Frisch identification techniques to the MISO case is straightforward; this is not the case for MIMO processes that face conceptual and practical congruence problems not present in the single–output case.

The multivariable identification problem

The MIMO dynamic systems considered in this section are described by the input–output model

$$P(z)\,\hat{y}(t) = Q(z)\,\hat{u}(t), \tag{58}$$

where $\hat{u}(t) \in \mathcal{R}^r$, $\hat{y}(t) \in \mathcal{R}^m$ and $P(z)$, $Q(z)$ are $(m \times m)$ and $(m \times r)$ left coprime polynomial matrices in the unitary advance operator z. By selecting a minimal parameterization [27], model (58) can be partitioned into the set of m relations

$$\hat{y}_i(t+\nu_i) = \sum_{j=1}^{m}\sum_{k=1}^{\nu_{ij}} \alpha_{ijk}\ \hat{y}_j(t+k-1) + \sum_{j=1}^{r}\sum_{k=1}^{\nu_i} \beta_{ijk}\ \hat{u}_j(t+k-1) \tag{59}$$

where the integers ν_i $(i = 1, \ldots, m)$ that appear in (59) and describe the structure of the model are the observability invariants of the system. The integers ν_{ij} are completely defined by these invariants through the relations

$$\nu_{ij} = \nu_i \qquad \text{for } i = j \tag{60}$$

$$\nu_{ij} = \min(\nu_i + 1, \nu_j) \qquad \text{for } i > j \tag{61}$$

$$\nu_{ij} = \min(\nu_i, \nu_j) \qquad \text{for } i < j\ . \tag{62}$$

For a complete description of the properties of the scalars $\{\nu_i, \alpha_{ijk}, \beta_{ijk}\}$ see [27]. The order of system (59) is given by

$$n = \sum_{i=1}^{m} \nu_i \ ; \tag{63}$$

ν_M will denote, in the following, the maximal observability index, i.e.

$$\nu_M = \max_i \{\nu_i, \ i = 1, \ldots, m\} \ . \tag{64}$$

In an errors–in–variables context, the noise–free signals $\hat{u}(t)$ and $\hat{y}(t)$ linked by model (59) are not directly accessible and only the noisy observations

$$u(t) = \hat{u}(t) + \tilde{u}(t) \tag{65}$$

$$y(t) = \hat{y}(t) + \tilde{y}(t) \ , \tag{66}$$

are available. In this paper the additive noises $\tilde{u}(t)$ and $\tilde{y}(t)$ satisfy the following assumptions.

1) The processes $\tilde{u}(t)$ and $\tilde{y}(t)$ are zero–mean, mutually uncorrelated white noise sequences, with unknown covariance matrices $\tilde{\varSigma}_u^* = \mathrm{diag}\,[\tilde{\sigma}_{u_1}^{2*}, \ldots \, \tilde{\sigma}_{u_r}^{2*}]$ and $\tilde{\varSigma}_y^* = \mathrm{diag}\,[\tilde{\sigma}_{y_1}^{2*}, \ldots \, \tilde{\sigma}_{y_m}^{2*}]$;

2) The processes $\tilde{u}(t)$ and $\tilde{y}(t)$ are uncorrelated with the the noise–free signal $\hat{u}(t)$.

The EIV MIMO identification problem can be stated as follows: given N noisy input–output observations $u(\cdot)$, $y(\cdot)$, estimate the noise covariance matrices $\tilde{\varSigma}_u^*$, $\tilde{\varSigma}_y^*$ and the coefficients α_{ijk}, β_{ijk} of model (59).

Properties of EIV MIMO systems

Consider the Hankel matrix

$$H_k(\hat{y}_i) = \begin{bmatrix} \hat{y}_i(1) & \ldots & \hat{y}_i(k) \\ \hat{y}_i(2) & \ldots & \hat{y}_i(k+1) \\ \vdots & & \vdots \\ \hat{y}_i(N) & \ldots & \hat{y}_i(k+N-1) \end{bmatrix} \ , \tag{67}$$

and the analogous matrices $H_k(y_i)$, $H_k(\tilde{y}_i)$, $H_k(\hat{u}_i)$, $H_k(u_i)$ and $H_k(\tilde{u}_i)$. Define also the multi–index $k^M = (k_1, \ldots, k_{m+r})$ and the matrix

$$\hat{H}(k^M) = \left[\, H_{k_1}(\hat{y}_1) \ldots H_{k_m}(\hat{y}_m) \, H_{k_m+1}(\hat{u}_1) \ldots H_{k_m+r}(\hat{u}_r) \,\right] .$$

Relations (59) can be used to write an overdetermined set of linear equations in the unknowns α_{ijk} and β_{ijk}. In fact, by considering the multi–index $\nu^M = (\nu_1 + 1, \ldots, \nu_m + 1, \nu_M, \ldots, \nu_M)$, relations (59) imply that

$$\hat{H}(\nu^M) \, \varTheta = 0 \ , \tag{68}$$

where

$$\Theta = \left[\, \theta_1 \; \theta_2 \, \cdots \, \theta_m \,\right] \, , \tag{69}$$

and

$$\begin{aligned} \theta_i = \big[\, & \alpha_{i11} \, \cdots \, \alpha_{i1\nu_{i1}} \; \underbrace{0 \cdots 0}_{(\nu_1+1-\nu_{i1})} \; | \cdots | \\ & | \, \alpha_{ii1} \, \cdots \, \alpha_{ii\nu_i} \; -1 \, | \cdots | \, \alpha_{im1} \, \cdots \, \alpha_{im\nu_{im}} \; \underbrace{0 \cdots 0}_{(\nu_m+1-\nu_{im})} \; | \\ & | \, \beta_{i11} \, \cdots \, \beta_{i1\nu_i} \; \underbrace{0 \cdots 0}_{(\nu_M-\nu_i)} \; | \cdots | \, \beta_{ir1} \, \cdots \, \beta_{ir\nu_i} \; \underbrace{0 \cdots 0}_{(\nu_M-\nu_i)} \,\big]^T . \end{aligned} \tag{70}$$

By defining the covariance matrix $\hat{\Sigma}(\nu^M)$ as

$$\hat{\Sigma}(\nu^M) = \frac{1}{N} \, \hat{H}(\nu^M)^T \hat{H}(\nu^M) = \begin{bmatrix} \hat{\Sigma}(\hat{y}_1\hat{y}_1) & \hat{\Sigma}(\hat{y}_1\hat{y}_2) & \ldots & \hat{\Sigma}(\hat{y}_1\hat{u}_r) \\ \hat{\Sigma}(\hat{y}_2\hat{y}_1) & \hat{\Sigma}(\hat{y}_2\hat{y}_2) & \ldots & \hat{\Sigma}(\hat{y}_2\hat{u}_r) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Sigma}(\hat{u}_r\hat{y}_1) & \hat{\Sigma}(\hat{u}_r\hat{y}_2) & \ldots & \hat{\Sigma}(\hat{u}_r\hat{u}_r) \end{bmatrix} , \tag{71}$$

equation (68) implies that

$$\hat{\Sigma}(\nu^M) \, \Theta = 0 \; . \tag{72}$$

Define now the point P^* as

$$P^* = \left(\tilde{\sigma}^{2*}_{y_1}, \ldots \, \tilde{\sigma}^{2*}_{y_m}, \, \tilde{\sigma}^{2*}_{u_1}, \ldots \, \tilde{\sigma}^{2*}_{u_r} \right); \tag{73}$$

the assumptions of noise additivity and independence at the basis of the Frisch scheme lead, for $N \to \infty$, to the decomposition

$$\Sigma\big(\nu^M\big) = \hat{\Sigma}\big(\nu^M\big) + \tilde{\Sigma}^*\big(\nu^M\big), \tag{74}$$

where

$$\tilde{\Sigma}^*(\nu^M) = \operatorname{diag} \left[\, \tilde{\sigma}^{2*}_{y_1} I_{\nu_1+1}, \ldots, \tilde{\sigma}^{2*}_{y_m} I_{\nu_m+1}, \, \tilde{\sigma}^{2*}_{u_1} I_{\nu_M}, \ldots, \tilde{\sigma}^{2*}_{u_r} I_{\nu_M} \right]. \tag{75}$$

Consider now the generic subsystem i described by relation (59), the multi–index

$$\nu_i^M = (\nu_{i1}, \ldots, \nu_i + 1, \ldots, \nu_{im}, \nu_i, \ldots, \nu_i) \tag{76}$$

and the i–th set of parameters

$$\eta_i = \left[\, \alpha_{i11}, \ldots, \alpha_{ii1}, \ldots, \alpha_{ii\nu_i}, -1, \ldots, \alpha_{im\nu_{im}}, \beta_{i11}, \ldots, \beta_{i1\nu_i}, \ldots, \beta_{ir\nu_i} \right]^T ; \tag{77}$$

then

$$\hat{\Sigma}(\nu_i^M) \, \eta_i = \left[\, \Sigma(\nu_i^M) - \tilde{\Sigma}^*(\nu_i^M) \,\right] \eta_i = 0. \tag{78}$$

By defining the relation between multi–indices

$$k^M < h^M \quad \text{if} \quad k_i < h_i \quad \text{for} \quad i = 1, \ldots, m + r \, , \tag{79}$$

it is possible to state the following theorems whose proofs can be carried out along the lines considered in [10, 12] for the MISO case.

Theorem 13 – For every structure $\xi = (\nu_1, \ldots, \nu_m)$, the admissible noise–space solutions associated with the i–th subsystem, i.e. the locus of points $(\tilde{\sigma}_1^2, \ldots\ \tilde{\sigma}_{m+r}^2)$ such that

$$\hat{\Sigma}(\nu_i^M) = \Sigma(\nu_i^M) - \tilde{\Sigma}(\nu_i^M) \geq 0 \ , \tag{80}$$

is a convex hypersurface $\mathcal{S}(\Sigma(\nu_i^M))$ belonging to the first orthant of $\mathcal{R}^{m+r}$ whose concavity faces the origin (singularity hypersurface).

Theorem 14 – If k_i^M and h_i^M are multi–indices with $h_i^M > k_i^M$, then $\mathcal{S}(\Sigma(h_i^M))$ lies under $\mathcal{S}(\Sigma(k_i^M))$.

Theorem 15 – All hypersurfaces $\mathcal{S}(\Sigma(k_i^M))$, $(i = 1, \ldots, m)$ with $k_i^M > \nu_i^M$ have the single common point P^* corresponding to the actual variances of the noise on the data.

Theorems 13, 14 and 15 give a picture of the multivariable case similar to the pictures of the SISO and MISO cases. The existence of a single point (exact noise variances) common to the singularity hypersurfaces associated with the different subsystems, allows to solve the MIMO identification problem in a way similar to the SISO and MISO cases, by computing, in a congruent way, the parameters of every subsystem, defined by the kernels of the matrices $\mathcal{S}(\hat{\Sigma}(\nu_i^M))$;

EIV MIMO identification

The identification of real processes requires, as in the SISO or MISO cases, the definition of suitable selection criteria since the common point described by Theorem 15 will no longer exist. The criteria described for the SISO (or MISO) case, as well as others, could be applied to every subsystem and this would lead to a complete parameterization of the multivariable model.

It must however be noted that a procedure of this kind would lead to an incongruent solution because the identification of the different subsystems would lead to the estimation of (slightly) different points in the noise space and the corresponding variances would be different and characterized by different ratios. As a consequence, the obtained model would be no longer associated with a single point in the noise space but with a set of points and this constitutes a serious limit to the congruence of the model and to some of its possible applications like, for instance, filtering.

The solution of this problem can be obtained by using radial parameterizations that allow to establish a one to one relation between *directions* in the noise space [25] and model parameters and by suitable extensions of the already described selection criteria. The identification of multivariable EIV models is described in detail in [25, 28].

4 Applications of Frisch identification techniques

A Blind identification of SIMO FIR systems

The blind identification of dynamic systems is of great relevance in many fields like telecommunications, sismology, radioastronomy, etc. The purpose is the reconstruction of the transfer function of a transmission channel starting from noisy measurements performed only on its output [29, 30].

Blind identification relies on linear models describing a set of parallel channels driven by an unknown sequence and characterized by a finite impulse response (FIR). These models can describe a single unknown source in presence of multiple spatially and/or temporally distributed sensors. In the two–channel case the process is described (see Fig. 7) by the model

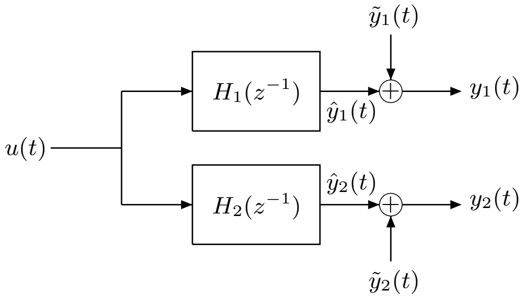


Fig. 7. Two–channel FIR system

$$\hat{y}_i(t) = H_i(z^{-1})\, u(t) = \sum_{k=0}^{n} h_i(k)\, u(t-k), \quad i = 1, 2 \tag{81}$$

$$H_i(z^{-1}) = h_i(0) + h_i(1)\, z^{-1} + \cdots + h_i(n)\, z^{-n}, \quad i = 1, 2 \tag{82}$$

$$y_i(t) = \hat{y}_i(t) + \tilde{y}_i(t), \quad i = 1, 2 \tag{83}$$

where $\tilde{y}_1(t)$ and $\tilde{y}_2(t)$ are mutually uncorrelated white noises, uncorrelated with $u(t)$ and with unknown variances $\tilde{\sigma}_{y1}^{2*}, \tilde{\sigma}_{y2}^{2*}$. Relations (81) allow mapping the blind identification problem into an errors–in–variables identification one that can be solved (in the case of two channels) by using the identification procedures described for the SISO case or with more specific procedures [31]. The multichannel case is more complex and cannot be reconducted to the MISO or MIMO cases; a procedure solving this problem is described in [32]. In both cases the proposed approaches extend the existing blind channel identification procedures to the case of unbalanced channel noises.

B Identification of noisy autoregressive models

Autoregressive (AR) models are commonly used in a wide range of engineering applications, like spectral estimation, speech and image processing, noise cancellation etc.

A considerable attention has been dedicated, in the literature, to the problem of estimating the AR parameters from signals corrupted by white noise.

In this case the estimates obtained with classical AR identification methods (least–squares, Yule–Walker equations) are poor, particularly for low signal–to–noise ratio conditions [33, 34]. Consider the noisy AR model

$$x(t) = \alpha_1 \, x(t-1) + \cdots + \alpha_n \, x(t-n) + e(t), \tag{84}$$

$$y(t) = x(t) + w(t), \tag{85}$$

where $x(t)$ is the noise–free AR signal, $e(t)$ is the driving noise and $y(t)$ is the available observation affected by the additive noise $w(t)$; $e(t)$ and $w(t)$ are zero–mean white processes, mutually uncorrelated, with unknown variances σ_e^{2*} and σ_w^{2*}.

The AR+noise identification problem consists in estimating $\alpha_1, \ldots, \alpha_n$ and σ_e^{2*}, σ_w^{2*} starting from the available measurements $y(1), y(2), \ldots, y(N)$. Also this problem can be mapped into an EIV identification problem whose solution must be searched on a limited portion of a suitable singularity surface by applying the already described procedures [35, 36]. Other applications of blind FIR identification and AR+noise procedures concern speech enhancement [37, 38, 39].

5 Conclusions

This paper has presented an overview of several results concerning the properties of the Frisch scheme and its application to the estimation of linear relations from data affected by unknown amounts of additive noise and to the identification of dynamic processes in an Errors–in–Variables context. While it does not present new results, it integrates in an unitary view many results previously scattered in different works and underlines the links, not previously described, between the algebraic and dynamic contexts where the Frisch scheme can be applied.

References

1. Kalman RE (1982) Identification from real data. In: *Hazewinkel M, Rinnooy Kan HG, Reidel D (eds) Current developments in the interface: Economics, Econometrics, Mathematics.* Dordrecht, The Netherlands

2. R. E. Kalman RE (to appear) *Nine lectures on identification.* Lecture notes on Economics and Mathematical Systems, Springer–Verlag, Berlin
3. Kalman RE (1982) System identification from noisy data. In: *Bednarek AR, Cesari L (eds) Dynamical Systems II*—, Academic Press, New York
4. Söderström T (2007) *Automatica*, 43:939–958
5. Frisch R (1934) Statistical confluence analysis by means of complete regression systems. *Pub. No. 5, Economic Institute, Oslo University*
6. Anderson BDO, Deistler M, Scherrer W (1996) *Automatica*, 32:1031–1035
7. Guidorzi R (1991) *Systems & Control Letters*, 17:415–424
8. Guidorzi R (1993) Errors–in–variables identification and model uniqueness. In: *Haagen K, Bartholomew DJ, Deistler M (eds) Statistical Modelling and Latent Variables*, North Holland, Amsterdam
9. Malinvaud E (1980) *Méthodes statistiques de l'économétrie*, 3rd edition. Dunod, Paris
10. Guidorzi R (1995) *Systems & Control Letters*, 24:159–166
11. Woodgate KG (1995) *Systems & Control Letters*, 24:153–158
12. Beghelli S, Guidorzi R, Soverini U (1990) *Automatica*, 26:171–176
13. Anderson BDO, Deistler M (1984) *Journal of Time Series Analysis*, 5:1–13
14. Deistler M (1986) In: *Bittanti S (ed) Time series and linear systems,* Lecture notes in Control and Information Sciences: 37–67, Springer–Verlag, Berlin
15. Stoica P, Nehorai A (1987) *Automatica*, 23:541–543
16. Beghelli S, Castaldi P, Guidorzi R, Soverini U (1993) *Proc. of the 9th International Conference on Systems Engineering*: 480–484, Las Vegas, Nevada
17. Diversi R, Guidorzi R, Soverini U (2004) Frisch scheme–based algorithms for EIV identification. *Proc. of the 12th IEEE Mediterranean Conference on Control and Automation*, Kusadasi, Turkey
18. Diversi R, Guidorzi R, Soverini U (2006) *Proc. of the 17th International Symposium on Mathematical Theory of Networks and Systems*: 391–395, Kyoto, Japan
19. Diversi R, Guidorzi R, Soverini U (2003) *Preprints of the 13th IFAC Symposium on System Identification*: 1993–1998, Rotterdam, The Netherlands
20. Guidorzi R, Diversi R, Soverini U (2003) *Automatica*, 39:281–289
21. Guidorzi R, Diversi R, Soverini U, Valentini A (2004) A noise signature approach to fault detection and isolation. *Proc. of the 16th International Symposium on Mathematical Theory of Networks and Systems*, Leuven, Belgium
22. Schachermayer W, Deistler M (1998). *Systems & Control Letters*, 34:101–104
23. Guidorzi R, Stoian A (1994) *Proc. of the 10th IFAC Symposium on System Identification*, 3:171–173, Copenhagen, Denmark
24. Guidorzi R, Pierantoni M (1995) *Proc. of the XII Int. Conf. on Systems Science*: 114–120, Wroclaw, Poland
25. Guidorzi R (1996) Identification of multivariable processes in the Frisch scheme context. *Proceedings of the MTNS '96*, St. Louis, Missouri
26. Guidorzi R, Diversi R (2006) *Proc. of the 17th International Symposium on Mathematical Theory of Networks and Systems*: 530–535, Kyoto, Japan
27. Guidorzi R (1989) *Kybernetica*, 25:233–257. Part II, Kybernetica 25:386–407
28. Guidorzi R, Soverini U, Diversi R (2002) Multivariable EIV identification. *Proc. of the 10th IEEE Mediterranean Conference on Control and Automation*, Lisboa, Portugal
29. Abed-Meraim K, Qiu W, Hua Y (1997) *Proc. of the IEEE*, 85:1310–1322
30. Tong L, Perreau S (1998) *Proc. of the IEEE*, 86:1951–1968

31. Diversi R, Guidorzi R, Soverini U (2005) *Signal Processing*, 85:215–225
32. Guidorzi R, Diversi R, Soverini U (2007) *Signal Processing*, 87:654–664
33. Kay SM (1979) *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27:478–485
34. Kay SM (1980) *IEEE Trans. on Acoustics, Speech and Signal Processing*, 28:292–303
35. Diversi R, Soverini U, Guidorzi R (2005) A new estimation approach for AR models in presence of noise. *Preprints of the 16th IFAC World Congress*, Prague, Czech Republic
36. Diversi R, Guidorzi R, Soverini U (2005) *Proc. of the 44th IEEE Conference on Decision and Control and 8th European Control Conference*: 4146–4151, Seville, Spain
37. Bobillet W, Grivel E, Guidorzi R, Najim M (2005) Noisy speech de-reverberation as a SIMO system identification issue. *Proc. of the IEEE Workshop on Statistical Signal Processing*, Bordeaux, France
38. Diversi R, Guidorzi R, Soverini U, Bobillet W, Grivel E, Najim M (2006) A new optimal smoothing approach for AR + noise models and application to single–microphone speech enhancement. *Proc. of the 2nd IEEE International Symposium on Communications, Control and Signal Processing*, Marrakech, Morocco
39. Bobillet W, Grivel E, Najim M, Diversi R, Guidorzi R, Soverini U (2006) Errors–in–variables based identification of autoregressive parameterers for speech enhancement using one microphone. *Proc. of the 2nd IEEE International Symposium on Communications, Control and Signal Processing*, Marrakech, Morocco

Multivariable Feedback with Geometric Tools

Giovanni Marro

Dipartimento di Elettronica, Informatica e Sistemistica, Università di Bologna,
Viale Risorgimento 2, 40136 Bologna, Italy
gmarro@deis.unibo.it

1 Introduction

I remember Toni for his personal warmth and innate respect; for a friendship you could feel. He would give important and less important bits of advice, throwing in the odd comment, or dropping a hint, never insisting, to show he meant no offence. I remember how attached he was to us in Bologna. During the first research doctorate the university seats of Bologna, Florence and Padua had formed a consortium, but he would still take the train to Bologna to hold his lectures because he liked to breathe the atmosphere of the city and walk across its main square, Piazza Maggiore.

I keep in my library a very simple, clear and attractive small book with a nice friendly hand-written dedication to me by Toni [1]. The book presents feedback in simple terms but reflects Toni's in-depth knowledge, highlighting the presence of feedback in many branches of the scientific world. Inspired by this book and recalling that many years ago Toni probably read our first paper on geometric tools in Italian [2], I decided to dedicate this contribution to feedback in geometric terms. In this short monograph I will try to explain how these tools not only enable a neat extension to the multivariable case of the most basic features of feedback control for single variable systems, including the internal model of the exosystem, hence steady-state robustness, but they are just within arm's reach using a certain number of algorithms available in a specific Geometric Approach toolbox for Matlab®. Although this approach sounds didactic and somewhat out of standard, it nevertheless contains an original idea: A proposal to extend model-following control to non minimal phase systems. Being a non optimal, but only pseudo-optimal solution, the proposal explains the detailed illustration of this toolbox and its use in the synthesis of multivariable control systems.

The geometric approach to system analysis and control suffered from a very slow and inconsistent growth for about four decades, with papers by numerous authors in many different styles, often aiming to impress readers (or reviewers) with intricate mathematics rather than giving them the most direct insight and feeling. This makes it very difficult now to gain a clean and simple outlook of the features and possibilities of the geometric environment from these papers. In most cases neither the authors nor the reviewers of these papers perceived that numerous, apparently different, problems could be framed together and led to the same unifying solution. Duality, in particular, detected early on as being a basic feature of the geometric approach [3], was almost never used or pointed out.

Notation

Through this contribution, $\mathbb{R}^n$ stands for the set of all n-tuples of real numbers and $\mathbb{C}_g$ denotes the open left-half complex plane. Script capitals, like $\mathcal{X}$, denote vector spaces and subspaces, while italic capitals, like A, denote matrices and linear maps. dim$\mathcal{X}$ and $\mathcal{X}^\perp$ denote the dimension and the orthogonal complement of subspace $\mathcal{X}$. The image, the null space, the transpose, the inverse and the pseudoinverse of a generic matrix A are denoted by $\operatorname{im} A$, $\ker A$, A^T, A^{-1} and $A^{\#}$ respectively, while the number of rows and columns of A are denoted by mA and nA.

Regarding the standard notation of the geometric approach [4, 5, 6], $\mathcal{V}$ and $\mathcal{S}$ stand for a generic controlled invariant and a generic conditioned invariant, $\mathcal{Z}(\mathcal{V})$ for the internal unassignable spectrum of $\mathcal{V}$. Referring to a given continuous or discrete-time LTI system $\varSigma \equiv (A, B, C, D)$, $\mathcal{V}^*$ and $\mathcal{S}^*$ denote the maximum output nulling controlled invariant and the minimum input containing conditioned invariant of $\varSigma$, while $\mathcal{Z}(\varSigma) = \mathcal{Z}(\mathcal{V}^*)$ is used for the set of all invariant zeros of $\varSigma$. For a quick review of numerous basic problems presented with this notation, refer to [7].

2 A short literature on model following

The *model following problem* (MFP), i.e. that of synthesizing a state or output feedback law for a given plant in order to make the impulse response matrix of the compensated system exactly equal to that of a prespecified model, has been studied for different classes of systems since the early 1970's [8, 9, 10, 11]. These early investigations were primarily concerned with problem solvability under various hypotheses.

In [12] the MFP was precisely stated by Morse, and necessary and sufficient geometric conditions for its resolution using algebraic feedforward compensation and a state feedback with some dynamics added were provided. The internal stability problem was discussed, but only a few preliminary results were given.

Several papers extended and refined Morse's early results. In [13] the set of all stable solutions to the MFP was presented in a parametric form. In [14] the extension to quadruples (A, B, C, D) was proposed, and necessary and sufficient conditions under which solutions exist using regular output feedback were provided. In [15] a new geometric proof for a solvability condition of the MFP expressed in terms of the infinite zero structure was presented.

Many variations of the original MFP were also proposed in the literature in the last decades. The most remarkable examples are the *partial state-feedback* MFP and the *partial* MFP, (see respectively [16] and [17] and the references therein).

Recently, in [18] a new procedure to synthesize minimal-order regulators for exact model following by output feedback with stability has been proposed. The approach, completely embedded in the geometric framework, exploits the properties of self-bounded controlled invariant subspaces and provides an effective treatment of nonminimum-phase systems.

Nevertheless the design strategy in [18] requires that all unstable invariant zeros of the nonminimum-phase system are replicated in the model, thus significantly constraining the performance of the resulting overall feedforward (or feedback) system.

In order to solve this problem, an original design layout for the solution of the feedforward and feedback model following problem for nonminimum-phase plants is presented in this contribution. This work was particularly inspired by the recent investigations on H_2 disturbance decoupling (see [19, 20]) and it proposes to solve the MFP for a new controlled system where the output matrix is replaced with an equally sized matrix allowing the H_2-optimal control problem to be approached as a disturbance decoupling while keeping the same relative degree and the steady-state gain of the original system. The new matrix is derived by applying the standard geometric approach tools to the Hamiltonian system instead of the original plant and using some refinements to adapt the solution of the H_2-optimal decoupling problem to the requirements of model following. The proposed constructive procedure to determine the output matrix can be readily implemented using the basic routines of the geometric approach.

Note that in [21, 22] geometric methods were also presented for the resolution of singular H_2 control problems. However the procedure adopted therein is less direct and very far from the standard geometric approach computational environment.

The rest of the work is organized as follows. In Sect. 3 the exact feedforward model following problem as presented in [18] is reviewed. In Sect. 4 the H_2-optimal decoupling problem is stated and its geometric solution is briefly recalled. In Sect. 5 the H_2-optimal model following problem is discussed and a convenient pseudo optimal solution is derived. In Sect. 6 a numerical example is presented in some detail. in Sect. 7 some information on the geometric algorithms is given, while in Sect. 8 the major contributions of the paper are summarized and some concluding remarks are provided.

3 A review of the exact model following problem

The purpose of this section is to briefly recall the various aspects of the exact model following problem and the corresponding solutions with the geometric approach tools.

A Exact feedforward model following

The most elementary model following problem is stated in the following terms.

Problem 1. (Exact feedforward model following) Refer to Fig. 1. Given a *plant* Σ described by

$$\begin{aligned} \dot{x}(t) &= A\,x(t) + B\,u(t) \\ y(t) &= C\,x(t) \end{aligned} \tag{1}$$

where $x \in \mathcal{X} = \mathbb{R}^n$, $u \in \mathbb{R}^p$ and $y \in \mathbb{R}^p$ denote the state, the control input and the controlled output, respectively, and a *model* Σ_m described by

$$\begin{aligned} \dot{x}_m(t) &= A_m\,x_m(t) + B_m\,h(t) \\ y_m(t) &= C_m\,x_m(t) \end{aligned} \tag{2}$$

where $x_m \in \mathcal{X}_m = \mathbb{R}^q$, $h \in \mathbb{R}^p$ and $y_m \in \mathbb{R}^p$ denote the state, the exogenous input and the output, design a linear dynamic stable feedforward compensator $\Sigma_c \equiv (A_c,\,B_c,\,C_c,\,D_c)$ such that the forced evolution of y is equal to that of y_m for every admissible (piecewise continuous) input h.

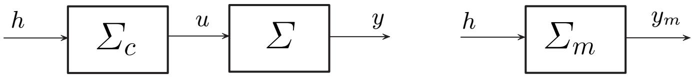


Fig. 1. Feedforward model following.

For the sake of simplicity both Σ and Σ_m are assumed to be minimal, stable, with no common poles and zeros, square, left and right invertible, (i.e., such that $\mathcal{V}^* \cap \mathcal{S}^* = \{0\}$, $\mathcal{V}^* + \mathcal{S}^* = \mathcal{X}$ and $\mathcal{V}_m^* \cap \mathcal{S}_m^* = \{0\}$, $\mathcal{V}_m^* + \mathcal{S}_m^* = \mathcal{X}_m$), and steady-state completely output controllable (i.e., with the matrices $C\,A^{-1}\,B$ and $C_m\,A_m^{-1}\,B_m$ finite and nonsingular)[1].

Theorem 1. *Problem 1 is solvable if*

$$\rho(\Sigma) \leq \mu(\Sigma_m) \tag{3}$$

$$\mathcal{Z}(\Sigma) \subseteq \mathbb{C}_g \tag{4}$$

[1] The left invertibility assumption can be overcome by using squaring down (see the routine *extendf* in Sect. B).

where ρ denotes the global relative degree of Σ and μ the minimum delay of Σ_m [2].

Proof. The relative degree ρ is the minimum value of the time derivative of any output function of Σ reproducible starting from the zero state by applying a suitable piecewise continuous input function. The minimum delay μ is the minimum value of the time derivative of any possible ouput function of Σ_m corresponding to a piecewise continuous input function. Hence condition (3) guarantees structural feasibility of exact model following. Condition (4) implies condition (8) of Corollary 2 below, that guarantees internal stability. ■

Conditions (3)-(4) are presented as only sufficient, but indeed they also are almost necessary, i.e., generally also necessary, except for some pathological cases whose investigation may be interesting, but beyond the introductory scope of this contribution.

Feedback solutions have been proposed in most of the numerous treatments of the model following problem available in the literature to include the overall stability requirement when Σ is unstable. However stabilization of Σ by feedback through a full or reduced order observer is unrelated to solvability of the model following problem. In fact, if Σ is not stable but stabilizable and detectable (recall that it has been assumed to be minimal in this case) it is possible to use a stabilizing feedback through a full or reduced order observer Σ_o as shown in Fig. 2, thus replacing it with a new stable system denoted here and thereafter by Σ, Σ_o. In fact, under the minimality assumption, Σ is both reachable and observable.

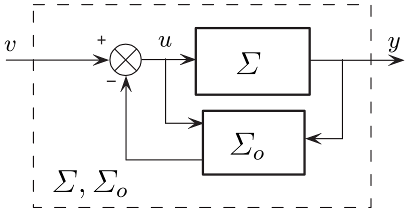


Fig. 2. Using a stabilizer.

Definition 1. *(Invariant zero structure) Refer to Fig. 3, representing an exosystem Σ_e, described by the differential equation $\dot{v}(t) = W\,v(t)$ with initial state v_0, algebraically connected to Σ. W is an* invariant zero structure *of Σ*

[2] Recall that the global relative degree is computed in geometric terms as the minimum value of i such that $\mathcal{V}^* + \mathcal{S}_i = \mathcal{X}$, where $\mathcal{S}_i$ denotes the i-th subspace in the well-known sequence for computing $\mathcal{S}^*$. The minimum delay of a triple is defined as the minimum value of i such that $C\,A^i\,B$ is nonzero.

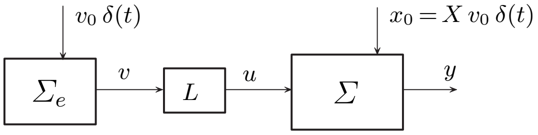


Fig. 3. The definition of invariant zero structure.

if and only if there exist nonzero matrices L and X such that the evolution $x(t)$ of the state of Σ from $x_0 = X\,v_0$ and with the control input $u(t) = L\,v(t)$ completely belongs to $\ker C$.

Let us refer again to Fig. 3 and assume that W and A have no common eigenvalues. By solving the Sylvester equation

$$A\,X - X\,W = -B\,L$$

where now L is given, we obtain an initial state $X\,v_0$ that makes Σ to behave as an algebraic connection (with all its modes equal to zero, already in the steady-state condition), described by $y(t) = -\,(C\,A^{-1}\,B\,L)\,v(t)$, where the matrix in brackets on the right can be made nonsingural by a suitable choice of L, owing to the left anf right invertibility of Σ. The following lemma is crucial to point out the pole-zero structure of the compensator Σ_c in the model following problem.

Lemma 1. *Let the LTI system $\Sigma \equiv (A, B, C)$ be stable, left and right invertible, steady-state output controllable, completely observable but not completely reachable and with the eigenvalues of the reachable subsystem different from those of the unreachable subsystem. Then, the eigenvalues of the unreachable subsystem are invariant zeros of Σ.*

Proof. Without any loss of generality, we can assume

$$A = \begin{bmatrix} A_{12} & A_{12} \\ O & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ O \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}$$

Note that the reachable subsystem $\Sigma_1 \equiv (A_1, B_1, C_1)$ must be left and right invertible by assumption, since (A_{22}, C_2) is neutral with respect to control. Let us refer to Fig. 4, representing the block diagram corresponding to $W = A_{22}$, $x_{0,2} = v_0$ and to choosing X_1 such that the modes of A_{11} are all zero at the initial time. This choice depends on L and A_{12}. By means of a suitable choice of L (with consequent adjustemnt of X_1) it is possible to have a complete control on the linear combination of the modes of W al the output, i.e., to attain the condition $y = 0$. Hence matrix A_{22} is an invariant zero structure of Σ owing to Definition 1. ∎

Corollary 1. *The overall system whose block diagram is shown in Fig. 2, with algebraic state feedback F through an observer, has the eigenvalues of the*

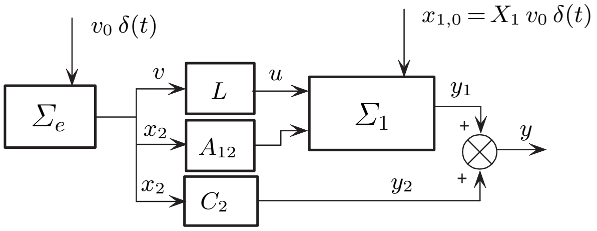


Fig. 4. The structure in Fig. 3 for an unreachable system.

observer and those of $A+BF$ as poles and the invariant zeros consisting of the invariant zeros of Σ and the poles of Σ_o. However, due to lack of controllability of the observer, the overall system is not minimal, but it is input-output equivalent to a n-th order system with the poles of $A+BF$ as poles and with the same invariant zeros as Σ.

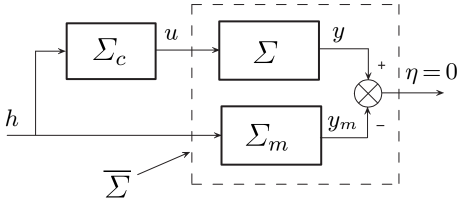


Fig. 5. Model following as a MDDPS.

It is well known that Problem 1 is a particular case of the standard (exact) measurable disturbance decoupling problem with stability (MDDPS) for an extended system $\overline{\Sigma}$ also including the model, as shown in Fig. 5. It is worth recalling herein the statement and the solution to this problem in geometric terms.

Problem 2. (Measurable signal decoupling with stability) Refer to Fig. 6, where Σ is assumed to be modelled by

$$\begin{aligned} \dot{x}(t) &= A\,x(t) + B\,u(t) + H\,h(t) \\ y(t) &= C\,x(t) \end{aligned} \tag{5}$$

Design a linear dynamic stable feedforward compensator $\Sigma_c \equiv (A_c,\, B_c,\, C_c,\, D_c)$ such that the forced evolution of y is identically zero.

Conditions ensuring the solvability of Problem 2 are stated by the following theorem, that is well known in the literature and is here recalled without proof.

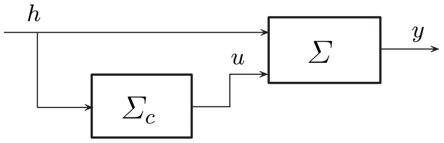


Fig. 6. Measurable signal decoupling.

Theorem 2. *Problem 2 is solvable if and only if*

$$\mathcal{H} \subseteq \mathcal{V}_g^* + \mathcal{B} \tag{6}$$

where $\mathcal{H} = im\, H$, $\mathcal{B} = im\, B$ *and, as it is usual,* $\mathcal{V}_g^*$ *denotes the restriction of* $\mathcal{V}^*$ *to include only the stable zeros of* Σ.

If Σ is minimum-phase clearly $\mathcal{V}_g^* = \mathcal{V}^*$. The conditions stated in the following corollary are equivalent to (6), but refer to an $(A, \mathcal{B})$-controlled invariant $\mathcal{V}_m \subseteq \mathcal{V}_g^*$, more convenient than $\mathcal{V}_g^*$ as a resolvent. In fact, using $\mathcal{V}_m$ as a resolvent instead of $\mathcal{V}_g^*$ may yield a compensator Σ_c with less state dimension.

Corollary 2. *Problem 2 is solvable if and only if*

$$\mathcal{H} \subseteq \mathcal{V}_{(\mathcal{B},\mathcal{C})}^* + \mathcal{B} \tag{7}$$

$$\mathcal{Z}(\mathcal{V}_m) \subseteq \mathbb{C}_g \tag{8}$$

with

$$\mathcal{V}_m := \mathcal{V}_{(\mathcal{B},\mathcal{C})}^* \cap \mathcal{S}_{(\mathcal{B}+\mathcal{H},\mathcal{C})}^* \tag{9}$$

Condition (7) was first stated in [23], while (8) is due to [24]. Note that solvability of Problem 1 as stated in Theorem 1 and of Problem 2 as stated in Corollary 2, like that of many other problems in the geometric control theory, requires both a *structural condition*, like (3) and (7), and a *stabilizability condition*, like (4) and (8). The subspace $\mathcal{V}_m$ defined in (9) is simply the reachable subspace on $\mathcal{V}_{(\mathcal{B},\mathcal{C})}^*$ with both inputs h and u.

Remark 1. Conditions (3)-(4), only sufficient but more directly intelligible as straightforward extensions of the SISO case, imply (7)-(8), that they are necessary and sufficient, but stated in strict geometric terms. Note that in (3)-(4) the *vector* (i.e., componentwise) relative degree and the vector minimum delay may be used instead of the *global* relative delay, thus achieving more definite conditions. However this is not valid in the extension to the non-minimum phase case considered in Sect. 5.

B From feedforward to feedback model following

The main contribution to exact model following is a very straightforward solution for the feedback connection shown in Fig. 7. In fact, it has also been pointed out in [18] that exact feedback model following can be obtained with some simple manipulation from the feedforward solution as stated in Problem 1 and under the conditions established in Theorem 1. Moreover, a multiple internal model with simple or multiple poles at the origin can be imposed in the feedback regulator $\varSigma_c$ if $\varSigma_m$ is suitably chosen with this aim. This makes the model following approach to multivariable regulation problems particularly attractive from the standpoint of engineering practice.

Consider the modified scheme shown in Fig. 8. Note that replacing the feed-

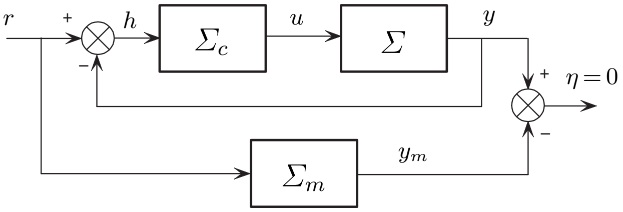


Fig. 7. Exact feedback model following.

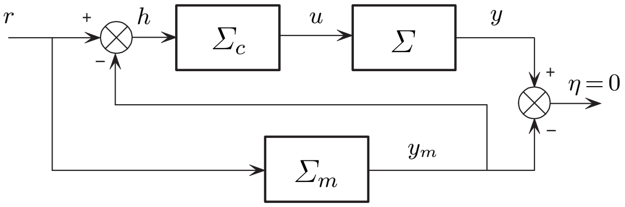


Fig. 8. A structurally equivalent connection.

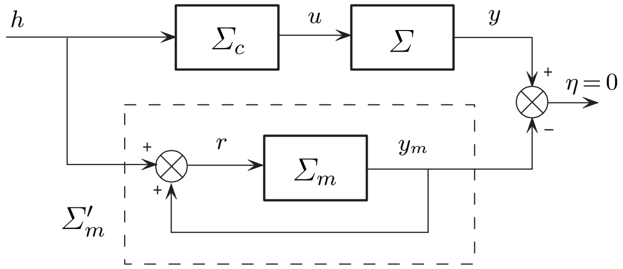


Fig. 9. Another structurally equivalent connection.

back connection in Fig. 7 with that shown in Fig. 8 does not affect the structural properties of the system since the signals y and y_m are still equal for all t, but may affect stability. The new block diagram represents a feedforward model following problem. In fact, note that h is obtained as the difference of r (applied to the input of the model) and y_m (the output of the model). This corresponds to the parallel connection of Σ_m and a diagonal algebraic system with gain -1, that is invertible, having zero relative degree. Its inverse is Σ_m with a feedback connection through a diagonal algebraic system with unit gain, as shown in Fig. 9.

It is thus evident that from a structural point of view, the modified feedback model following in Fig. 7 is equivalent to a modified feedforward problem which refers to the model $\Sigma'_m \equiv (A_m + B_m C_m, B_m, C_m)$. It is then possible to bring back the feedback model following problem to the feedforward problem discussed in Sect. A by means of the above algebraic manipulations, that guarantee existence of equal solutions of the differential equations describing the overall two systems (called herein *structural equivalence* of the two involved systems), but not guarantee stability. In fact, the condition $\eta(t) = 0$ for all $t \geq 0$ in the modified system may be obtained as the difference of diverging signals y and y_m. However, stability is recovered when going back to the original feedback connection represented in Fig. 7, that ensures that all the modes present in the impulse response of Σ_m are stable.

Adding row-by-row decoupling and including an internal model

Row-by-row decoupling is a standard problem of control theory. It is also called Morgan's problem in the literature, from the name of the first author who approached it using the state-space formulation [25]. Morgan's solution was based on state algebraic feedback and algebraic precompensation. More recently the problem has been discussed and solved in geometric terms [26].

The previous solutions, that appear very intricate, are overcome by the approach presented herein, since row-by-row decoupling can be very easily obtained in feedforward or feedback model following by simply assuming a model Σ_m consisting of q independent SISO systems with suitable relative degrees.

It is also possible to include in every feedback loop concerning Σ'_m (Fig. 9) poles at the origin with arbitrary multiplicity γ by assuming these SISO systems with transfer functions having the γ coefficients corresponding to the lower powers of s (including the constant term) equal in the numerator and denominator polynomials. These poles at the origin are repeated as an *internal model* in the compensator, so that both Σ'_m and the compensator may be unstable systems. However, as previously pointed out, the original feedback connection in Fig. 7 is stable since it has the same poles as Σ_m.

4 From exact decoupling to $\mathbf{H_2}$-optimal decoupling

Let us consider again Problem 2 and assume that Σ is non-minimum phase, so that condition (6) or (7)-(8) may be (and in general are) not satisfied. In this case a possible resort is to consider the following problem as "the best we can do".

Problem 3. (H_2-optimal decoupling with stability) Refer again to Fig. 6, where Σ is assumed to be modelled by (5). Design a dynamic stable feedforward compensator $\Sigma_c \equiv (A_c, B_c, C_c, D_c)$ such that the H_2 norm of the overall system from input h to output y is minimal.

Let us recall that the H_2-norm of a triple $\Sigma \equiv (A, B, C)$ is defined as

$$\|\Sigma\|_2 = \sqrt{\operatorname{tr}\Big(\int_0^\infty g(t)\, g^T(t)\, dt\Big)} \tag{10}$$

where "tr" denotes the trace of a matrix and $g(t)$ denotes the impulse response of the system.

Problem 3 is traced back to Problem 2. In fact, by using the standard optimal control framework it is easily shown that "almost exact decoupling" corresponds to exact decoupling for the Hamiltonian system $\widehat{\Sigma}$ defined by

$$\begin{aligned} \dot{\hat{x}}(t) &= \widehat{A}\,\hat{x}(t) + \widehat{B}\,u(t) \\ 0 &= \widehat{C}\,\hat{x}(t) \end{aligned} \quad , \qquad \hat{x} = \begin{bmatrix} x \\ p \end{bmatrix} \tag{11}$$

with

$$\widehat{A} = \begin{bmatrix} A & 0 \\ -C^T C & -A^T \end{bmatrix}, \quad \widehat{B} = \begin{bmatrix} B \\ O \end{bmatrix}, \qquad \widehat{C} = \begin{bmatrix} O & B^T \end{bmatrix} \tag{12}$$

It is well known ([27] (adapted from discrete to continuous-time systems) that if Σ is left invertible, $\widehat{\Sigma}$ is both right and left invertible, and if Σ is left and right invertible, $\mathcal{Z}(\widehat{\Sigma})$ consists of all the elements of $\mathcal{Z}(\Sigma)$ and their opposites. The set of equations (11) can be considered to refer to an LTI dynamic system whose output is constrained to be zero. Denote by $\widehat{\mathcal{V}}^*$ the maximum output nulling controlled invariant subspace of $\widehat{\Sigma}$, whose internal eigenvalues are the elements of $\mathcal{Z}(\widehat{\Sigma})$, and by

$$\widehat{\mathcal{V}}_g^* = \operatorname{im} \begin{bmatrix} V_1 \\ P_1 \end{bmatrix} \tag{13}$$

its restriction to the stable internal eigenvalues. Clearly, the dimension of $\widehat{\mathcal{V}}_g^*$ is equal to that of $\mathcal{V}^*$. It follows that minimizing the H_2 norm of the overall system in Fig. 6 is equivalent to the perfect decoupling problem for the triple $(\widehat{A}, \widehat{B}, \widehat{C})$. The necessary and sufficient condition (6), adapted to this case, is

$$\mathrm{im}\begin{bmatrix} H \\ O \end{bmatrix} \subseteq \widehat{\mathcal{V}}_g^* \tag{14}$$

Thus we have proven the following theorem.

Theorem 3. *Problem 3 is solvable if and only if*

$$\mathcal{H} \subseteq \mathcal{V}_1 + \mathcal{B} \tag{15}$$

where $\mathcal{V}_1 := \mathit{im}\, V_1$ *with matrix* V_1 *defined in (13).*

Remark 2. Owing to the structure of the Hamiltonian system (12) $\mathcal{V}_1$ is clearly an $(A, \mathcal{B})$-controlled invariant, not necessarily contained in $\ker C$.

Remark 3. Under the assumed condition that Σ is both left and right invertible the inclusion $\mathcal{Z}(\mathcal{V}^*) \subseteq \mathcal{Z}(\mathcal{V}_1)$ holds, so that if Σ is minimum-phase, inclusion (6) implies (15) and $\mathcal{V}_1$ is contained in $\ker C$. In other terms, Problem 3 can be regarded as a straightforward extension of Problem 2.

5 H$_2$-pseudo optimal model following

According to the above considerations, when Σ is nonminimum-phase, it appears to be natural to replace the exact model following problem (Problem 1) with a more general H_2-optimal model following problem. This can easily be stated as the problem of minimizing the H_2-norm of the overall system in Fig. 5 from input h to output η. This solution is feasible, but has several drawbacks in practice. The most important are considered in the following list.

1. Zero steady-state tracking error, that is an essential requirement in every control system design, is not guaranteed. In fact, the H_2 norm as defined in (10) refers to the impulse response, not to the step response.
2. Contrary to the exact case, the relative degree condition (3) does not imply the necessary and sufficient condition (6). Actually, it can be proved that when $\rho(\Sigma) \geq 2$ the problem does not admit solutions of the type sketched in Fig. 5 (since Σ_c cannot generate distributions).
3. Extension to the feedback case, including row-by-row decoupling and inclusion of an internal model in the compensator, as described in Sect. B, is not possible, since the output η of the overall system is not any longer identically zero.

In order to solve these problems, a suitable *pseudo optimal* solution to the problem is proposed herein.

A Introducing a new output matrix $\mathbf{C_1}$

Refer to the modified feedforward scheme shown in Fig. 10. Replace matrix C with another matrix C_1 such that the new system Σ_1, corresponding to the triple (A, B, C_1), is minimum-phase, the ℓ_2 norm of output y_1 is minimal for the maximum set of initial states compatible with a control without distributions and the steady-state gain from v to y_1 is equal to that from v to y. This auxiliary output may be provided by a full or reduced order observer Σ_o connected to the input and output of the original plant (see Fig. 2). Σ_o could also be used to stabilize the system or give it a more favourable pole placement for optimal (H_2 or H_∞) rejection of inaccessible disturbances directly acting on Σ. C_1 can be computed as shown in the constructive proof of the following theorem.

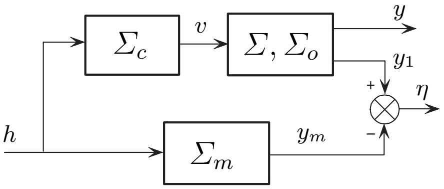


Fig. 10. Modified feedforward model following.

Theorem 4. *Refer to system Σ and assume that it is nonminimum-phase, with no purely imaginary invariant zeros. A unique matrix C_1 exists such that:*

1 - the exact disturbance decoupling problem with state feedback for any input disturbance matrix H such that $\mathrm{im}H \subseteq \mathcal{V}_1^*$, where $\mathcal{V}_1^*$ is referred to the triple (A, B, C_1), corresponds to a minimal H_2-norm solution for the triple (A, B, C);

2 - the number of outputs of (A, B, C_1) is equal to those of (A, B, C);

3 - the triple (A, B, C_1) is minimum-phase;

4 - the triple (A, B, C_1) has the same global relative degree as (A, B, C);

5 - the steady-state gain of (A, B, C_1) is equal to that of (A, B, C).

Proof. Refer to equation (13). $\mathcal{V}_1^* := \mathrm{im}V_1$ is an (A, B)-controlled invariant that, by construction, satisfies property 1 of the statement. The dimension of $\mathcal{V}_1^*$ is equal to that of $\mathcal{V}^*$, but $\mathcal{V}_1^*$ is not contained in $\ker C$ if and only if Σ is nonminimum-phase. Let us refer the symbol $\mathcal{S}^*$ to the triple (A, B, C) and $\mathcal{S}_1^*$ to (A, B, C_1). It will be shown now that

$$\mathcal{V}_1^* \cap \mathcal{S}^* = \{0\} \tag{16}$$

From (12) and from the inclusion $\widehat{\mathcal{V}}_g^* \subseteq \widehat{\mathcal{V}}^*$, it follows that

$$\operatorname{im} B \subseteq (\operatorname{im} P_1)^{\perp} \tag{17}$$

Moreover, from the cost formulation of the optimal control problem, it is well known that $V_1^T P_1 \neq 0$ since in our case the cost (the H_2 norm) is nonzero, and consequently

$$\operatorname{im} V_1 \subseteq (\operatorname{im} P_1)^{\perp} \tag{18}$$

From (17) and (18), one obtains $\operatorname{im} V_1 \cap \operatorname{im} B = \{0\}$, hence

$$\mathcal{V}_1^* \cap \mathcal{S}_1^* = \{0\} \tag{19}$$

To prove (16), it still remains to show that $\mathcal{S}_1^* = \mathcal{S}^*$. From the algorithm for the computation of the minimum input containing conditioned invariant [5], it follows that

$$\mathcal{S}_1^* \supseteq \mathcal{S}^* \tag{20}$$

while from a simple consideration on subspaces dimensions, i.e. $\dim \mathcal{S}_1^* = \dim \mathcal{S}^*$, from (20) it follows that $\mathcal{S}_1^* = \mathcal{S}^*$. This completes the proof of (16).
Let $\mathcal{C} := \ker C$. To conclude the proof of the theorem, consider the identity

$$\mathcal{C} = \mathcal{V}^* + (\mathcal{S}^* \cap \mathcal{C}) \tag{21}$$

obtained by intersection of $\mathcal{V}^* + \mathcal{S}^* = \mathcal{X}$ with $\mathcal{C}$, and set

$$\mathcal{C}_1 = \mathcal{V}_1^* + (\mathcal{S}^* \cap \mathcal{C}) \tag{22}$$

Owing to (16) and the dimensions of $\mathcal{V}^*$ and $\mathcal{V}_1^*$ being equal, the subspaces $\mathcal{C}_1$ and $\mathcal{C}$ have the same dimension. Let $\bar{C}_1$ be the transpose of any basis matrix of the orthogonal complement of the subspace $\mathcal{C}_1$ defined in (22), hence having $\mathcal{C}_1$ as its kernel. The new matrix C_1 is defined as

$$C_1 = G\,\bar{G}^{-1}\,\bar{C}_1 \tag{23}$$

where G denotes the steady-state gain of the triple (A, B, C) and $\bar{G}$ that of the triple $(A, B, \bar{C}_1)$.

- Property 1 in the statement is satisfied because the maximum (A, B)-controlled invariant contained in $\ker C_1 = \mathcal{C}_1$ is $\mathcal{V}_1^*$, since $\mathcal{V}_1^* \cap \mathcal{S}^* = \{0\}$.

- Property 2 is satisfied since $\mathcal{C}$ and $\mathcal{C}_1$ have the same dimension.

- Property 3 is satisfied since the invariant zeros of (A, B, C_1) are the internal eigenvalues of $\mathcal{V}_1^*$, stable by construction.

- Property 4 is satisfied since both the triples (A, B, C_1) and (A, B, C) have $\mathcal{S}^*$ as the minimum conditioned invariant containing $\operatorname{im} B$ and, since both $\mathcal{V}_1^*$ and $\mathcal{V}^*$ have null intersection with $\mathcal{S}^*$, the dimensions of $\mathcal{V}_1^* + \mathcal{S}_i$ and $\mathcal{V}^* + \mathcal{S}_i$ are each other equal for all i.

- Property 5 is due to relation (23). Uniqueness of matrix C_1 is due to uniqueness of all the subspaces involved in the constructive procedure described above and to relation (23) that sets a unique value for the steady-state gain of (A, B, C_1). ■

Remark 4. Refer to the block diagram in Fig. 10 and recall Property 1. If the auxiliary output y_1 is obtained through an observer, connected to Σ as shown in Fig. 2, due to lack of controllability of the observer, the overall system $(\Sigma, \Sigma_o$ in the figure) is not minimal, but it is input-output equivalent to a n-th order system with the poles of $A + BF$ as poles and with the same invariant zeros as Σ_1, hence minimum-phase.

Remark 5. It can be proven that the original triple (A, B, C) and the corresponding minimum-phase triple (A, B, C_1) have equal H_2 norms. However the proof of this property is beyond the scope and space of this treatise.

6 A worked example

The purpose of this section is to show how the tools of the Geometric Approach toolbox can be used for the complete synthesis of a multivariable feedback regulator for a nonminimum-phase controlled system[3].

Suppose that the plant Σ is defined by

$$A = \begin{bmatrix} -5 & 1.00 & 2.00 & 3.00 & 4.00 \\ 0 & -8.95 & -6.45 & 0 & 0 \\ 0 & 2.15 & -0.35 & 0 & 0 \\ 0 & -10.89 & -40.94 & -16.10 & -7.95 \\ 0 & 8.17 & 28.87 & 7.07 & -0.20 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 10 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 5 & 6 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

First, we check whether the conditions stated in Problem 1 are satisfied by the controlled system Σ.

```
>> V=vstar(A,B,C,D);
>> S=sstar(A,B,C,D);
>> Yc=sums(V,S);
>> Uo=ints(V,S);
>> p=eig(A);
>> z=gazero(A,B,C,D);
>> rho=reldeg(A,B,C,D);
```

[3] The main commands of the toolbox are detailed in Sect. 7. The software is freely downloadable from the web page:
`http://www3.deis.unibo.it/Staff/FullProf/GiovanniMarro/downloads.htm`
To understand the synthesis procedure considered in this section, refer to the *help* command of the routines as they are cited herein.

The plant Σ is stable, since $p^T = \left[-5 \;\; -2.5 \;\; -10.795 \;\; -6.7 \;\; -5.505 \right]$, right-invertible, being Y_c a full rank matrix (hence with $\operatorname{im} Y_c = \mathcal{X}$), left invertible, being $U_o = \{0\}$, nonminimum-phase, since $z^T = \left[14.7907 \;\; -9.7907 \right]$, and with relative degree $\rho = 2$. From the step response shown in Fig. 11 it appears that Σ is very far from being row-by-row decoupled.

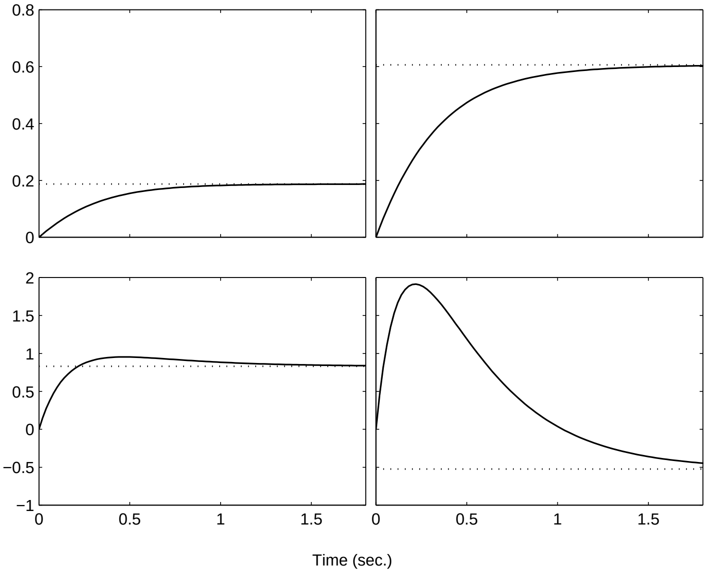


Fig. 11. The step response of the plant Σ.

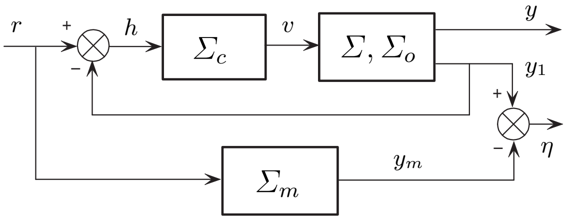


Fig. 12. Modified feedback model following.

For this system, we want to design a feedback regulator of the type shown in Fig. 12, with a partial-order observer Σ_o (with state dimension $nA - nC = 3$) included in the feedback loop, a model Σ_m SISO-parallel (hence steady-state row-by row decoupling), and a type one (i.e., with single poles at the origin)

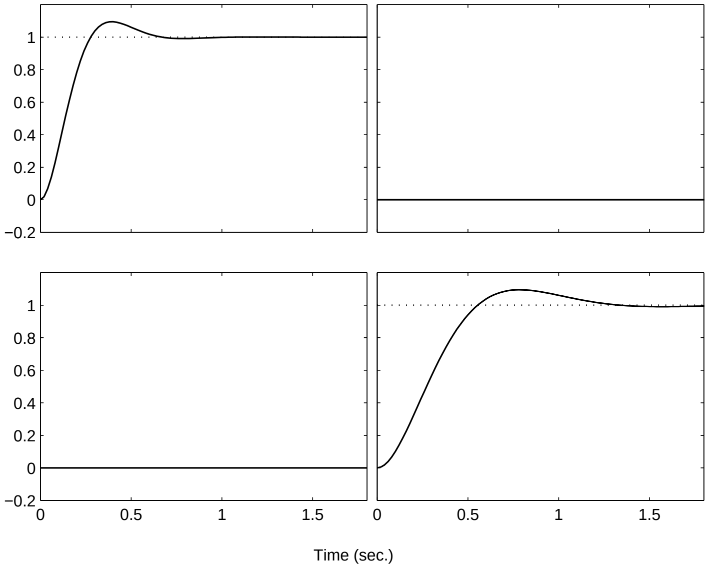


Fig. 13. The step response of the model Σ_m.

internal model, to achieve steady-state robustness. Let assume for the two rows of the model the transfer functions

$$G_1(s) = \frac{1}{s^2 + 1.2\,s + 1}\,, \quad G_2(s) = \frac{0.25}{s^2 + 0.6\,s + 0.25}$$

corresponding to the state space realization

$$A_m = \begin{bmatrix} 0 & -100 & 0 & 0 \\ 1 & -12 & 0 & 0 \\ 0 & 0 & 0 & -25 \\ 0 & 0 & 1 & 6 \end{bmatrix}, \quad B_m = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$

$$C_m = \begin{bmatrix} 0 & 100 & 0 & 0 \\ 0 & 0 & 0 & 25 \end{bmatrix}, \quad D_m = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and to the step response shown in Fig. 13. To justify our choice of a pseudo optimal solution instead of a optimal one, it is interesting to check whether relation (15) is satisfied in the case under examination. Refer to Fig. 5 and define system $\overline{\Sigma} \equiv (\bar{A}, \bar{B}, \bar{C})$. The check is done through the following Matlab lines.

```
>> nA=length(A); nB=size(B,2); mC=size(C,1);
>> nAm=length(Am); nBm=size(Bm,2);
>> Abar=[A zeros(nA,nAm); zeros(nAm,nA) Am];
>> Bbar=[B; zeros(nAm,nB)];
>> Cbar=[C, Cm];
>> Dbar=zeros(mC,nB);
>> Hbar=[zeros(nA,nBm); Bm];
>> V1bar=vstargh2(Abar,Bbar,Cbar,Dbar);
>> VB=sums(V1bar,Bbar);
>> VBH=sums(VB,Hbar);
```

The routine *vstargh2* provides matrix V_1 in (13), that is a basis matrix of $\mathcal{V}_1$. Matrix VB, whose image is $\mathcal{V}_1 + \mathcal{B}$, has less columns than VBH, whose image is $\mathcal{V}_1 + \mathcal{B} + \mathcal{H}$, hence (15) is not satisfied and exact H_2-optimal model following is not feasible in this case.

Let compute the matrix C_1 with the procedure outlined in the proof of Theorem 4 in Sect. A.

```
>> V=vstar(A,B,C,D);
>> S=sstar(A,B,C,D);
>> V1=vstargh2(A,B,C,D);
>> Vr=sums(V1,ints(S,ker(C)));
>> C1=ortco(Vr)';
>> G=-C*inv(A)*B; G1=-C1*inv(A)*B;
>> C1=G*inv(G1)*C1;
>> z1=gazero(A,B,C1,D);
```

We have $z_1^T = \begin{bmatrix} -14.7907 & -9.7907 \end{bmatrix}$, hence the new system $\Sigma_1 \equiv (A, B, C_1, D)$ is minimum-phase. We can easily check that it is left and right invertible and that its relative degree is 2.

We go ahead with the synthesis of a partial-order observer, connected to the system as shown in Fig. 2. It provides an estimate of the state, that is used to obtain the new output y_1. To design the observer, the following command is used.

```
>> [Ao,Bo1,Bo2,Co,Do1,Do2]=redobs(A,B,C,D);
```

The routine *redobs* implements a standard reduced-order state observer. In the output list the matrices B_{01} and D_{01} refer to input u, while B_{02} and D_{02} refer to input y. While running, the poles to be assigned for the dynamics of the observer are asked for. In our design we choose the values (−10, −12, −15). The following command generates the overall system $\Sigma_n \equiv (A_n, B_n, C_n, D_n)$ that provides an estimate of the state as output.

```
>> nA=length(A); nAo=length(Ao);
>> An=[A zeros(nA,nAo); Bo2*C Ao];
>> Bn=[B; Bo1+Bo2*D];
>> Cn=[Do2*C Co];
>> Dn=Do1+Do2*D;
```

If we want to change the poles of the minimal form of Σ_n (that are those of Σ by now), we have to add the commands

```
>> F=-place(A,B,p);
>> An=An+Bn*F*Cn;
```

(where p is the vector of the poles to be assigned) that correspond to the feedback connection shown in Fig. 2. The overall new system for the design of the compensator Σ_c is $\Sigma_t \equiv (A_t, B_t, C_t)$, with

```
>> At=An; Bt=Bn; Ct=C1*Cn;
```

To simplify computations, we can also directly use the minimal realization of Σ_t and define

```
>> At=A-B*F; Bt=B; Ct=C1;
```

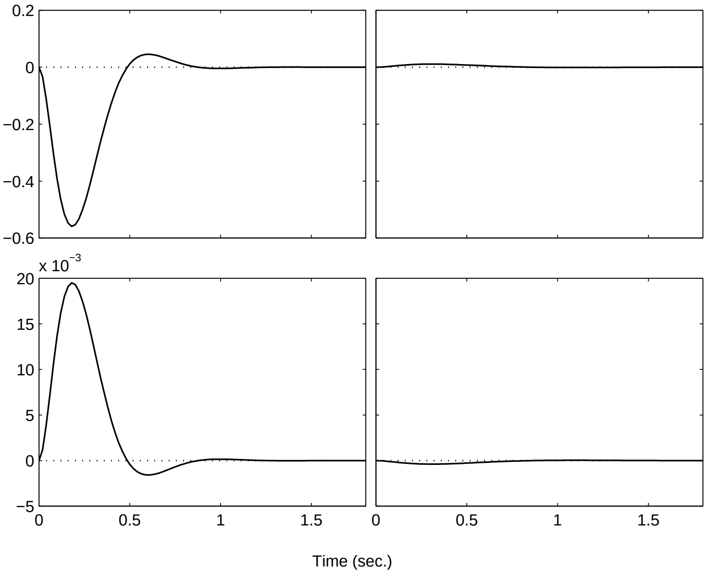


Fig. 14. The actual step response tracking error.

Now we perform on the model the feedback connection shown in Fig. 9 through

```
>> Am=Am+Bm*Cm;
```

and we are ready to obtain the regulator Σ_c with the final list of commands

```
>> nAt=length(At); nBt=size(Bt,2);
>> nAm=length(Am); nBm=size(Bm,2);
>> Ahat=[At zeros(nAt,nAm); zeros(nAm,nAt) Am];
>> Bhat=[Bt;zeros(nAm,nBt)];
>> Hhat=[zeros(nAt,nBm);Bm];
>> Chat=[Ct -Cm];
>> [Ac,Bc,Cc,Dc]=hud(Ahat,Bhat,Chat,Hhat);
```

The routine *hud* solves Problem 2, i.e., it checks if the conditions (7)-(9) are satisfied, and, if so, it computes the measurable signal decoupling compensator by assuming $\mathcal{V}_m$ as the resolving controlled invariant. Thus the final result $\Sigma_c \equiv (A_c, B_c, C_c, D_c)$ is obtained. The eigenvalues of A_c are $[-14.7907$, -9.7907, -12, -6, 0, $0]$, hence an eigenvalue at the origin (the internal model) has been obtained for every feedback loop, while the invariant zeros of Σ_c are equal to the poles of Σ, thus reproducing the classical pole-zero cancellation layout of the SISO case.

Fig. 14 shows the overall tracking error $y - y_m$ for the complete system-model layout (see Fig. 12). From an inspection of the four plots, it is evident that the proposed H_2-pseudo optimal solution guarantees an excellent transient response with almost perfect row-by-row decoupling.

7 The Matlab toolbox "GA"

The first edition of our book on controlled and conditioned invariants [5] enclosed a diskette with some Matlab routines for analysis and synthesis of LTI systems. The algorithms that were used at that time have been improved in the meanwhile, so a quick review of the corresponding computational processes seems to be in order.

A The basic operations on subspaces

The subspaces are numerically expressed through orthonomal basis matrices. Operations on subspaces are performed by using standard routines whose workings are briefly recalled in this section. The overall numerical robustness is held up by the first routine, *ima*, where the crucial decision on linear independence of vectors is taken.

- `Q=ima(A,[fl])` performs the orthonormalization of a set of vectors given as colums of the matrix A and returns them as the columns of matrix Q. The flag *fl* refers to the possible re-ordering of vectors during the orthonormalization process: if it is absent or $fl = 1$ re-ordering is allowed, if $fl = 0$ it is not.
- `Q=ortco(A)` computes the orthogonal complement of im$A$ as

  ```
  X=ima([A,eye(na)],0); Q=X(:,nA+1:mA);
  ```

- `Q=sums(A,B)` computes the sum of im$A$ and im$B$ as

  ```
  Q=ima([A,B]);
  ```

- `Q=ints(A,B)` computes the intersection of im$A$ and im$B$ as

  ```
  Q=ortco(sums(ortco(A),ortco(B)));
  ```

- `Q=invt(A,X)` computes the subspace $A^{-1}\,\mathcal{X}$, inverse transform of $\mathcal{X} = \mathrm{im}X$ with respect to the linear transformation expressed by matrix A. It uses

```
Q=ortco(A'*ortco(X));
```

- `Q=ker(A)` computes the nullspace of matrix A through

```
Q = ortco(A');
```

B The specific tools of the geometric approach

The following routines *miinco* and *mainco* implement exactly the algorithms that were proposed by Basile and Marro in their first paper on the geometric approach [3]. These algorithms have been proved adequate to build up a complete toolbox. An alternative computational option is described in [28], but apparently had no sequel.

- `Q=miinco(A,C,X)` computes a basis matrix Q of the minimum $(A,\mathcal{C})$-conditioned invariant containing $\mathcal{X}$. The subspaces $\mathcal{C}$ and $\mathcal{X}$ are defined as $\mathcal{C}=\text{im}C$ and $\mathcal{X}=\text{im}X$. It operates through the sequence of subspaces

$$\begin{aligned}\mathcal{S}_1 &= \mathcal{X} \\ \mathcal{S}_i &= \mathcal{X} + A\,(\mathcal{S}_{i-1}\cap\mathcal{C}) \quad i=2,3,...\end{aligned} \tag{24}$$

  with stop condition $\mathcal{S}_{i+1} = \mathcal{S}_i$.
- `Q=mainco(A,B,X)` computes a basis matrix Q of the maximum $(A,\mathcal{B})$-controlled invariant contained in $\mathcal{X}$. It utilizes the duality expressed by

```
Q = ortco(miinco(A',ortco(B),ortco(X)));
```

- `[V,F]=vstar(A,B,C[,D])` or `[V,F]=vstar(sys)` provides as V a basis matrix of $\mathcal{V}^\star$, the maximum output nulling subspace of the LTI system `sys=ss(A,B,C,D)` and as F a corresponding *friend*, i.e., a state feedback matrix such that $(A+BF)\,\mathcal{V}^* \subseteq \mathcal{V}^*$.

  - *Computation of* $\mathcal{V}^*$: If D is absent or $D=O$, the routine uses `V=mainco(A,B,ker(C))`, while, if if $D\neq O$ the computation is referred to the extended system

$$A_1 = \begin{bmatrix} A\ O \\ C\ O \end{bmatrix}, \quad B_1 = \begin{bmatrix} B \\ D \end{bmatrix}, \quad C_1 = \begin{bmatrix} O\ I \end{bmatrix} \tag{25}$$

  and matrix V is derived through

```
V1=mainco(A1,B1,ker(C1)); V=V1(1:nA,:);
```

  - *Computation of* F: Recall that for any $(A,\mathcal{B})$-controlled invariant $\mathcal{V}=\text{im}V$ there exist (possibly non-unique) matrices X and U such that[4]

$$A\,V = V\,X + B\,U \tag{26}$$

  Hence X and U can be determines as

[4] See [5], Property 4.1.4.

$$\begin{bmatrix} X \\ U \end{bmatrix} = \begin{bmatrix} V \; B \end{bmatrix}^{\#} A\,V \tag{27}$$

(recall that the symbol $^{\#}$ denotes the pseudoinverse). Then compute

$$F = -U\,V^{\#} \tag{28}$$

In fact, equation (26) can also be written as

$$(A - BUV^{\#})\,V = V\,X \quad \text{or} \quad V^{\#}(A+BF)V = X \tag{29}$$

that proves that $\mathcal{V}$ is an $(A+BF)$-invariant with internal dynamics expressed by matrix X. If the system is not left invertible, i.e., if $\mathcal{V}^* \cap \mathcal{B} \neq \{0\}$, the matrices on the left of (27) and (28) are non-unique, In fact, in this case the internal dynamics of $\mathcal{R}^* = \mathcal{V}^* \cap \mathcal{S}^*$ is completely assignable through a suitable choice of F [5].

- `[S,G]=sstar(A,B,C[,D])` or `[S,G]=sstar(sys)` provides as S a basis matrix of $\mathcal{S}^\star$, the maximum input containing subspace of the LTI system `sys=ss(A,B,C,D)`, and as G an output injection matrix such that $(A+GC)\,\mathcal{S}^* \subseteq \mathcal{S}^*$. By duality, the computation is simply done through the statements

```
[V,F] = vstar(A’,C’,B’,D’)); S=ortco(V); G=F’;
```

- `[z,X]=gazero(A,B,C[,D])` or `[z,X]=gazero(sys)` gives as z the column vector of the invariant zeros and as X the matrix of the invariant zero structure of the LTI system `sys=ss(A,B,C,D)` [6]. Let us recall that the invariant zeros are the unassignable eigenvalues of $\mathcal{V}^*$. If the considered system is not left invertible, i.e., if $\mathcal{R}^* = \mathcal{V}^* \cap \mathcal{B} \neq \{0\}$, there are in $\mathcal{V}^*$ nR assignable eigenvalues and $nV - nR$ unassignable eigenvalues. Consider the change of basis defined by transformation $T = [T_1, T_2, T_3]$ with $\operatorname{im} T_1 = \mathcal{R}^*$, $\operatorname{im}[T_1\,T_2] = \mathcal{V}^*$, then

$$T^{-1}(A+BF)T = \begin{bmatrix} A'_{11} & A'_{12} & A'_{13} \\ 0 & A'_{22} & A'_{23} \\ 0 & 0 & A'_{33} \end{bmatrix} , \quad T^{-1}B = \begin{bmatrix} B'_1 \\ O \\ B'_3 \end{bmatrix}$$

where F denotes any friend of $\mathcal{V}^*$. The invariant zeros are the eigenvalues of matrix A'_{22} and the invariant zero structure is represented by A'_{22} itself. It can be proven that the pair (A'_{11}, B'_1) is always reachable and (A'_{33}, B'_3) is reachable or stabilizable if the considered system is so. Another procedure, more strictly related to the above computational hints, is the following:

a) determine $\mathcal{R}^* = \mathcal{V}^* \cap \mathcal{S}^*$ and denote by R and V the basis matrices obtained for $\mathcal{R}^*$ and $\mathcal{V}^*$, and by nR, nV their numbers of columns;

[5] This can be done by using the routine *effesta*.

[6] To check whether a given controlled invariant $\mathcal{V} = \operatorname{im} V$ is internally stabilizable, the command `z=gazero(A,B,ortco(V)')` can be used.

b) use `V=ima([R,V],0])` to obtain a suitably ordered basis of $\mathcal{V}^*$;

c) use (27) and denote by X_{22} the $(nV - nR) \times (nV - nR)$ submatrix extracted from the bottom/right corner of X. The invariant zeros are the eigenvalues of X_{22} and the invariant zero structure is represented by X_{22} itself.

- `F=effesta(A,B,V)` for any $(A,\mathcal{B})$-controlled invariant $\mathcal{V} = \mathrm{im}V$ a state feedback matrix F is given such that $(A+BF)\,\mathcal{V} \subseteq \mathcal{V}$ and the eigenvalues of $(A+BF)|_{\mathcal{R}^*}$ are assigned in interactive mode. The computational procedure, similar to that of *gazero*, is

  a) determine $\mathcal{R}^* = \mathcal{V}^* \cap \mathcal{S}^*$ and denote by R and V the basis matrices obtained for $\mathcal{R}^*$ and $\mathcal{V}^*$, and by nR the number of columns of R;

  b) use `V=ima([R,V],0])` to obtain a suitably ordered basis of $\mathcal{V}^*$;

  c) use (27) and
  $$K = \ker \begin{bmatrix} X \\ U \end{bmatrix} \tag{30}$$
  and denote by A_1 the $nR \times nR$ submatrix extracted from the top/left corner of the matrix on the left of (27) and by B_1 the matrix consisting of the first nR rows of K.
  d) since (A_1, B_1) is a controllable pair, it is possible to define a matrix F such that $M := A_1 + BF$ has arbitrary eigenvalues, defined in interactive mode;
  e) replace the $nR \times nR$ submatrix at the top/left corner of the matrix on the left of (27) with M and use (28).

- `[V,F]=vstarg(A,B,C,D[,-1)` or `[V,F]=vstarg(sys)` computes a basis matrix V of the maximum output nulling controlled invariant subspace of the LTI system `sys=ss(A,B,C,D[,-1])`. The optional fifth argument "`-1`" is to refer to a discrete-time system. The computational procedure is strongly related to that of *gazero*: When matrix X_{22} has been determined at step d of the algorithm, use the utility *subsplit*[7] to derive the subspace W_s of the stable modes of X_{22}. Through a suitable change of basis relating to V (the basis matrix of $\mathcal{V}^*$), make W_s to be the image of the first nW_s columns of X_{22} and assume as the first output matrix the first $nR + nW_s$ columns of the new V. The friend F is computed like in *vstar*.
- `[V,F,X]=vstargh2(A,B,C,D])` or `[V,F,X]=vstargh2(sys)` computes the maximum internally stabilizable controlled invariant minimizing the output ℓ_2 norm of continuous-time LTI system `sys=ss(A,B,C,D)`. Matrix F is a friend of V and matrix X is used to compute the cost as shown below. The algorithm is based on a generalization of the geometric approach

[7] The routine `[As,Au]=subsplit(A[,-1])` computes basis matrices A_s, A_u of the subspaces of stable and unstable modes of a real square matrix A by using the Schur form.

applied to the Hamiltonian system described in Sect. 4. The quadruple (A,B,C,D) is assumed to be left invertible and stabilizable. Let us refer to the the Hamiltonian system

$$\begin{aligned} \dot{\hat{x}}(t) &= \widehat{A}\,\hat{x}(t) + \widehat{B}\,u(t) \\ 0 &= \widehat{C}\,\hat{x}(t) + \widehat{D}\,u(t) \end{aligned} \qquad \hat{x} = \begin{bmatrix} x \\ p \end{bmatrix} \tag{31}$$

with

$$\begin{aligned} \widehat{A} &= \begin{bmatrix} A & 0 \\ -C^T C & -A^T \end{bmatrix}, \quad \widehat{B} = \begin{bmatrix} B \\ -C^T D \end{bmatrix} \\ \widehat{C} &= \begin{bmatrix} D^T C & B^T \end{bmatrix}, \quad \widehat{D} = D^T D. \end{aligned} \tag{32}$$

that extends (11,12) for quadruples. By using *vstarg*, compute matrices $\widehat{V}$, $\widehat{F}$ and partition them as in (13), i.e., $\widehat{V}^T = \begin{bmatrix} V_1 \; P_1 \end{bmatrix}$ and $\widehat{F} = \begin{bmatrix} F_1 \; F_2 \end{bmatrix}$. Let us assume $V = V_1$. It is a basis matrix of the subspace of the admissible trajectories of system (31) relative to x, that are expressible as $x(t) = V\,\alpha(t)$, with $\alpha(t)$ satisfying the differential equation

$$\dot{\alpha}(t) = V^{\#}\,(A + BF)\,V\,\alpha(t)\,, \quad \alpha(0) = \alpha_0 \tag{33}$$

where F is defined as $F = (F_1 + F_2\,P_1\,V_1^{\#})$, while $X = V_1^T\,P_1$ is the matrix of the cost, that is computable as $c_{(0,\infty)} = \alpha_0^T\,X\,\alpha_0$.

- `r=reldeg(A,B,C[,D])` or `r=reldeg(sys)` provides the relative degree of the LTI system `sys=ss(A,B,C,D)`. Let us momentarily assume that $D = O$ and that the system is right-invertible, i.e., $\mathcal{V}^* + \mathcal{S}^* = \mathcal{X}$, where $\mathcal{X}$ denotes the whole space. In this case the relative degree is the minimum value of i in the sequence (24) such that $\mathcal{V}^* + \mathcal{S}_i = \mathcal{X}$. If the system is not right-invertible but left invertible, compute the relative degree referring to the dual system (A^T, C^T, B^T, D^T). If the system is neither right nor left invertible, use squaring down provided by the routine *extendf*. If $D \neq O$, the relative degree is computed for the auxiliary system (25) and lowered by one.
- `r=rhomin(A,B,C[,D])` or `r=rhomin(sys)` provides the minimum delay of the LTI system `sys=ss(A,B,C,D)`. When $D = O$ the minimum delay is computed as the minimum values of i such that $C\,A^i\,B \neq O$. When $D \neq O$ the minimum delay is computed referring to the auxiliary system (25) and lowered by one.
- `[Am,Bm,Cm,Dm,Fs,Us]=extendf(A,B,C,D)` given an LTI system described by the quadruple (A,B,C,D) that is non-left invertible (i.e., typically with more inputs than outputs) computes F_s and U_s such that the new system `sysm=ss(Am,Bm,Cm,Dm)` is left invertible and, besides the zeros of (A,B,C,D) has a number of new zeros equal to $\dim \mathcal{R}^*$, defined in interactive mode.

The new system is defined by $A_m = A + B\,F_s$; $B_m = B\,U_s$; $C_m = C + D\,F_s$; $D_m = D\,U_s$. When a state feedback matrix F_m referred to (A_m, B_m, C_m, D_m)

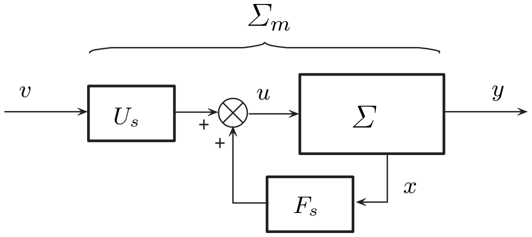


Fig. 15. Squaring down of a non-left invertible system.

has been derived, the solution referred to (A,B,C,D) is recovered by using `F=Us*Fm+Fs`. This procedure is called *squaring down* and is used to deal with non-left invertible systems when sythesizing controllers with computational processes that require left invertibility[8]. If $D=O$, F_s is computed as `Fs=effesta(A,B,Rv)`, where R_v denotes a basis matrix of $\mathcal{R}^*=\mathcal{V}^*\cap\mathcal{S}^*$ [9], while B_m is defined as $B_m=B\,U$, where U is a basis matrix of $(B^{-1}\,\mathcal{V}^*)^\perp$, computable with the Matlab command `Bm=B*ortco(invt(B,vstar(A,B,C)))`. If $D\neq O$, the above procedure is still valid if used for the triple (A_1,B_1,C_1) defined in (25).

- `[Ac,Bc,Cc,Dc]=hud(A,B,C,H[,D,G])` computes a feedforward decoupling compensator solving Problem 2. The first four matrices in the call list are those appearing in equations (5), while D and G refer to possible feedthroughs from u and h to y. First, assume that these feedthrough terms are absent. Assume for the system is left invertible, since, if not, a prelimunary squaring down can be used. First, condition (7) in Corollary (2) is checked and, if it is satisfied, a basis matrix V of the $(A,\mathcal{B})$-controlled invariant $\mathcal{V}_m$ defined in (9) is computed. Then use (27) and set $A_c=X$, $C_c=-U$. Owing to the left-invertibility assumption, the projections of H on $\mathcal{V}_m$ and $\operatorname{im}B$ are unique, so that matrices B_c and D_c can be derived as

$$\begin{bmatrix} B_c \\ -D_c \end{bmatrix} = \begin{bmatrix} V & B \end{bmatrix}^{\#} H$$

If the matrices D and G are present and nonzero, substitute the quadruple (A,B,C,H) with (A_1,B_1,C_1,H_1) defined by

$$A_1=\begin{bmatrix} A & O \\ C & O \end{bmatrix} \qquad B_1=\begin{bmatrix} B \\ D \end{bmatrix} \qquad C_1=\begin{bmatrix} O & I \end{bmatrix} \qquad H_1=\begin{bmatrix} H \\ G \end{bmatrix} \tag{34}$$

[8] Typical examples are the Matlab routines *care* and *dare* to solve the infinite-horizon LQR problem for continuous and discrete-time systems, respectively.

[9] Recall that $\mathcal{R}^*\neq\{0\}$ if and only if $\mathcal{V}^*\cap\operatorname{im}B\neq\{0\}$ (see [5], Sect. 4.1.2).

8 Conclusions

This short monograph proposed a new geometric solution to the feedforward and feedback model following problems. The design strategy is based on the replacement of the output matrix of the controlled system with an equally dimensioned matrix ensuring H_2-optimality in the standard disturbance decoupling problem, while maintaining the same global relative degree, tha same steady-state gain and the same H_2 norm as the original system. The new matrix is derived by applying the standard geometric approach tools to the Hamiltonian system instead of the original plant and using some refinements to adapt the solution of the H_2-optimal decoupling problem to the requirements of the model following.

Aknowledgment

My warmest thanks to Fabio Morbidi and Domenico Prattichizzo of the University of Siena for help and suggestions, Augusto Ferrante of the University of Padua for the enlightening discussions on some topics considered herein.

References

1. R. Calimani and A. Lepschy. (1990). *Feedback*. Garzanti, Milano, Italy.
2. G. Basile, R. Laschi, and G. Marro. (1969). *L'Elettrotecnica*, 51(1):1–5.
3. G. Basile and G. Marro. (1969). *Journal of Optimization Theory and Applications*, 3(5):306–315.
4. W.M. Wonham. (1985). *Linear Multivariable Control: A Geometric Approach*. Springer Verlag, New York, 3rd edition.
5. G. Basile and G. Marro. (1992). *Controlled and Conditioned Invariants in Linear System Theory*. Prentice Hall, Englewood Cliffs, NJ.
6. H.L. Trentelman, A.A. Stoorvogel, and M. Hautus. (2001). *Control theory for linear systems*. Communications and Control Engineering. Springer, Great Britain.
7. G. Marro. (2007) The geometric approach to control. In S. Bittanti, editor, *Control Science Evolution*. CNR Publications, Rome.
8. W.A. Wolovich. (1972) *SIAM J. Contr.*, 10(3):512–523.
9. S.H. Wang and C.A. Desoer. (1972). *IEEE Trans. Automat. Contr.*, 17(3):347–349.
10. B.C. Moore and L.M. Silverman. (1972). *IEEE Trans. Automat. Contr.*, 17(4):491–497.
11. S.H. Wang and E.J. Davison. (1972). *IEEE Trans. Automat. Contr.*, 17(4):574.
12. A.S. Morse. (1973). *IEEE Trans. Automat. Contr.*, 18(4):346–354.
13. B.D Anderson and R.W. Scott. (1977). *IEEE Trans. Automat. Contr.*, 22(1):137–138.
14. M. Malabre. (1982). *IEEE Trans. Automat. Contr.*, 27(2):458–461.
15. M. Malabre and V. Kučera. (1984). *IEEE Trans. Automat. Contr.*, 29(3):266–268.
16. J. Descusse and M. Malabre. (1981). *IEEE Trans. Automat. Contr.*, 26(3):791–795.

17. J.C. Martínez García, M. Malabre, and V. Kučera. (1995). *Syst. & Contr. Lett.*, 24(1):61–74.
18. G. Marro and E. Zattoni. (2005). *J. Optim. Theory Appl.*, 125(2):409—-429.
19. A. Saberi, Z. Lin, and A. Stoorvogel. (1995). In *Proc. American Control Conf*, volume 5, pages 3414–3418.
20. G. Marro, D. Prattichizzo, and E. Zattoni. (2002). *Kybernetika*, 4(38):479–492.
21. A. Stoorvogel. (1992). *Automatica*, 28(3):627–631.
22. A. Stoorvogel, A. Saberi, and B.M. Chen. (1993). *Int. J. Contr.*, 58(4):803–834.
23. S.P. Bhattacharyya. (1974). *Int. J. Systems Science*, 5(7):931–943.
24. G. Basile, G. Marro, and A. Piazzi. (1984). In *Proceedings of the '84 International AMSE Conference on Modelling and Simulation*, volume 1.2, pages 19–27, Athens. GAN.
25. B.S. Morgan. (1964). *IEEE Trans. Autom. Contr.*, pages 405–411.
26. J. Descusse, J.F. Lafay, and M. Malabre. (1988). *IEEE Trans. Autom. Contr.*, pages 732–739.
27. G. Marro, D. Prattichizzo, and E. Zattoni. (2002). *IEEE Trans. Automat. Contr.*, 47(1):102–107.
28. B.C. Moore and A.J. Laub. (1978). *IEEE Trans. Autom. Contr.*, pages 783–792.

On the optimality of the Karhunen-Loève approximation

Giorgio Picci

Dipartimento di Ingegneria dell'Informazione, Università di Padova, via Gradenigo 6/B, 35131 Padova, Italy
picci@dei.unipd.it

1 Introduction

This paper is dedicated to the memory of Toni Lepschy, man of many interests, founder of the control group in Padova and above all, an exquisite gentleman. One of his recurrent research interests during his academic life has been model reduction. From the very early papers on Padè approximation to the more recent work, say on L^2 approximation by rational functions, one can see a long thread of contributions to this topic which spans his whole academic career.

In an attempt to adhere to the spirit of this work, this article will also be about a class of model reduction problems. It will actually survey a basic and well-known technique of signal approximation which goes under the names of *Principal Component Analysis (PCA)* (discrete time) or *Karhunen-Loève expansion* (continuous time). These techniques are important and widely used in many applications ranging from data compression, source coding, filtering, data classification and pattern recognition.

Specifically, the problem addressed in this paper is the approximation of random signals by linear combinations of certain deterministic functions of time which we call the *modes* of the signal. We shall first discuss an exact representation of an arbitrary random signal in terms deterministic functions of time, which has certain natural properties. The functions of time appearing in this expansion are the modes of the signal. In general an exact representation requires "too many" modes (infinitely many in continuous time) and the key question is the optimal approximation of the signal in terms of a small number of modes.

This is the signal model reduction problem we shall discuss. A key question, which to our knowledge is scarcely addressed in the literature, is to make sense of this approximation problem in a formal way. Which criterion are we seeking to optimize by this truncation or in what sense is this approximation optimal and what are the error bounds. We shall provide a simple and natural answer to these questions.

2 Discrete-time: Principal Component Analysis

We shall initially discuss the discrete time case since it can basically be treated by linear algebra and elementary Hilbert space techniques.

Assume $\{\mathbf{y}(1),\dots,\mathbf{y}(t),\ \dots,\mathbf{y}(T)\}$, is a random signal defined on a finite time interval $[1,\,T]$; in other words a finite sequence of random variables which for convenience we shall write as a T-dimensional (random) column vector denoted $\mathbf{y}$. Since subtracting the means does not change the construction we shall describe below, without loss of generality we shall assume that $\mathbf{y}$ has zero mean. Denote by

$$\Sigma = \mathbb{E}\,\mathbf{y}\mathbf{y}^\top$$

the covariance matrix of $\mathbf{y}$, which will be assumed to be positive definite and given to us. Denote also by $\mathbf{H}(\mathbf{y})$ the (finite dimensional) Hilbert space linearly generated by the scalar components of $\mathbf{y}$

$$\mathbf{H}(\mathbf{y}) = \{\sum_t \alpha_t \mathbf{y}(t)\,;\, \alpha_t \in \mathbb{R},\, t=1,\dots,T\} := \operatorname{Span}\{\mathbf{y}(t)\,;\, t=1,\dots,T\} \quad (1)$$

with inner product

$$\langle \boldsymbol{\xi},\, \boldsymbol{\eta}\rangle := E\boldsymbol{\xi}\boldsymbol{\eta}\,. \quad (2)$$

We shall look for a linear expansion of $\mathbf{y}$ in terms of a family of deterministic time functions (also written as column vectors), $u_k = [u_k(1)\dots u_k(T)]^\top$, $k = 1,2,\dots$. These are the candidate modes of the signal. Naturally the coefficients of this expansion will have to be random.

Let the random variables $\{\mathbf{x}_k; k=1,\dots,T\}$ form an orthonormal basis of $\mathbf{H}(\mathbf{y})$. We shall consider expansions of the form

$$\mathbf{y} = \sum_{k=1}^{T} \alpha_k\, \mathbf{x}_k u_k, \qquad u_k \in \mathbb{R}^T \quad (3)$$

where α_k are real numbers. Evidently the candidate modes $u_k = [u_k(1)\dots u_k(T)]^\top$ must obey the condition

$$\alpha_k\, u_k := E\{\mathbf{y}\mathbf{x}_k\}, \qquad k=1,\dots,T.$$

The expansion (3) is called *biorthogonal* if the $\{\mathbf{x}_k; k=1,\dots,T\}$ form an orthonormal basis of $\mathbf{H}(\mathbf{y})$ and the modes u_k are orthonormal with respect to the Euclidean inner product in $\mathbb{R}^T$; i.e.

$$u_k^\top u_j = \delta_{k,j} \qquad k,j = 1,2,\dots,T$$

$\delta_{k,j}$ being the Kronecker delta.

Proposition 1. *The random vector* $\mathbf{y}$ *admits a biorthogonal expansion of the form* (3) *if and only if the modes* $\{u_k\}$ *form a system of normalized eigenvectors for the covariance matrix,* Σ*, of* $\mathbf{y}$.

In this case, letting $\lambda_k > 0$ denote the eigenvalue of Σ corresponding to the eigenvector u_k, it holds that $\alpha_k = \sqrt{\lambda_k}$ and the random variables $\mathbf{x}_k$ are given by the formula

$$\mathbf{x}_k = \frac{1}{\sqrt{\lambda_k}}\, u_k^\top\, \mathbf{y} = \frac{1}{\sqrt{\lambda_k}} \sum_{t=1}^{T} u_k(t)\, \mathbf{y}(t), \qquad k = 1, \dots, T. \tag{4}$$

Proof. Sufficiency: Assume the u_k's are chosen to be the orthonormal eigenvectors of Σ. From

$$\Sigma u_k = \lambda_k u_k, \qquad k = 1, \dots, T$$

one easily sees that the variables $\mathbf{x}_k$ defined in (4) are orthonormal with respect to the inner product (2). In fact,

$$E\mathbf{x}_k\, \mathbf{x}_j = \frac{1}{\sqrt{\lambda_k \lambda_j}} u_k' \Sigma u_j = \delta_{k,j}$$

and since by construction they belong to $\mathbf{H}(\mathbf{y})$, they must form an orthonormal basis for this space; i.e. $\mathbf{H}(\mathbf{y}) = \mathbf{H}(\mathbf{x})$ where, in vector notation, $\mathbf{x} := [\mathbf{x}_1, \dots, \mathbf{x}_T]'$. Hence $\mathbf{y}$ coincides with its orthogonal projection (wide-sense conditional expectation) onto $\mathbf{H}(\mathbf{x})$,

$$\mathbf{y} = E\,[\mathbf{y} \mid \mathbf{x}] = \sum_{k=1}^{T} E\{\mathbf{y}\mathbf{x}_k\}\, \mathbf{x}_k\, .$$

Since $E\{\mathbf{y}\mathbf{x}_k\} = \frac{1}{\sqrt{\lambda_k}} E\{\mathbf{y}\mathbf{y}'\}\, u_k = \sqrt{\lambda_k}\, u_k$ we find a representation of the form (3) in which $\alpha_k = \sqrt{\lambda_k}$.

Necessity: Assume $\mathbf{y}$ admits a biorthogonal expansion (3). From this the following expression for the covariance matrix is found

$$\begin{aligned} \Sigma &= \textstyle\sum_{k,j=1}^{T} \alpha_k \alpha_j\, E[\mathbf{x}_k \mathbf{x}_j] u_k u_j^\top \\ &= \textstyle\sum_{k=1}^{T} \lambda_k u_k\, E[\mathbf{x}_k]^2 u_k^\top \\ &:= U \text{ diag } \{\lambda_1, \dots, \lambda_T\}\, U' \end{aligned}$$

where $U := [u_1, \dots, u_T]$ is an orthogonal matrix (i.e. a matrix with orthonormal columns). It follows that the columns of U must be the normalized eigenvectors of Σ. □

The vectors u_k are usually called the *principal components* of the signal,although *proper modes* would perhaps be a more descriptive denomination.

The eigenvalues of Σ will be listed in *decreasing order*; i.e.

$$\lambda_1 \geq \ldots \geq \lambda_T > 0$$

accordingly, this ordering is transmitted to the random coefficients $\mathbf{x}_k$ and to the corresponding modes u_k. With this convention, taking $\alpha_k = +\sqrt{\lambda_k}$, the biorthogonal expansion (3) is *unique*. It is commonly called the *Principal Components Analysis (PCA)* of the signal $\mathbf{y}$. It is actually just the discrete-time version of the *Karhunen-Loève* expansion which will be discussed section 2.

Model approximation

It often happens that the "statistical energy" of the signal

$$E\,\|\mathbf{y}\|^2 = \sum_{k=1}^{T} \lambda_k E\{\mathbf{x}_k\}^2 = \sum_{k=1}^{T} \lambda_k$$

is concentrated on a few proper modes. In other words it often happens that the eigenvalues of index larger than some $n < T$ are (relatively) small, for example such that

$$\lambda_1 + \ldots + \lambda_n \gg \lambda_{n+1} + \ldots + \lambda_T$$

and their contribution to the expansion (3) can be therefore be neglected. The resulting approximate expansion,

$$\mathbf{y} \simeq \hat{\mathbf{y}}_n := \sum_{k=1}^{n} \sqrt{\lambda_k}\; \mathbf{x}_k u_k = \sum_{k=1}^{n} (u_k^\top \mathbf{y})\, u_k \tag{5}$$

is universally used in data compression, source coding, data storage and especially in pattern recognition. In practice the covariance matrix Σ is not known but can usually be estimated from experimental data. For example from N independent measurements of the same signal, say $\{y_1, \ldots, y_N\}$, one can form the sample covariance estimate

$$\hat{\Sigma} = \frac{1}{N} \sum_{k=1}^{N} y_k y_k^\top$$

and use this estimate in place of the true Σ.

Note that the approximation error vector $\tilde{\mathbf{y}}_n := \mathbf{y} - \hat{\mathbf{y}}_n$ is orthogonal to $\hat{\mathbf{y}}_n$, so that

$$\Sigma = \mathrm{Var}\ \mathbf{y} = \mathrm{Var}\ \hat{\mathbf{y}}_n + \ \mathrm{Var}\ \tilde{\mathbf{y}}_n \ := \ \hat{\Sigma}_n + \tilde{\Sigma}_n$$

and the variance matrix of the approximant $\hat{\mathbf{y}}_n$ can be expresses as

$$\hat{\Sigma}_n = \sum_{k=1}^{n} \lambda_k\, u_k E\{\mathbf{x}_k^2\}\, u_k' = \sum_{k=1}^{n} \lambda_k\, u_k\, u_k'.$$

A very well-known property of the approximation is recalled in the following proposition.

Proposition 2. *The variance matrix $\hat{\Sigma}_n$, of $\hat{\mathbf{y}}_n$, is the best symmetric positive semidefinite approximant of rank n, (either in the ℓ^2 or Frobenius norm) of the original variance matrix Σ.*

Proof. The statement follows from the well-known optimal approximation property of the truncated Singular Value Decomposition of a matrix [4]. One just needs to apply the result to a square root (say the Cholesky factor) of Σ. □

It is normally claimed and often given for granted in the literature that the approximation procedure described above does provide an *optimal* approximate representation of the signal. However besides the optimality of the covariance approximation described above, to our best knowledge there is no satisfactory discussion of what this optimality should be in terms of *random signal approximation* nor explicit proof that the truncation (5) is actually optimal with respect to a criterion of this kind.

Below we shall try to understand what kind of approximation criterion should be natural and reasonable to use in this context. To begin with, let us observe that the second member of (5) can be seen as a linear transformation acting on the random vector $\mathbf{y}$, represented by a certain deterministic matrix say M, which is symmetric, positive semidefinite and of rank n.

Any M of this kind can be written in factorized form $M = WW^\top$ where W is $T \times n$ and of full column rank. That M has rank n $(\leq T)$, implies that the approximation $\hat{\mathbf{y}} := M\mathbf{y}$, generates an n-dimensional subspace of $\mathbf{H}(\mathbf{y})$. In this sense we can say that (5) provides an approximation $\hat{\mathbf{y}} := M\mathbf{y}$, *of rank* n, of $\mathbf{y}$.

Motivated from the above, let us consider a problem of optimal rank n approximation of the random vector $\mathbf{y}$, having the following natural formulation.

Problem 1. Find a matrix $M \in \mathbb{R}^{T\times T}$ of rank n, solving the following minimum problem

$$\min_{\mathrm{rank}\,(\,M\,)\,=\,n} E\{\|\mathbf{y} - M\,\mathbf{y}\|^2\} \tag{6}$$

Note that an equivalent geometric formulation is to look for an optimal n-dimensional subspace of $\mathbf{H}(\mathbf{y})$ onto which $\mathbf{y}$ should be projected in order to minimize the approximation error variance. Let us stress that this is quite different from the usual least squares approximation problem which amounts to projecting onto a *given subspace*.

As for (5), minimizing the square distance in (6) requires that the approximation $M\mathbf{y}$ should be uncorrelated with the approximation error; namely

$$\mathbf{y} - M\mathbf{y} \;\perp\; M\mathbf{y} \tag{7}$$

which is equivalent to

$$M\Sigma - M\Sigma M^\top = 0\,.$$

Introducing a square root $\Sigma^{1/2}$ of Σ and defining $\hat{M} := \Sigma^{-1/2} M \Sigma^{1/2}$, this condition is seen to be equivalent to

$$\hat{M} = \hat{M}\,\hat{M}^{\top}$$

which just says that $\hat{M}$ must be symmetric and *idempotent* (i.e. $\hat{M} = \hat{M}^2$), in other words an *orthogonal projection* from $\mathbb{R}^T$ onto some n-dimensional subspace. Hence M must have the following structure

$$M \;=\; \Sigma^{1/2}\,\Pi\,\Sigma^{-1/2}, \qquad \Pi \;=\; \Pi^2 \qquad \Pi \;=\; \Pi^{\top} \tag{8}$$

where $\Sigma^{1/2}$ is any square root of Σ and Π is an orthogonal projection matrix of rank n.

Theorem 1. *The solutions of the signal approximation problem* (6) *are of the form*

$$M = W\,W^{\top}, \qquad W = U_n Q_n$$

where U_n is a $T \times n$ matrix whose columns are the first n normalized eigenvectors of Σ, ordered according to the descending magnitude ordering of the corresponding eigenvalues and Q_n is an arbitrary $n \times n$ orthogonal matrix.

Proof. Let $\Lambda := \operatorname{diag}\{\lambda_1, \ldots, \lambda_T\}$ and $\Sigma \;=\; U \Lambda U^{\top}$ the spectral decomposition of Σ in which U is an orthogonal matrix of eigenvectors. We can pick as a square rot of Σ the matrix $\Sigma^{1/2} \;:=\; U\Lambda^{1/2}$.

Now, no matter how $\Sigma^{1/2}$ is chosen, the random vector $\mathbf{e} := \Sigma^{-1/2}\,\mathbf{y}$ has orthonormal components. Hence using (8) the cost function of our minimum problem can be rewritten as

$$\begin{aligned} E\{\|\mathbf{y} - M\,\mathbf{y}\|^2\} \;&=\; E\{\|\Sigma^{1/2}\mathbf{e} - \Sigma^{1/2}\,\Pi\,\Sigma^{-1/2}\mathbf{y}\|^2\} \\ &=\; E\{\|\Sigma^{1/2}(\mathbf{e} - \Pi\,\mathbf{e}\,)\|^2\} \;=\; E\{\|\Lambda^{1/2}(\mathbf{e} - \Pi\,\mathbf{e}\,)\|^2\} \\ &=\; E\,(\mathbf{e} - \Pi\,\mathbf{e}\,)^{\top}\Lambda\,(\mathbf{e} - \Pi\,\mathbf{e}\,) \\ &=\; \operatorname{Tr}\,\left[\Lambda\,E(\mathbf{e} - \Pi\,\mathbf{e}\,)(\mathbf{e} - \Pi\,\mathbf{e}\,)^{\top}\right] \end{aligned}$$

where $\operatorname{Tr} A := \sum a_{kk}$ is the trace of A. Our minimum problem can therefore be rewritten as

$$\min_{\operatorname{rank}\,(\,\Pi\,) \,=\, n} \operatorname{Tr}\,\{\Lambda\Pi^{\perp}\}$$

where $\Pi^{\perp} := I - \Pi$ is the orthogonal projection onto the orthogonal complement of the subspace $\operatorname{Im}\,\Pi$.

Since the eigenvalues are ordered in decreasing order; i.e. $\{\lambda_1 \geq \ldots \geq \lambda_T\}$, one sees that the minimum of this function of Π is reached when Π projects onto the subspace spanned by the first n coordinate axes. In other words, $\Pi_{optimal} = \operatorname{diag}\{I_n,\, 0\}$ the minimum being $\lambda_{n+1} + \ldots + \lambda_T$. It is then evident that

$$M \,=\, U\Lambda^{1/2}\,\Pi_{optimal}\Lambda^{-1/2}U^{\top} = U_n U_n^{\top}.$$

Naturally, multiplying U_n by any orthogonal $n \times n$ matrix does not change the result. □

The theorem may possibly be a novel contribution of this paper; it confirms in particular that the truncated expansion (5) is optimal in the sense that it provides the best M and the best approximation subspace for the criterion (6). This characterization can be exploited when dealing with subspace approximation problems; see e.g. [5].

We should finally remark that, although the paper [3] comes very close, in mathematical terms, to our problem formulation above, its proof of optimality rests on rather deep results of linear algebra which do not seem to generalize so easily to the continuous time case. Our approach is instead completely elementary and the generalization is straightforward.

3 Continuous time: the Karhunen-Loève expansion

The Karuhnen-Loève expansion is the analog of PCA for continuous time signals. Although from a conceptual point of view there are no novelties, more sophisticated mathematics is needed; in particular the spectral decomposition of the covariance matrix $\varSigma$ must be replaced by an eigenfunction expansion of a certain integral operator.

Let us consider a continuous time random signal $\mathbf{y} := \{\mathbf{y}(t);\ t \in T\}$, the variabile t now ranging on some interval of the real line which we denote T. As before we assume zero mean (w.l.o.g.). The covariance function

$$R(t,s) := E\{\mathbf{y}(t)\mathbf{y}(s)\} \tag{9}$$

is assumed to be continuous in both arguments. This condition is equivalent to mean square continuity of the process. An essential technical assumption is that

$$\int_T \int_T R(t,s)^2 \, dtds \ < \ \infty \tag{10}$$

Clearly, when T is a finite interval this condition is automatically satisfied. Let us now consider the inner product space space $C^2[T]$ of continuous (deterministic) signals endowed with the inner product

$$\langle f,\, g\rangle \ := \int_T f(t)\, g(t)\, dt$$

This space is not complete (i.e. Hilbert) in general. It is immediate to check that, in force of condition (10), the linear operator $\varSigma_R$ defined by

$$[\varSigma_R\, f](t) \ := \int_T R(t,s)\, f(s)\, ds \tag{11}$$

maps $C^2[T]$ into itself (inn fact $\varSigma_R$ is a *compact operator* in force of condition (10) . The eigenvalues and the corresponding eigenfunctions of an integral operator of this kind are pairs λ, φ with $0 < \|\varphi\|_{L^2[T]} < \infty$, which satisfy

$$[\Sigma_R\,\varphi](t) \;=\; \int_T R(t,s)\,\varphi(s)\,ds \;=\; \lambda\varphi(t) \qquad t\in T \tag{12}$$

Although eigenvalues may in general not exist at all, it is well-known that under the compactness condition (10), the operator Σ_R does admit eigenvalues. In fact it behaves virtually like a finite dimensional operator described by a symmetric positive definite matrix. The following is the central result of the theory; see e.g. [1, 2] for the complete story.

Theorem 2 (Mercèr). *Under the stated assumptions, the following holds:*

1. *The eigenvalue problem* (12) *admits solutions and all eigenvalues are real.*
2. *There is a maximal eigenvalue* λ_0 *given by the formula*

$$\lambda_0 = \max_{\|\varphi\|_{L^2[T]}=1} \langle \Sigma_R\varphi,\,\varphi\rangle \;=\; \max_{\|\varphi\|_{L^2[T]}=1} \int_T\int_T R(t,s)\,\varphi(t)\varphi(s)\,dtds \tag{13}$$

   The corresponding eigenfunction, $\varphi_0(t)$*, is a continuous function and belongs to* $C^2[T]$.
3. *The function* $R_1(t,s) := R(t,s) - \lambda_0\varphi_0(t)\varphi_0(s)$ *is still a covariance function (of positive type) satisfying the compactness condition* (10)*. Hence the eigenvalue problem*

$$[\Sigma_{R_1}\,\varphi](t) \;:=\; \int_T R_1(t,s)\,\varphi(s)\,ds \;=\; \lambda\,\varphi(t) \tag{14}$$

   still has a maximal eigenvalue, λ_1*, given by*

$$\lambda_1 = \max_{\|\varphi\|_{L^2[T]}=1} \langle \Sigma_{R_1}\varphi,\,\varphi\rangle \;=\; \max_{\|\varphi\|_{L^2[T]}=1} \int_T\int_T R_1(t,s)\,\varphi(t)\varphi(s)\,dtds \tag{15}$$

   and $\lambda_1 \le \lambda_0$.
4. *The procedure can be iterated. The eigenvalues of the problem* (12) *form a monotone nonincreasing sequence of positive numbers (not necessarily distinct), whose only accumulation point can be* 0*. The corresponding eigenfunctions are all continuous an belong to* $C^2[T]$*; they can be made orthonormal so that*

$$\int_T \varphi_k(t)\varphi_j(t)\,dt \;=\; \delta_{k,j}\,.$$

5. *The covariance function* (9) *admits the following expansion*

$$R(t,s) \;=\; \sum_{k=0}^{\infty} \lambda_k\,\varphi_k(t)\varphi_k(s), \qquad t,s\,\in\,T\times T \tag{16}$$

   the series being pointwise uniformly convergent on $T\times T$.

By truncating the expansion (16) to the first $n+1$ terms, one can obtain an approximation of rank $n+1$ of the covariance function $R(t,s)$,

$$R(t,s) \simeq R_n(t,s) := \sum_{k=0}^{n} \lambda_k \, \varphi_k(t)\varphi_k(s), \tag{17}$$

It is possible to show that this approximation is the best possible in a variety of ways. For example, the linear operator Σ_{R_n} defined by

$$[\Sigma_{R_n} f](t) := \int_T R_n(t,s)\, f(s)\, ds$$

is the best approximant of rank $n+1$ of Σ_R, in the sense that it solves the constrained optimum problem

$$\min_{\text{rank}\,(\,\Sigma\,)\,=\,n+1} \|\Sigma_R - \Sigma\| \tag{18}$$

the minimum being exactly λ_{n+1}, the first neglected eigenvalue. Here the norm is the operator norm (defined e.g. in [1])

The following is the continuous time analog of Proposition 1

Proposition 3. *If the covariance function of the random process* $\mathbf{y}$ *is a continuous function saisfying* (10)*, then* $\mathbf{y}$ *admits the biorthogonal expansion*

$$\mathbf{y}(t) = \sum_{k=0}^{+\infty} \sqrt{\lambda_k}\, \varphi_k(t)\, \mathbf{x}_k \tag{19}$$

where $\lambda_k;\ k = 0,1,\ldots$ *are the eigenvalues of the operator* Σ_R*, ordered in decreasing magnitude,* $\varphi_k;\ k = 0,1,\ldots$ *the corresponding normalized eigenfunctions and the random variables* $\mathbf{x}_k;\ k = 0,1,\ldots$ *are defined by*

$$\mathbf{x}_k = \frac{1}{\sqrt{\lambda_k}} \int_T \varphi_k(t)\, \mathbf{y}(t)\, dt \tag{20}$$

These random variables form an orthonormal basis for the Hilbert space $\mathbf{H}(\mathbf{y})$*, generated by the process* $\mathbf{y}$*. The expansion converges in quadratic mean, uniformly in* $t \in T$*.*

The above is the celebrated *Karhunen-Loève expansion* of the process $\mathbf{y}$. In analogy to the discrete time case, the expansion (19) is normally truncated to a finite number of terms, leading to the approximate description

$$\mathbf{y}_n(t) = \sum_{k=0}^{n} \sqrt{\lambda_k}\, \varphi_k(t)\, \mathbf{x}_k \tag{21}$$

in terms of the first $n+1$ modes. The quality of the approximation can be measured in terms of statistical energy content. Since

$$E\int_T \mathbf{y}(t)^2\,dt \;=\; \int_T E\,\mathbf{y}(t)^2\,dt \;=\; \sum_{k=0}^{+\infty}\lambda_k$$

the energy of $\mathbf{y}_n$ is just given by the sum above truncated to the first $n+1$ terms. One sees that the energy of the approximation error $\mathbf{y}-\mathbf{y}_n$ decreases with n at the same rate as the residual sum of eigenvalues of the operator Σ_R. In relative terms the energy of the error can be expresed as the ratio

$$\frac{\sum_{k=n+1}^{+\infty}\lambda_k}{\sum_{k=0}^{+\infty}\lambda_k} \;=\; 1-\frac{\sum_{k=0}^{n}\lambda_k}{\sum_{k=0}^{+\infty}\lambda_k}.$$

In fact, one can show, in perfect analogy to Theorem 1, that the truncated expansion (21) provides the best approximant of rank $n+1$ of $\mathbf{y}$ in the sense that it minimizes the norm

$$\|\mathbf{y}-\hat{\mathbf{y}}\|^2 \;:=\; \int_T E\,|\mathbf{y}(t)-\hat{\mathbf{y}}(t)|^2\,dt$$

where $\hat{\mathbf{y}}$ is a m.s. continuous process with $H(\hat{\mathbf{y}})\subset H(\mathbf{y})$ a subspace of dimension $n+1$.

Note that in general the expansion in sinusoidal modes provided by a pathwise Fourier analysis of the signal has worse approximation properties than the expansion (21).

K-L expansion of stationary processes

When $\mathbf{y}$ is a stationary process one has $R(t,s)=R(t-s)$ and it is easy to check that the condition (10) can hold only when T is a *bounded interval*. If the interval T is unbounded, say $T=[0,+\infty)$, the process has no K-L expansion.

On a finite interval, say $T=[-a,\,a]$, the covariance function can be expanded in Fourier series

$$\Sigma(\tau)=\sum_{k=0}^{+\infty}\sigma_k\,\cos\frac{k\pi\tau}{a}$$

and substituting this expression in the integral equation (12) and taking into account the orthogonality of the cosine functions, one readily sees that the eigenvalues of the operator Σ_R are simply the Fourier coefficients of R; i.e. $\lambda_k=\sigma_k$, while the normalized eigenfunctions are

$$\varphi_k(t) \;=\; \frac{1}{\sqrt{a}}\cos\frac{k\pi\tau}{a}$$

Hence the random variables $\mathbf{x}_k$ are just the (random) Fourier coefficients of the signal $\mathbf{y}$. In this case the K-L expansion coincides with the Fourier representation.

References

1. Akhiezer N.I., Glazman I.M. (1966). *Theory of Linear Operators in Hilbert Spaces*, voll. I e II, Ungar, New, York.
2. Dieudonnè J. A. (1969). *Foundations of modern analysis, Chap XI*, Academic Press.
3. Dür A. (1998). *SIAM J. Control and Optimiz.*, 36, pp. 1937-1939.
4. Golub G.H, Van Loan C. F. (1989). *Matrix Computations*, Johns Hopkins U. P., Baltimore.
5. Yang B. (1995). *IEEE Transactions on Signal Processing*, 43, 95–107.

On the reachability properties of continuous-time positive switched systems

Paolo Santesso and Maria Elena Valcher

Dipartimento di Ingegneria dell'Informazione, Università di Padova, via Gradenigo 6/B, 35131 Padova, Italy
santesso@dei.unipd.it, meme@dei.unipd.it

> *È difficile cercare di ricordare Toni e quello che ha rappresentato per me e per tutti noi, qui al DEI di Padova, senza correre il rischio di essere prolissi, forse un po' retorici e, cosa che più temo, poco efficaci. Vorrei ricordare di Toni una cosa sola, che mi ha fatto osservare Gianni Marchesini quando ormai Toni non c'era più, e che mi stupisco di non aver notato da sola: la porta di Toni era sempre aperta. E non c'è stata una sola volta in cui io mi sia presentata alla sua soglia e l'abbia trovata chiusa o lui mi abbia chiesto di tornare più tardi. Chiunque giungesse alla sua porta, con una richiesta, una storia da raccontare, una curiosità da soddisfare, un consiglio da chiedere, veniva invitato ad entrare, ad accomodarsi, ed aveva la sua attenzione.*
> *Vedere quella porta chiusa per tanto tempo è stato il modo in cui ho preso coscienza, giorno dopo giorno, del gran vuoto che aveva lasciato.*
> *M.E.V.*

1 Introduction

Modeling of physical phenomena typically comes as the result of a pondered balance among different, and often conflicting, needs. The first natural goal one pursues, when describing a physical system, is accuracy, which is generally ensured by resorting to computationally demanding solutions. As a consequence, this requirement is often weakened in order to achieve feasible solutions, which are more suitable to real-time implementation. Under this point of view, the case often occurs that a complex nonlinear model, which provides a good description of the real system dynamics, can be efficiently replaced by a family of simpler and possibly linear models, each of them appropriate for describing the system evolution under specific working conditions.

This simple fact stimulated, in the last ten-fifteen years, a long stream of research concerned with the analysis and design of "switched systems", by this meaning systems whose describing equations change, according to some switching law, within a (possibly infinite) family of mathematical models. In particular, switched linear systems consist of a family of (linear) subsystems and a switching law, specifying when and how the switching among the various subsystems takes place. Research efforts in this area were first oriented to the investigation of stability and stabilizability issues [5, 6, 15], and it was only a few years later that structural properties, like reachability, controllability and observability, were initially addressed [4, 13, 14, 16, 17].

On the other hand, the positivity requirement is often introduced in the system models whenever the physical nature of the describing variables constrains them to take only positive (or at least nonnegative) values. Positive linear systems have received considerable attention, as they naturally arise in various fields such as bioengineering (compartmental models), economic modelling, behavioral science, and stochastic processes (Markov chains), where the state variables represent quantities, like pressures, population levels, concentrations, etc., that have no meaning unless nonnegative [3].

In this perspective, switched positive systems are mathematical models which keep into account two different needs: the need for a system model which is obtained as a family of simple subsystems, each of them accurate enough to capture the system laws under specific operating conditions, and the need to introduce the nonnegativity constraint the physical variables are subject to. This is the case when trying to describe certain physiological and pharmacokinetic processes. For instance, the insulin-sugar metabolism is captured by two different compartmental models: one valid in steady-state and the other (more complex) describing the evolution under perturbed conditions, following an oral assumption or an intravenous injection.

Of course, the need for this class of systems in specific research contexts has stimulated an interest in theoretical issues related to them. Specifically, structural properties of continuous-time positive switched systems have been recently investigated in [8, 9]. While controllability analysis is quite immediate, the study of reachability has required a certain number of preliminary steps. In particular, necessary conditions for reachability have been investigated in [8] (monomial reachability) and in [9] (pattern reachability), while detailed results regarding the dominant modes of the exponential matrix of a Metzler matrix have been presented in [10]. As it will be clear from this paper, they represent a necessary first step toward the complete problem solution.

The aim of this contribution is to define reachability properties for continuous-time positive switched systems, and to provide some characterizations of these properties. Monomial reachability is investigated in section 3, while pattern reachability is addressed in section 4. Some necessary and/or sufficient conditions for reachability are derived in section 5, while in section 6, by making use of the asymptotic exponential cone of a Metzler matrix, further sufficient conditions for reachability, rather easy to check, are provided.

Before proceeding, we introduce some notation. For every $k \in \mathbb{N}$, we set $\langle k \rangle := \{1, 2, \dots, k\}$. In the sequel, the (i, j)th entry of a matrix A is denoted by $[A]_{i,j}$. If A is block partitioned, we denote its (i, j)th block by $\mathrm{block}_{(i,j)}[A]$.

Given a matrix $A \in \mathbb{R}^{q\times r}$, by the *nonzero pattern* of A we mean the set of index pairs corresponding to its nonzero entries, namely $\overline{\mathrm{ZP}}(A) := \{(i,j) : [A]_{i,j} \neq 0\}$. Conversely, its *zero pattern* $(\mathrm{ZP}(A))$ is the set of indices corresponding to its zero entries. The adaptation to the vector case is straightforward.

The symbol $\mathbb{R}_+$ denotes the semiring of nonnegative real numbers. A matrix A_+ with entries in $\mathbb{R}_+$ is a *nonnegative matrix* $(A_+ \geq 0)$; if $A_+ \geq 0$ and at least one entry is positive, A_+ is a *positive matrix* $(A_+ > 0)$, while if all its entries are positive it is a *strictly positive matrix* $(A_+ \gg 0)$. The same notation is adopted for nonnegative, positive and strictly positive vectors. The *spectral radius* $\rho(A_+)$ of a nonnegative matrix A_+ is the modulus of its largest eigenvalue. The Perron-Frobenius Theorem [1, 2, 7] ensures that $\rho(A_+)$ is always an eigenvalue of A_+, corresponding to a positive eigenvector. We let $\mathbf{e}_i$ denote the ith vector of the canonical basis in $\mathbb{R}^n$ (where n is always clear from the context), whose entries are all zero except for the ith one which is unitary. If $\mathcal{S} \subseteq \langle n\rangle$, we let $\mathbf{e}_{\mathcal{S}}$ denote the vector $\sum_{i\in\mathcal{S}} \mathbf{e}_i$.

A *Metzler matrix* is a real square matrix, whose off-diagonal entries are nonnegative. If A is an $n \times n$ Metzler matrix, then there exist a nonnegative matrix $A_+ \in \mathbb{R}_+^{n\times n}$ and a nonnegative number α such that $A = A_+ - \alpha I_n$. As a consequence, the spectrum of A, $\sigma(A)$, is obtained from the spectrum of A_+ by simple translation. This ensures, in particular, that [12]:
1) $\lambda_{\max}(A) = \rho(A_+) - \alpha \in \sigma(A)$ is a real dominant eigenvalue, by this meaning that $\lambda_{\max}(A) > \mathrm{Re}(\lambda), \forall\, \lambda \in \sigma(A), \lambda \neq \lambda_{\max}(A)$;
2) there exists a positive eigenvector $\mathbf{v}_1$ corresponding to $\lambda_{\max}(A)$.

To every $n \times n$ Metzler matrix A we associate [2, 11] a *directed graph* $\mathcal{G}(A)$ of *order* n, with vertices indexed by $1, 2, \ldots, n$. There is an arc (j, i) from j to i if and only if $[A]_{ij} \neq 0$. We say that vertex i is *accessible* from j if there exists a path (i.e., a sequence of adjacent arcs $(j, i_1), (i_1, i_2), \ldots, (i_{k-1}, i)$) in $\mathcal{G}(A)$ from j to i (equivalently, $\exists\, k \in \mathbb{N}$ such that $[A^k]_{ij} \neq 0$). Two distinct vertices i and j are said to *communicate* if each of them is accessible from the other. Each vertex is assumed to communicate with itself. The concept of communicating vertices allows to partition the set of vertices $\langle n\rangle$ into *communicating classes*, say $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_\ell$. The *reduced graph* $\mathcal{R}(A)$ [11] associated with A (with $\mathcal{G}(A)$) is the (acyclic) graph having the classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_\ell$ as vertices. There is an arc (j, i) from $\mathcal{C}_j$ to $\mathcal{C}_i$ if and only if $\mathrm{block}_{(i,j)}[A] \neq 0$. With any class $\mathcal{C}_i$ we associate two index sets:

$$\mathcal{A}(\mathcal{C}_i) := \{j : \text{ the class } \mathcal{C}_j \text{ has access to the class } \mathcal{C}_i\}$$
$$\mathcal{D}(\mathcal{C}_i) := \{j : \text{ the class } \mathcal{C}_j \text{ is accessible from the class } \mathcal{C}_i\}.$$

Each class $\mathcal{C}_i$ is assumed to have access to itself. Any (acyclic) path $(i_1, i_2), (i_2, i_3), \ldots, (i_{k-1}, i_k)$ in $\mathcal{R}(A)$ identifies a *chain of classes* $(\mathcal{C}_{i_1}, \mathcal{C}_{i_2}, \ldots, \mathcal{C}_{i_k})$, having $\mathcal{C}_{i_1}$ as *initial class* and $\mathcal{C}_{i_k}$ as *final class*. An $n \times n$ Metzler matrix A is *reducible* if there exists a permutation matrix P such that

$$P^T A P = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

where A_{11} and A_{22} are square (nonvacuous) matrices, otherwise it is *irreducible*. It follows that 1×1 matrices are always irreducible. In general, given a square Metzler matrix A, a permutation matrix P can be found such that

$$P^T A P = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1\ell} \\ & A_{22} & \dots & A_{2\ell} \\ & & \ddots & \vdots \\ & & & A_{\ell\ell} \end{bmatrix}, \tag{1}$$

where each A_{ii} is irreducible. (1) is usually known as *Frobenius normal form* of A [7]. Clearly, the directed graphs $\mathcal{G}(A)$ and $\mathcal{G}(P^T A P)$ are isomorphic and the irreducible matrices $A_{11}, A_{22}, \dots, A_{\ell\ell}$ correspond to the communicating classes $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_\ell$ of $\mathcal{G}(P^T A P)$ (coinciding with those of $\mathcal{G}(A)$, after a suitable relabelling). When dealing with the graph of a matrix in Frobenius normal form (1), for every $i \in \langle \ell \rangle$, $\mathcal{A}(\mathcal{C}_i) \subseteq \{i, i+1, \dots, \ell\}$, while $\mathcal{D}(\mathcal{C}_i) \subseteq \{1, 2, \dots, i\} = \langle i \rangle$, so that $\mathcal{A}(\mathcal{C}_i) \cap \mathcal{D}(\mathcal{C}_i) = \{i\}$. On the other hand, if $i > j$ then $\mathcal{A}(\mathcal{C}_i) \cap \mathcal{D}(\mathcal{C}_j) = \emptyset$, while if $i < j$ then $\mathcal{A}(\mathcal{C}_i) \cap \mathcal{D}(\mathcal{C}_j) \neq \emptyset \Leftrightarrow i \in \mathcal{D}(\mathcal{C}_j) \Leftrightarrow j \in \mathcal{A}(\mathcal{C}_i)$. If A is irreducible ($\mathcal{G}(A)$ has a single communicating class), then $\lambda_{\max}(A)$ is a simple eigenvalue and the corresponding nonnegative eigenvector $\mathbf{v}_1$ is strictly positive.

Basic definitions and results about cones may be found, for instance, in [1]. We recall here only those facts that will be used within this paper. A set $\mathcal{K} \subset \mathbb{R}^n$ is said to be a *cone* if $\alpha\mathcal{K} \subset \mathcal{K}$ for all $\alpha \geq 0$; a cone is *convex* if it contains, with any two points, the line segment between them. A convex cone $\mathcal{K}$ is *solid* if the interior of $\mathcal{K}$ is nonempty, and it is *pointed* if $\mathcal{K} \cap \{-\mathcal{K}\} = \{0\}$. A closed, pointed, solid convex cone is called a *proper cone.* A cone $\mathcal{K}$ is said to be *polyhedral* if it can be expressed as the set of nonnegative linear combinations of a finite set of *generating vectors.* This amounts to saying that a positive integer k and an $n \times k$ matrix C can be found, such that $\mathcal{K}$ coincides with the set of nonnegative combinations of the columns of C. In this case, we adopt the notation $\mathcal{K} := \mathrm{Cone}(C)$. A proper polyhedral cone $\mathcal{K}$ in $\mathbb{R}^n$ is said to be *simplicial* if it admits n linearly independent generating vectors. In other words, $\mathcal{K} := \mathrm{Cone}(C)$ for some nonsingular square matrix C.

2 Switched positive systems: main definitions

A *continuous-time positive switched system* is described by the equation

$$\dot{x}(t) = A_{\sigma(t)} x(t) + B_{\sigma(t)} u(t), \qquad t \in \mathbb{R}_+, \tag{2}$$

where $x(t)$ and $u(t)$ denote the n-dimensional state variable and the m-dimensional input, respectively, at the time instant t, while σ is a switching sequence, taking values in a finite set $\mathcal{P} = \{1, 2, \dots, p\}$.

We assume that the switching sequence is piece-wise continuous, and hence in every time interval $[0, t]$ there is a finite number of discontinuities, which correspond to a finite number of switching instants $0 = t_0 < t_1 < \dots < t_k < t$.

Also, we assume that, at the switching times t_ℓ, σ is right continuous. For each $i \in \mathcal{P}$, the pair (A_i, B_i) represents a continuous-time positive system, which means that A_i is an $n \times n$ Metzler matrix and B_i an $n \times m$ nonnegative matrix. The initial condition $x(0)$ and the input function $u(\cdot)$ are constrained to take nonnegative values, thus ensuring that $x(t) > 0$ at every time instant $t \in \mathbb{R}_+$.

The definition of reachability for positive switched systems may be given by suitably adjusting the analogous definition given in [4, 13], in order to introduce the nonnegativity constraint on the state and input variables.

Definition 1. *A state $x_f \in \mathbb{R}^n_+$ is said to be* (positively) reachable *if there exist a time instant $t_f > 0$, a switching sequence $\sigma : [0, t_f] \to \mathcal{P}$ and an input $u : [0, t_f] \to \mathbb{R}^m_+$, that lead the state trajectory from $x(0) = 0$ to $x(t_f) = x_f$.*

A positive switched system (2) is said to be (positively) reachable *if every state $x_f \in \mathbb{R}^n_+$ is (positively) reachable.*

In the sequel, the specification "positively" will be omitted. Clearly, among all positive vectors one may want to reach there is the class of monomial vectors. For standard positive systems, "monomial reachability" represents a necessary and sufficient condition for reachability. When dealing with switched positive systems this is not the case, as we will see in the following. For this reason, monomial reachability deserves an independent investigation.

Definition 2. *A positive switched system (2) is said to be* monomially reachable *if every monomial vector $\alpha\, \mathbf{e}_i \in \mathbb{R}^n_+, i = 1, 2, \ldots, n, \alpha \in \mathbb{R}_+$, is reachable.*

In order to investigate monomial reachability and (general) reachability, we preliminary introduce the expression of the forced state evolution of the system at an arbitrary time instant $t > 0$. Given a time interval $[0, t]$ and a switching sequence $\sigma : [0, t] \to \mathcal{P}$, corresponding to a set of switching instants $\{t_0,\, t_1, \ldots,\, t_k\}$ satisfying $0 = t_0 < t_1 < \cdots < t_k < t$, the state at the time instant t, starting from $x(0) = 0$ and under the action of the soliciting input $u(\tau), \tau \in [0, t]$, can be expressed as follows [4, 17]:

$$\begin{aligned} x(t) = {} & e^{A_{i_k}(t-t_k)} \ldots e^{A_{i_1}(t_2-t_1)} \int_{t_0}^{t_1} e^{A_{i_0}(t_1-\tau)} B_{i_0} u(\tau) d\tau + \\ & + e^{A_{i_k}(t-t_k)} \ldots e^{A_{i_2}(t_3-t_2)} \int_{t_1}^{t_2} e^{A_{i_1}(t_2-\tau)} B_{i_1} u(\tau) d\tau + \\ & + \ldots + \int_{t_k}^{t} e^{A_{i_k}(t-\tau)} B_{i_k} u(\tau) d\tau, \text{ where } i_\ell = \sigma(t_\ell), \ell = 0, 1, \ldots, k. \end{aligned}$$

3 Monomial reachability

The first goal we pursued [8] was that of providing a family of equivalent conditions for monomial reachability. They represent the natural starting point for reachability analysis.

Proposition 1. *[8] Given a positive switched system (2) the following conditions are equivalent:*

i) the system is monomially reachable;
ii) $\forall\, i \in \langle n\rangle$, there exist $k \in \mathbb{Z}_+$ and indices $i_1, i_2, \dots, i_{k-1}, i_k, j \in \mathcal{P}$, such that $e^{A_{i_k}} e^{A_{i_{k-1}}} \dots e^{A_{i_1}} e^{A_j} B_j$ has an i-monomial column;
iii) $\forall\, i \in \langle n\rangle$, there exist indices $j \in \mathcal{P}$ and $r \in \langle m\rangle$ such that $A_j \mathbf{e}_i = \alpha_i \mathbf{e}_i$, and $B_j \mathbf{e}_r = \beta_i \mathbf{e}_i$, for some $\alpha_i \geq 0$ and $\beta_i > 0$.

Remark. i) It is worthwhile to remark that condition iii) in the previous proposition necessarily constrains the matrix A_j to be reducible. Consequently, all subsystems (A_i, B_i) with A_i irreducible, play no role in the monomial reachability.
ii) As it clearly follows from the previous Proposition, when dealing with single-input systems, condition iii) must be verified for n distinct indices j in $\mathcal{P}$. As a consequence, a necessary condition for a positive switched system to be monomially reachable (and henceforth reachable) is that $p \geq n$.
iii) Unfortunately, except for the case of 2-dimensional systems, none of the equivalent conditions of Proposition 1 is in general also sufficient for reachability. The interested reader is referred to [8] for the details.

4 Pattern reachability

The concept of monomial reachability naturally extends to the broader concept of “pattern reachability”.

Definition 3. *A positive switched system (2) is said to be* pattern reachable *if for every nonempty set $\mathcal{S} \subseteq \langle n\rangle$ there exists a vector $x_f \in \mathbb{R}^n_+$, with $\overline{\mathrm{ZP}}(x_f) = \mathcal{S}$, which is reachable.*

Even if pattern reachability, like monomial reachability, represents only a necessary condition for reachability, the study of this property enlightens certain features which will be useful for the characterization of reachability.

Lemma 1. *Given a positive switched system (2), a subset $\mathcal{S} \subseteq \langle n\rangle$, an integer $k \in \mathbb{Z}_+$ and indices $i_0, i_1, \dots, i_k \in \mathcal{P}$, there exists a vector $x_f \in \mathbb{R}^n_+$, with $\overline{\mathrm{ZP}}(x_f) = \mathcal{S}$, which can be reached by applying (a suitable nonnegative input and) a switching sequence σ, ordinately taking the values $i_0, i_1, \dots, i_k$, if and only if the cone generated by the columns of the* continuous-time reachability matrix associated with the switching sequence $(i_0, i_1, \dots, i_k)$, *i.e.*

$$\mathcal{R}(i_0, i_1, \dots, i_k) = \left[e^{A_{i_k}} B_{i_k} \ e^{A_{i_k}} e^{A_{i_{k-1}}} B_{i_{k-1}} \ \dots \ e^{A_{i_k}} \dots e^{A_{i_0}} B_{i_0}\right],$$

includes a vector v with $\overline{\mathrm{ZP}}(v) = \mathcal{S}$ (equivalently, there is a selection of the columns in $\mathcal{R}(i_0, i_1, \dots, i_k)$ which sums up to a vector w with $\overline{\mathrm{ZP}}(w) = \mathcal{S}$).

Proof. Assume $\mathcal{S} \subseteq \langle n \rangle$ and let $x_f \in \mathbb{R}^n_+$ be any vector with $\overline{\text{ZP}}(x_f) = \mathcal{S}$. By resorting to equation (3), with $x(t) = x_f$, it is clear that for x_f to be reachable it is necessary and sufficient that there exists a finite number of nonzero matrix products in (3) of the following type

$$e^{A_{i_k}(t-t_k)} \dots e^{A_{i_l}(t_{l+1}-t_l)} \int_{t_{l-1}}^{t_l} e^{A_{i_{l-1}}(t_l-\tau)} B_{i_{l-1}} u(\tau) d\tau \tag{3}$$

which sum up to x_f. We may easily observe that, when our interest is only in nonzero patterns, the role of the nonnegative input $u(t)$ in every time interval $[t_{l-1}, t_l)$ is just that of "selecting" the columns of $B_{i_{l-1}}$. So, if we restrict our attention to input functions $u(t)$ whose entries are either zero or unitary in each time interval $[t_{l-1}, t_l)$ (we denote by u_{l-1} the vector value the function $u(t)$ takes in that interval) the class of all nonzero patterns attainable by means of arbitrary inputs coincides with the class of all nonzero patterns attainable by means of piece-wise constant binary inputs.

Even more, due to the fact that the integral operator does not change the zero pattern properties and that the zero pattern of the exponential matrix is invariant, as t ranges over $\{t \in \mathbb{R}_+, t > 0\}$ (see Lemma A.1, point ii)), it follows that there exists a positive vector x_f with $\overline{\text{ZP}}(x_f) = \mathcal{S}$ which can be reached by means of a switching sequence σ, ordinately taking the values $i_0, i_1, \dots, i_k$, if and only if there is a finite number of matrix products like $e^{A_{i_k}} \dots e^{A_{i_l}} e^{A_{i_{l-1}}} B_{i_{l-1}} u_{l-1}$ which sum up to a vector w with $\overline{\text{ZP}}(w) = \mathcal{S}$.

Proposition 2. *A positive switched system (2) is pattern reachable if and only if for every $\mathcal{S} \subseteq \langle n \rangle$ there exist $k < |\mathcal{S}|$ and indices $i_0, i_1, \dots, i_k \in \mathcal{P}$ such that the cone generated by the columns of the reachability matrix $\mathcal{R}(i_0, i_1, \dots, i_k)$ contains a vector v with $\overline{\text{ZP}}(v) = \mathcal{S}$.*

Proof. By Lemma 1, it is clear that the positive switched system (2) is pattern reachable if and only if for every $\mathcal{S} \subseteq \langle n \rangle$ there exist k and indices $i_0, i_1, \dots, i_k \in \mathcal{P}$ such that the cone generated by the columns of $\mathcal{R}(i_0, i_1, \dots, i_k)$ contains a vector v with $\overline{\text{ZP}}(v) = \mathcal{S}$. So, we only need to verify the upper bound on the index k in the "only if" part, namely to verify that, under the pattern reachability assumption, the index k in the proposition's statement may be chosen smaller than $|\mathcal{S}|$.

We prove this result by induction on the cardinality s of the set $\mathcal{S}$. If $s = 1$, then $\mathcal{S} = \{i\}$, for some index $i \in \langle n \rangle$, and we are dealing with monomial reachability. As we have seen in Proposition 1 iii), monomial reachability ensures the existence of indices $j \in \mathcal{P}$ and $r \in \langle m \rangle$, such that $\overline{\text{ZP}}(A_j \mathbf{e}_i) = \{i\}$ as well as $\overline{\text{ZP}}(B_j \mathbf{e}_r) = \{i\}$. Consequently, $\overline{\text{ZP}}(e^{A_j} B_j \mathbf{e}_r) = \{i\}$, and this ensures, by Lemma 1, that we need a constant switching sequence (namely, $\sigma(t) = j$ for every $t \geq 0$) in order to reach vectors with a single positive entry.

We assume, now, by induction, that given any subset $\mathcal{S}'$ of $\langle n \rangle$, with $|\mathcal{S}'| < s$, there exists a vector $v' \geq 0$, with $|\overline{\text{ZP}}(v')| = \mathcal{S}'$, that belongs to the reachability cone of a switching sequence taking no more than $|\mathcal{S}'|$ values (equivalently, commuting no more than $|\mathcal{S}'| - 1$ times). We aim to prove that the result extends to all subsets $\mathcal{S}$ of $\langle n\rangle$, with $|\mathcal{S}| = s$. Indeed, consider the smallest nonnegative index k for which indices $i_0, i_1, \dots, i_k$ in $\mathcal{P}$, and a positive vector v, with $\overline{\mathrm{ZP}}(v) = \mathcal{S}$, can be found, such that

$$v = e^{A_{i_k}} e^{A_{i_{k-1}}} \dots e^{A_{i_0}} B_{i_0}\bar{u}_0 + \dots + e^{A_{i_k}} e^{A_{i_{k-1}}} B_{i_{k-1}}\bar{u}_{k-1} + e^{A_{i_k}} B_{i_k}\bar{u}_k, \tag{4}$$

for suitable nonnegative vectors $\bar{u}_i \geq 0$. Since each of these terms is left multiplied by $e^{A_{i_k}}$, it follows that v can be expressed as $v = e^{A_{i_k}}\mathcal{B}_k$, with

$$\mathcal{B}_k := e^{A_{i_{k-1}}} \dots e^{A_{i_1}} e^{A_{i_0}} B_{i_0}\bar{u}_0 + \dots + e^{A_{i_{k-1}}} B_{i_{k-1}}\bar{u}_{k-1} + B_{i_k}\bar{u}_k.$$

By Lemma A.3, then, $\mathcal{S} = \overline{\mathrm{ZP}}(v) = \overline{\mathrm{ZP}}(e^{A_{i_k}}\mathcal{B}_k)$ implies $\mathcal{S} \supseteq \mathcal{S}' := \overline{\mathrm{ZP}}(\mathcal{B}_k)$. Clearly, it cannot be $\mathcal{S} = \mathcal{S}'$, otherwise condition

$$\mathcal{B}_k \in \mathrm{Cone}(\mathcal{R}(i_0, i_1, \dots, i_{k-1})), \qquad \text{with } \overline{\mathrm{ZP}}(\mathcal{B}_k) = \mathcal{S},$$

would contradict the minimality assumption on the index k. So, it must be $\mathcal{S} \supset \mathcal{S}'$. By the inductive assumption, there exists a vector v', with $|\overline{\mathrm{ZP}}(v')| = |\mathcal{S}'| < |\mathcal{S}| = s$, that can be reached by resorting to a switching sequence $(j_0, j_1, \dots, j_l)$ with $l + 1 \leq |\mathcal{S}'| \leq s - 1$, i.e.

$$v' = e^{A_{j_l}} \dots e^{A_{j_1}} e^{A_{j_0}} B_{j_0} u_0 + \dots + e^{A_{j_l}} B_{j_l} u_l$$

for suitable $u_i \geq 0$. Since $\mathcal{S} = \overline{\mathrm{ZP}}(e^{A_{i_k}}\mathcal{B}_k) = \overline{\mathrm{ZP}}(e^{A_{i_k}}(v' + B_{i_k}0))$, we have found a switching sequence $(j_0, j_1, \dots, j_l, i_k)$ taking no more than s values that allows to reach the pattern $\mathcal{S}$.

We are, now, in a position to provide the final characterization of pattern reachability. Even though the result could be easily given for multiple input systems, for the sake of simplicity we state it for single input systems.

Proposition 3. *A single-input positive switched system (2) is pattern reachable if and only if for every set $\mathcal{S} \subseteq \langle n\rangle$ there exist an integer $\ell \leq |\mathcal{S}|$, indices $j_1, j_2, \dots, j_\ell$, and a subset sequence $\mathcal{S}_0 \subseteq \mathcal{S}_1 \subset \mathcal{S}_2 \subset \dots \subset \mathcal{S}_\ell = \mathcal{S}$, such that*

$$\overline{\mathrm{ZP}}(e^{A_{j_h}}\mathbf{e}_{\mathcal{S}_{h-1}}) = \mathcal{S}_h, \qquad \forall\ h \in \langle \ell\rangle \tag{5}$$

$$\emptyset \neq \overline{\mathrm{ZP}}(B_{j_1}) \subseteq \mathcal{S}_1. \tag{6}$$

Proof. If system (2) is pattern reachable, then, by Proposition 2, for every $\mathcal{S} \subseteq \langle n\rangle$ there exist $k < |\mathcal{S}|$, indices $i_0, i_1, \dots, i_k \in \mathcal{P}$, and a positive vector v, with $\overline{\mathrm{ZP}}(v) = \mathcal{S}$, such that

$$v = e^{A_{i_k}} e^{A_{i_{k-1}}} \dots e^{A_{i_1}} e^{A_{i_0}} B_{i_0} u_0 + \dots + e^{A_{i_k}} e^{A_{i_{k-1}}} B_{i_{k-1}} u_{k-1} + e^{A_{i_k}} B_{i_k} u_k,$$

where, w.l.o.g., the scalars $u_0, u_1, \dots, u_k$ take values in $\{0, 1\}$, and k is the smallest such index. Set $\ell := \min\{d \geq 1 : u_{k-d+1} = 1\}$, and $j_h := i_{k-\ell+h}$ for $h = 1, 2, \dots, \ell$. Then:

$$v = e^{A_{j_\ell}} e^{A_{j_{\ell-1}}} \dots e^{A_{j_1}} \Big[e^{A_{i_{k-\ell}}} \dots e^{A_{i_0}} B_{i_0} u_0 + \dots + e^{A_{i_{k-\ell}}} B_{i_{k-\ell}} u_{k-\ell} + B_{j_1} u_{k-\ell+1} \Big].$$

Set $\mathcal{B}_0 := e^{A_{i_{k-\ell}}} \dots e^{A_{i_1}} e^{A_{i_0}} B_{i_0} u_0 + \dots + e^{A_{i_{k-\ell}}} B_{i_{k-\ell}} u_{k-\ell} + B_{j_1} u_{k-\ell+1}$ and $\mathcal{B}_h := e^{A_{j_h}}\mathcal{B}_{h-1}$, $h = 1, 2, \dots, \ell$. Notice that $\mathcal{B}_\ell = v$. Set, finally, $\mathcal{S}_h := \overline{\mathrm{ZP}}(\mathcal{B}_h)$. By recursively applying Lemma A.3, we can prove that

$$\mathcal{S} = \overline{\mathrm{ZP}}(v) = \overline{\mathrm{ZP}}(\mathcal{B}_\ell) \supseteq \overline{\mathrm{ZP}}(\mathcal{B}_{\ell-1}) \cdots \supseteq \overline{\mathrm{ZP}}(\mathcal{B}_1) \supseteq \overline{\mathrm{ZP}}(B_{j_1}).$$

On the other hand, all the inequalities $\mathcal{S}_h \supseteq \mathcal{S}_{h-1}, h = 2, 3, \dots, \ell$, must be strict, otherwise the sequence could be shortened. Therefore $\ell \leq |\mathcal{S}|$ and (5) holds. Finally, condition $\mathcal{S}_1 \supseteq \mathcal{S}_0 = \overline{\mathrm{ZP}}(\mathcal{B}_0) \supseteq \overline{\mathrm{ZP}}(B_{j_1})$ ensures that (6) holds.

Assume, now, that (5)-(6) hold. We prove that the system is pattern reachable by induction on $s := |\mathcal{S}|$. To this end, consider, first the case $s = 1$, namely $\mathcal{S} = \{i\}$ for some $i \in \langle n \rangle$. If so, $\ell = 1$ and there exists an index j_1 and sets $\mathcal{S}_0 = \mathcal{S}_1 = \mathcal{S}$ such that $\overline{\mathrm{ZP}}(e^{A_{j_1}}\mathbf{e}_i) = \{i\}$, and $\emptyset \neq \overline{\mathrm{ZP}}(e^{B_{j_1}}) = \{i\}$. So, by Proposition 1, the system is monomially reachable.

Suppose, now, that for every set $\mathcal{S}'$ of cardinality smaller than s, there exists a reachable vector v' with $\overline{\mathrm{ZP}}(v') = \mathcal{S}'$. Consider an arbitrary set $\mathcal{S}$ of cardinality s and let $\ell, j_1, \dots, j_\ell, \mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_\ell$ be the corresponding indices and sets as they appear in the proposition's statement. Consider the (possibly empty) set $\mathcal{S}' = \mathcal{S}_0 \setminus \overline{\mathrm{ZP}}(B_{j_1})$ whose cardinality is smaller than s. By the inductive assumption, there exist indices $i_0, i_1, \dots, i_k$ in $\mathcal{P}$ such that the cone generated by the columns of the reachability matrix $\mathcal{R}(i_0, i_1, \dots, i_k)$ includes a vector v' with $\overline{\mathrm{ZP}}(v') = \mathcal{S}'$ (if $\mathcal{S}' = \emptyset$, simply choose $v' = 0$). Let u be a binary vector such that $\overline{\mathrm{ZP}}(\mathcal{R}(i_0, i_1, \dots, i_k)u) = \mathcal{S}'$. Then, the vector

$$v = e^{A_{j_\ell}} e^{A_{j_{\ell-1}}} \dots e^{A_{j_1}} \left[\mathcal{R}(i_0, i_1, \dots, i_k)u + B_{j_1}\right]$$

satisfies $\overline{\mathrm{ZP}}(v) = \mathcal{S}$. This ensures that a vector with nonzero pattern $\mathcal{S}$ is reachable through the switching sequence $(i_0, i_1, \dots, i_k, j_1, j_2, \dots, j_\ell)$.

5 Necessary and/or sufficient conditions for reachability

As a result of the pattern reachability analysis, given a single-input switched system (2), switching among p positive subsystems $(A_i, b_i), i \in \mathcal{P}$, a positive vector v, with $\overline{\mathrm{ZP}}(v) = \mathcal{S}$, is reachable only if there is an index $j = j(\mathcal{S}) \in \mathcal{P}$ such that $\overline{\mathrm{ZP}}(e^{A_{j(\mathcal{S})}}\mathbf{e}_\mathcal{S}) = \mathcal{S}$. Consequently, a necessary condition for reachability is that, for every $\mathcal{S} \subseteq \langle n \rangle$, the set $\mathcal{I}_\mathcal{S} := \{i \in \langle p \rangle : \overline{\mathrm{ZP}}(e^{A_i}\mathbf{e}_\mathcal{S}) = \mathcal{S}\} \neq \emptyset$. Note that, if $\mathcal{S} = \langle n \rangle$, $\mathcal{I}_\mathcal{S} = \mathcal{P}$, and the previous condition is trivially satisfied.

In this section we focus on the derivation of necessary and/or sufficient conditions for reachability, by restricting our attention to single-input systems and, occasionally, on single-input systems of size n which commute among $p = n$ subsystems. As we have seen, this represents the minimum number of subsystems among which a single-input positive switched system has to commute in order to be reachable. The first result of the section is a sufficient condition for reachability.

Proposition 4. *Consider a positive switched system (2), switching among p single-input subsystems $(A_i, b_i), i \in \mathcal{P}$. If $\forall\, \mathcal{S} \subseteq \langle n \rangle$, $\exists\, j(\mathcal{S}) \in \mathcal{I}_\mathcal{S}$ such that*

$$\overline{\mathrm{ZP}}(e^{A_{j(\mathcal{S})}}\mathbf{e}_\mathcal{S}) = \mathcal{S} \text{ and } \overline{\mathrm{ZP}}(b_{j(\mathcal{S})}) \subseteq \mathcal{S}, \text{ with } |\overline{\mathrm{ZP}}(b_{j(\mathcal{S})})| = 1,$$

then the switched system is reachable.

Proof. Given any positive vector $v \in \mathbb{R}^n_+$, set $r := |\overline{\mathrm{ZP}}(v)|$. Set, now, $\mathcal{S}_r := \overline{\mathrm{ZP}}(v)$, and let $j(\mathcal{S}_r)$ be an index which makes the Proposition assumption satisfied, and hence $\overline{\mathrm{ZP}}(e^{A_{j(\mathcal{S}_r)}}\mathbf{e}_{\mathcal{S}_r}) = \mathcal{S}_r$, and $\{i_r\} := \overline{\mathrm{ZP}}(b_{j(\mathcal{S}_r)}) \subseteq \mathcal{S}_r$. For each $h \in \langle r-1\rangle$, we may recursively define sets $\mathcal{S}_h$ and indices i_h, as

$$\mathcal{S}_h := \mathcal{S}_{h+1} \setminus \{i_{h+1}\}, \qquad \{i_h\} := \overline{\mathrm{ZP}}(b_{j(\mathcal{S}_h)}).$$

Notice that, by the way the sets $\mathcal{S}_h$ are defined, $|\mathcal{S}_h| = h$. Moreover, when $h \neq q$ we have $j(\mathcal{S}_h) \neq j(\mathcal{S}_q)$. Now, we show that by suitably choosing a final time instant $t_r > 0$, the values of the switching instants $t_i, i = 0, 1, \dots, r-1$, with $0 = t_0 < \dots < t_{r-1} < t_r$, and positive input values $\bar{u}_i$ in every time interval $[t_{i-1}, t_i)$, we may ensure that

$$v = e^{A_{j(\mathcal{S}_r)}(t_r - t_{r-1})} e^{A_{j(\mathcal{S}_{r-1})}(t_{r-1} - t_{r-2})} \dots e^{A_{j(\mathcal{S}_2)}(t_2 - t_1)} \int_{t_0}^{t_1} e^{A_{j(\mathcal{S}_1)}(t_1 - \tau)} d\tau \, b_{j(\mathcal{S}_1)} \bar{u}_1 + \dots + \int_{t_{r-1}}^{t_r} e^{A_{j(\mathcal{S}_r)}(t_r - \tau)} d\tau \, b_{j(\mathcal{S}_r)} \bar{u}_r \qquad (7)$$

By the previous considerations, every term in (7) has a nonzero pattern included in $\mathcal{S}$. Moreover, by Lemma A.2, it is easy to conclude that, since every exponential matrix can be made as close as we want to the identity matrix and since $b_{j(\mathcal{S}_\ell)}$ is an i_ℓ-monomial vector, then each positive term

$$e^{A_{j(\mathcal{S}_r)}(t_r - t_{r-1})} e^{A_{j(\mathcal{S}_{r-1})}(t_{r-1} - t_{r-2})} \dots e^{A_{j(\mathcal{S}_{\ell+1})}(t_{\ell+1} - t_\ell)} \int_{t_{\ell-1}}^{t_\ell} e^{A_{j(\mathcal{S}_\ell)}(t_\ell - \tau)} d\tau \, b_{j(\mathcal{S}_\ell)} \qquad (8)$$

can be made as close as we want to the monomial vector $\mathbf{e}_{i_\ell}$ (and, of course, its nonzero pattern is included in $\mathcal{S}$), by suitably choosing the time intervals between two consecutive switching instants sufficiently small. If we assume that the switching time instants are given, in order to ensure that the aformentioned terms are desired approximations of selected monomial vectors, the only values we have to choose are the constant values $\bar{u}_\ell$, and the problem we have to solve can be seen as that of solving an algebraic equation of the following type: $A\bar{u} = \bar{v}$, where $\bar{v} \in \mathbb{R}^r_+$ is the (strictly positive) vector consisting of the nonzero entries of v, $A \in \mathbb{R}^{r \times r}_+$ is the positive matrix whose columns are those terms (8) (approximating the monomial vectors) which pertain to the indices in $\mathcal{S}$, and $\bar{u}$ is the vector containing the associated input vectors $\bar{u}_\ell$. By Lemma A.4, this linear equation admits a positive solution, and hence the vector $v \in \mathbb{R}^n_+$ is reachable.

Remark. It is worthwhile noticing that, when the previous sufficient condition holds, all states in $\mathbb{R}^n_+$ are reached by resorting to a suitable switching sequence $(i_1, i_2, \dots, i_k)$ and by applying a nonnegative input which is surely nonzero during the last switching interval (when the system has commuted to the i_kth subsystem). Of course, this is not the general case, and a state may be reached even by eventually leaving the system freely evolve (meaning that no soliciting input is applied during the last part of the time interval), meanwhile commuting from one subsystem to another. Consequently, the above condition is only sufficient for reachability, as shown in the following example.

Example 1. Consider the positive switched system (2), switching among the following three subsystems

$$(A_1, B_1) = \left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right) \qquad (A_2, B_2) = \left(\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right)$$

$$(A_3, B_3) = \left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right).$$

Note that the hypothesis of Proposition 4 is fulfilled $\forall \mathcal{S} \neq \{1, 2\}$. Therefore, in order to show that the switched system is reachable, we only need to prove that every vector v with $\overline{\text{ZP}}(v) = \{1, 2\}$ is reachable. Observe now that

$$e^{A_3 t} = \begin{bmatrix} e^t & t\,e^t & 0 \\ 0 & e^t & 0 \\ 0 & 0 & e^t \end{bmatrix}.$$

Hence, given $v = \begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix}$, with $v_1,\, v_2 \neq 0$, set $t = \frac{v_1}{v_2} + 1,\ t_1 = 1,\ t_0 = 0$. Introduce the piece-wise constant input function and the switching sequence:

$$u(t) = \begin{cases} \frac{v_2}{(e-1)e^{\frac{v_1}{v_2}}}, & \text{for } 0 \leq t < t_1; \\ 0, & \text{for } t_1 \leq t < t; \end{cases} \qquad \sigma(t) = \begin{cases} 2, & \text{for } 0 \leq t < t_1; \\ 3, & \text{for } t_1 \leq t < t. \end{cases}$$

By referring to equation (3), we get

$$\begin{aligned} x(t) &= e^{A_3(t-t_1)}\int_{t_0}^{t_1} e^{A_2(t_1-\tau)} B_2 u(\tau) d\tau + \int_{t_1}^{t} e^{A_3(t-\tau)} B_3 u(\tau) d\tau \\ &= e^{A_3 \frac{v_1}{v_2}} \int_0^1 \begin{bmatrix} 0 \\ e^{1-\tau} \\ 0 \end{bmatrix} d\tau \, \frac{v_2}{(e-1)e^{\frac{v_1}{v_2}}} + 0 = \frac{v_2\,(e-1)}{(e-1)e^{\frac{v_1}{v_2}}} \begin{bmatrix} \frac{v_1}{v_2} e^{\frac{v_1}{v_2}} \\ e^{\frac{v_1}{v_2}} \\ 0 \end{bmatrix} = v. \end{aligned}$$

As a consequence, the switched system is reachable.

Aiming to provide an equivalent condition for reachability, we first introduce a technical lemma which allows us to use, when dealing with single-input systems, only piece-wise constant input signals.

Lemma 2. *Consider an n-dimensional monomially reachable positive switched system (2), switching among n single-input subsystems $(A_i, b_i), i \in \langle n \rangle$, with*[1]

$$A_i \mathbf{e}_i = \alpha_i \mathbf{e}_i, \qquad b_i = \beta_i \mathbf{e}_i, \qquad \exists\ \alpha_i \geq 0 \text{ and } \beta_i > 0. \tag{9}$$

[1] Notice that this assumption is by no means restrictive, since, by Proposition 1, we can always reduce ourselves to this case by means of a simple relabeling.

Given $t > 0$, $v \in \mathbb{R}^n_+$, $k \in \mathbb{N}$, time instants $0 = t_0 < t_1 < \ldots < t_k < t$ and indices $i_0, i_1, \ldots, i_k \in \langle n \rangle$, if there exists a nonnegative input $u(\cdot)$ such that:

$$v = e^{A_{i_k}(t-t_k)} e^{A_{i_{k-1}}(t_k - t_{k-1})} \ldots e^{A_{i_1}(t_2-t_1)} \int_{t_0}^{t_1} e^{A_{i_0}(t_1-\tau)} b_{i_0} u(\tau) d\tau \\ + \ldots + \int_{t_k}^{t} e^{A_{i_k}(t-\tau)} b_{i_k} u(\tau) d\tau, \tag{10}$$

then there exists a piece-wise constant input $u(\cdot)$, taking some suitable constant value $u_i \geq 0$ in every time interval $[t_i, t_{i+1})$, such that

$$v = e^{A_{i_k}(t-t_k)} e^{A_{i_{k-1}}(t_k - t_{k-1})} \ldots e^{A_{i_1}(t_2-t_1)} \int_{t_0}^{t_1} e^{A_{i_0}(t_1-\tau)} b_{i_0} d\tau \cdot u_0 \\ + \ldots + \int_{t_k}^{t} e^{A_{i_k}(t-\tau)} b_{i_k} d\tau \cdot u_k. \tag{11}$$

Proof. By the assumption (9), $e^{A_i t} b_i = e^{\alpha_i t} \beta_i \mathbf{e}_i, \forall t \in \mathbb{R}_+$. Consequently,

$$\begin{aligned} \int_{t_i}^{t_{i+1}} e^{A_i(t_{i+1}-\tau)} b_i u(\tau) d\tau &= \int_{t_i}^{t_{i+1}} e^{\alpha_i(t_{i+1}-\tau)} \beta_i \mathbf{e}_i u(\tau) d\tau \\ &= \left[\int_{t_i}^{t_{i+1}} e^{\alpha_i(t_{i+1}-\tau)} u(\tau) d\tau \right] \cdot \beta_i \mathbf{e}_i, \end{aligned} \tag{12}$$

where the term inside the square brackets is a nonnegative number. But then, a nonnegative coefficient u_i can always be found such that

$$\int_{t_i}^{t_{i+1}} e^{\alpha_i(t_{i+1}-\tau)} u(\tau) d\tau = \int_{t_i}^{t_{i+1}} e^{\alpha_i(t_{i+1}-\tau)} d\tau \cdot u_i. \tag{13}$$

This immediately implies the lemma statement.

From the previous lemma, we get the following Proposition.

Proposition 5. *Consider an n-dimensional positive switched system (2), switching among n single-input systems $(A_i, b_i), i \in \langle n \rangle$, and suppose that for every index $i \in \langle n \rangle$ the pair (A_i, b_i) satisfies (9). The system is reachable if and only if for every positive vector $v \in \mathbb{R}^n_+$ there exist $k \in \mathbb{N}$, strictly positive intervals $\tau_1, \ldots, \tau_k$ and switching values $i_0, i_1, \ldots, i_k \in \langle n \rangle$, such that*

$$\begin{aligned} v \in\ & \mathrm{Cone}[e^{A_{i_k}\tau_k} b_{i_k} | e^{A_{i_k}\tau_k} e^{A_{i_{k-1}}\tau_{k-1}} b_{i_{k-1}} | \ldots | e^{A_{i_k}\tau_k} \ldots e^{A_{i_1}\tau_1} e^{A_{i_0}\tau_0} b_{i_0}] \\ &= \mathrm{Cone}[\mathbf{e}_{i_k} | e^{A_{i_k}\tau_k} \mathbf{e}_{i_{k-1}} | \ldots | e^{A_{i_k}\tau_k} \ldots e^{A_{i_1}\tau_1} \mathbf{e}_{i_0}]. \end{aligned}$$

Proof. By the assumption on the n subsystems (A_i, b_i), the identity

$$\mathrm{Cone}[e^{A_{i_k}\tau_k} b_{i_k} | e^{A_{i_k}\tau_k} e^{A_{i_{k-1}}\tau_{k-1}} b_{i_{k-1}} | \ldots | e^{A_{i_k}\tau_k} \ldots e^{A_{i_1}\tau_1} e^{A_{i_0}\tau_0} b_{i_0}] = \\ \mathrm{Cone}[\mathbf{e}_{i_k} | e^{A_{i_k}\tau_k} \mathbf{e}_{i_{k-1}} | \ldots | e^{A_{i_k}\tau_k} \ldots e^{A_{i_1}\tau_1} \mathbf{e}_{i_0}]$$

immediately follows. So, in the sequel, we only refer to the latter expression.

[Necessity] If the system is reachable, then $\forall\, v \in \mathbb{R}^n_+$ there exist parameters t, t_j, i_j (endowed with suitable properties) and an input $u(\cdot) \in \mathbb{R}_+$ such that:

$$v = e^{A_{i_k}(t-t_k)} e^{A_{i_{k-1}}(t_k - t_{k-1})} \dots e^{A_{i_1}(t_2-t_1)} \int_{t_0}^{t_1} e^{A_{i_0}(t_1-\tau)} b_{i_0} u(\tau) d\tau \\ + \dots + \int_{t_k}^{t} e^{A_{i_k}(t-\tau)} b_{i_k} u(\tau) d\tau. \tag{14}$$

But then, by Lemma 2, this means that there exist suitable $u_j \geq 0$ such that

$$v = e^{A_{i_k}(t-t_k)} e^{A_{i_{k-1}}(t_k - t_{k-1})} \dots e^{A_{i_1}(t_2-t_1)} \int_{t_0}^{t_1} e^{A_{i_0}(t_1-\tau)} b_{i_0} d\tau \cdot u_0 \\ + \dots + \int_{t_k}^{t} e^{A_{i_k}(t-\tau)} b_{i_k} d\tau \cdot u_k = e^{A_{i_k}\tau_k} e^{A_{i_{k-1}}\tau_{k-1}} \dots e^{A_{i_1}\tau_1} \mathbf{e}_{i_0} c_{i_0} + \dots + \mathbf{e}_{i_k} c_{i_k},$$

where $t_{k+1} := t$, $\tau_j := t_{j+1} - t_j$ and $c_{i_j} = \int_{t_j}^{t_{j+1}} e^{\alpha_{i_j}(t_{j+1}-\tau)} \beta_{i_j} d\tau \cdot u_j$. Hence,

$$v \in \text{Cone}[\mathbf{e}_{i_k} | e^{A_{i_k}\tau_k} \mathbf{e}_{i_{k-1}} | \dots | e^{A_{i_k}\tau_k} \dots e^{A_{i_1}\tau_1} \mathbf{e}_{i_0}]. \tag{15}$$

[Sufficiency] Conversely, suppose that for every positive vector v we can find $k \in \mathbb{N}$, intervals $\tau_1, \dots, \tau_k > 0$ and switching values $i_0, i_1, \dots, i_k \in \langle n \rangle$, such that (15) holds. Let $c_{i_j}, j = 0, 1, \dots, k$, be nonnegative coefficients such that $v = e^{A_{i_k}\tau_k} e^{A_{i_{k-1}}\tau_{k-1}} \dots e^{A_{i_1}\tau_1} \mathbf{e}_{i_0} c_{i_0} + \dots + \mathbf{e}_{i_k} c_{i_k}$. Set, now, $t_0 := 0, t_1 := 1$ and $t_{j+1} := t_j + \tau_j$ for every $j \in \langle k \rangle$. Then, by assuming $u_j := \dfrac{c_{i_j}}{\displaystyle\int_{t_j}^{t_{j+1}} e^{\alpha_{i_j}(t_{j+1}-\tau)} d\tau \beta_{i_j}}$, we get

$$e^{A_{i_k}(t_{k+1}-t_k)} e^{A_{i_{k-1}}(t_k - t_{k-1})} \dots e^{A_{i_1}(t_2-t_1)} \int_{t_0}^{t_1} e^{A_{i_0}(t_1-\tau)} b_{i_0}\ u_0\ d\tau + \dots + \\ + \int_{t_k}^{t_{k+1}} e^{A_{i_k}(t-\tau)} b_{i_k}\ u_k\ d\tau = e^{A_{i_k}\tau_k} e^{A_{i_{k-1}}\tau_{k-1}} \dots e^{A_{i_1}\tau_1} \mathbf{e}_{i_0} c_{i_0} + \dots + \mathbf{e}_{i_k} c_{i_k} = v$$

thus proving that v is reachable.

Proposition 6. *Given an n-dimensional positive switched system (2), commuting among n single-input subsystems $(A_i, b_i), i = 1, 2, \dots, n$, the following facts are equivalent:*

i) the switched system (2) is reachable;
ii) for every proper subset $\mathcal{S} \subset \langle n \rangle$ we have:
  iia) if $|\mathcal{S}| = 1$, then $\exists\, j(\mathcal{S}) \in \mathcal{I}_{\mathcal{S}}$ such that ($\overline{\text{ZP}}(e^{A_{j(\mathcal{S})}} \mathbf{e}_{\mathcal{S}}) = \mathcal{S}$ and) $\overline{\text{ZP}}(b_{j(\mathcal{S})}) = \mathcal{S}$;
  iib) if $|\mathcal{S}| > 1$, then $\mathcal{I}_{\mathcal{S}} \neq \emptyset$, and either
    1. $\exists\, j(\mathcal{S}) \in \mathcal{I}_{\mathcal{S}}$ such that[2] *$\overline{\text{ZP}}(b_{j(\mathcal{S})}) \subset \mathcal{S}$,* or

[2] Note that condition iia) together with the fact that we are switching among n subsystems ensure that $\overline{\text{ZP}}(b_{j(\mathcal{S})}) \neq \emptyset$.

2. *$\forall\, v \in \mathbb{R}^n_+$, with $\overline{\mathrm{ZP}}(v) = \mathcal{S}$, there exist $m \in \mathbb{N}$, $\tau_1, \dots, \tau_m > 0$ and $i_1, \dots, i_m \in \mathcal{I}_{\mathcal{S}}$, such that v can be obtained as the nonnegative combination of no more than $|\mathcal{S}| - 1$ columns of $e^{A_{i_m}\tau_m} \dots e^{A_{i_1}\tau_1} P_{\mathcal{S}}$, where $P_{\mathcal{S}}$ is the selection matrix which selects all the columns corresponding to the indices [3] appearing in $\mathcal{S}$.*

Proof. i) ⇒ ii) Suppose, first, that system (2) is reachable. Since condition iia) is equivalent to monomial reachability, its necessity has already been proved, and we may assume, as usual, that each pair (A_i, b_i) satisfies (9).

Now, let $\mathcal{S}$ be any subset of $\langle n \rangle$ with cardinality $|\mathcal{S}| > 1$, and let v be any positive vector with nonzero pattern $\overline{\mathrm{ZP}}(v) = \mathcal{S}$. By the reachability assumption and by Proposition 5, there exist $k \in \mathbb{N}$, positive intervals $\tau_0, \tau_1, \dots, \tau_k$, switching values $i_0, i_1, \dots, i_k \in \langle n \rangle$ (with $i_j \neq i_{j+1}$ w.l.o.g.), and nonnegative coefficients c_{i_j}, $j = 0, 1, \dots, k$, such that

$$\begin{aligned} v &= e^{A_{i_k}\tau_k} \dots e^{A_{i_1}\tau_1} e^{A_{i_0}\tau_0} b_{i_0} c_{i_0} + .. + e^{A_{i_k}\tau_k} e^{A_{i_{k-1}}\tau_{k-1}} b_{i_{k-1}} c_{i_{k-1}} + e^{A_{i_k}\tau_k} b_{i_k} c_{i_k} \\ &= e^{A_{i_k}\tau_k} \left[e^{A_{i_{k-1}}\tau_{k-1}} \dots e^{A_{i_1}\tau_1} e^{A_{i_0}\tau_0} b_{i_0} c_{i_0} + \dots + b_{i_k} c_{i_k} \right]. \end{aligned} \tag{16}$$

Clearly, by Lemma A.3, $\overline{\mathrm{ZP}}(e^{A_{i_k}} \mathbf{e}_{\mathcal{S}}) = \mathcal{S}$. So, the set $\mathcal{I}_{\mathcal{S}}$ is nonempty. If there exist $j(\mathcal{S}) \in \mathcal{I}_{\mathcal{S}}$ such that $\overline{\mathrm{ZP}}(b_{j(\mathcal{S})}) \subset \mathcal{S}$ we fall in case 1. of iib). Suppose, now, that $\forall\, j(\mathcal{S}) \in \mathcal{I}_{\mathcal{S}}, \overline{\mathrm{ZP}}(b_{j(\mathcal{S})}) \not\subset \mathcal{S}$. Consequently, in (16), $c_{i_k} = 0$, and hence (16) becomes $v = e^{A_{i_k}\tau_k} \mathcal{B}_k$, with

$$\mathcal{B}_k := e^{A_{i_{k-1}}\tau_{k-1}} \dots e^{A_{i_1}\tau_1} e^{A_{i_0}\tau_0} b_{i_0} c_{i_0} + e^{A_{i_{k-1}}\tau_{k-1}} b_{i_{k-1}} c_{i_{k-1}}. \tag{17}$$

From Lemma A.3, $\mathcal{S}_k := \overline{\mathrm{ZP}}(\mathcal{B}_k) \subseteq \mathcal{S}$. Now, either $\mathcal{S}_k \subsetneq \mathcal{S}$ or $\mathcal{S}_k = \mathcal{S}$.

1. If $\mathcal{S}_k \subsetneq \mathcal{S}$, then v lies on a face of $\mathrm{Cone}(e^{A_{j_i}\tau_k} P_{\mathcal{S}}), \exists\, \tau_k > 0$, namely it can be obtained by combining no more than $|\mathcal{S}| - 1$ columns of $e^{A_{j_i}\tau_k} P_{\mathcal{S}}$.
2. If $\mathcal{S}_k = \mathcal{S}$ then $\overline{\mathrm{ZP}}(e^{A_{i_{k-1}}} \mathbf{e}_{\mathcal{S}}) = \mathcal{S}$, which, in turn, implies that $i_{k-1} \in \mathcal{I}_{\mathcal{S}}$. But since $\overline{\mathrm{ZP}}(b_{i_{k-1}}) \not\subset \mathcal{S}$ (and hence $c_{i_{k-1}} = 0$), it follows that we can iterate this reasoning until we find some index ℓ such that $i_k, i_{k-1}, \dots, i_\ell \in \mathcal{I}_{\mathcal{S}}$, $i_{\ell-1} \notin \mathcal{I}_{\mathcal{S}}$ and $v = e^{A_{i_k}\tau_k} \dots e^{A_{i_\ell}\tau_\ell} \mathcal{B}_\ell$, for some suitable $\mathcal{B}_\ell$ with $\overline{\mathrm{ZP}}(\mathcal{B}_\ell) = \mathcal{S}_\ell \subsetneq \mathcal{S}$. Consequently, again, v can be obtained by combining no more than $|\mathcal{S}| - 1$ columns of $e^{A_{i_k}\tau_k} \dots e^{A_{i_\ell}\tau_\ell} P_{\mathcal{S}}$.

In both cases we fall in case 2. of iib).

ii) ⇒ i) Let us see, now, whether condition iia) and iib) are also sufficient for reachability. We prove this fact by induction on the cardinality of the nonzero pattern $|\mathcal{S}| = |\overline{\mathrm{ZP}}(v)|$ of any vector $v \in \mathbb{R}^n_+$. If $|\mathcal{S}| = 1$, condition iia), corresponding to monomial reachability, ensures that v is reachable.

[3] Notice that since $\overline{\mathrm{ZP}}(e^{A_{i_h}} \mathbf{e}_{\mathcal{S}}) = \mathcal{S}$ for $h = 1, 2, \dots, m$, the polyhedral cone $\mathrm{Cone}(e^{A_{i_m}\tau_m} \dots e^{A_{i_1}\tau_1} P_{\mathcal{S}})$ is generated by $|\mathcal{S}|$ linearly independent vectors whose nonzero pattern is included in $\mathcal{S}$. Indeed, $\mathrm{Cone}(P_{\mathcal{S}}^T e^{A_{i_m}\tau_m} \dots e^{A_{i_1}\tau_1} P_{\mathcal{S}})$ is a simplicial cone in $\mathbb{R}^{|\mathcal{S}|}_+$ and it coincides with $\mathrm{Cone}(e^{\tilde{A}_{i_m}\tau_m} \dots e^{\tilde{A}_{i_1}\tau_1})$, $\tilde{A}_{i_h} = P_{\mathcal{S}}^T A_{i_h} P_{\mathcal{S}}$.

Suppose now that, under the assumptions ii), every positive vector w, with $|\overline{\mathrm{ZP}}(w)| < s$, is reachable. Let v be a positive vector with $|\mathcal{S}| = |\overline{\mathrm{ZP}}(v)| = s$. If for the set $\mathcal{S}$ the case 1. applies, it has been already proved in Proposition 4 that v is reachable. Suppose now that only case 2. holds. Then $\exists\, m > 0$, $\exists\; i_1, \dots, i_m \in \mathcal{I}_{\mathcal{S}}$, $\exists\; \tau_1, \dots, \tau_m > 0$ such that v is obtained by combining no more than $r-1$ columns of $e^{A_{i_m}\tau_m} \dots e^{A_{i_1}\tau_1} P_{\mathcal{S}}$ and hence $\exists\, w \geq 0$, with $\overline{\mathrm{ZP}}(w) \subsetneq \mathcal{S}$ (and therefore $|\overline{\mathrm{ZP}}(w)| < r$), such that $v = e^{A_{i_m}\tau_m} \dots e^{A_{i_1}\tau_1} w$. Since vector w is reachable for hypothesis, also v is. Indeed, upon reaching w, we switch ordinately to the subsystems $i_1, i_2, \dots, i_m$ and leave the system freely evolve at each stage for a lapse of time equal to τ_i.

6 The asymptotic exponential cone and a sufficient condition for reachability

Not every condition provided in Proposition 6 can be easily verified. Specifically, there is no obvious way of testing whether indices $i_1, \dots, i_m$ and positive time intervals $\tau_1, \dots, \tau_m$ can be found, such that a given vector $v > 0$, with $\overline{\mathrm{ZP}}(v) = \mathcal{S}$, can be obtained by combining less than $|\mathcal{S}|$ columns of $e^{A_{i_m}\tau_m} \dots e^{A_{i_1}\tau_1} P_{\mathcal{S}}$.

As a first step toward the general problem solution, in this section we explore the restrictive case when $m = 1$. In other words, we are interested in investigating when a positive vector v, with $\mathcal{S} := \overline{\mathrm{ZP}}(v)$ of cardinality s, can be expressed as the positive combination of at most $s-1$ columns of $e^{A_{j(\mathcal{S})}\tau} P_{\mathcal{S}}$, for some suitable $j(\mathcal{S}) \in \mathcal{I}_{\mathcal{S}}$ and $\tau > 0$ (as usual, $P_{\mathcal{S}}$ is the selection matrix that singles out the columns indexed on $\mathcal{S}$). Notice, though, that this is equivalent to investigating when the restriction of v to its positive entries (which is a strictly positive vector, say $v_{\mathcal{S}}$, of size s) belongs to the boundary of the simplicial cone, $\mathrm{Cone}[P_{\mathcal{S}}^T e^{A_{j(\mathcal{S})}\tau} P_{\mathcal{S}}]$. Thus, our problem may be restated in a just apparently restrictive, but in fact absolutely general, formulation, by assuming $\mathcal{S} = \langle n \rangle$ and $I_{\mathcal{S}} = \mathcal{P}$ (and, consequently, $v_{\mathcal{S}} = v$).

Problem Statement: search for conditions ensuring that every strictly positive vector $v \in \mathbb{R}^n_+$ can be obtained as

$$v = e^{A_{j(\mathcal{S})}\tau}\mathbf{u}, \qquad \exists\; \tau > 0, \text{ and } \mathbf{u} \in \mathbb{R}^n_+ \text{ with } \mathrm{ZP}(\mathbf{u}) \neq \emptyset. \tag{18}$$

To solve this problem, we introduce a new concept which turns out to be very meaningful for our investigation.

Definition 4. *Given an $n \times n$ Metzler matrix A, we define its* asymptotic exponential cone, $\mathrm{Cone}_\infty\big(e^{At}\big)$*, as the polyhedral cone generated by the (normalized) vectors v_j^∞, i.e. the asymptotic directions of the columns of e^{At}, i.e.*

$$v_j^\infty := \lim_{t\to\infty} \frac{e^{At}\mathbf{e_j}}{\|e^{At}\mathbf{e_j}\|}, \qquad j = 1, 2, \dots, n.$$

It is not hard to prove that $\mathrm{Cone}_\infty\big(e^{At}\big)$ always exists, it is a polyhedral convex cone in $\mathbb{R}^n_+$, and it is never the empty set. Moreover, except for the case of a diagonal matrix A (in which case $\mathrm{Cone}(e^{At}) = \mathrm{Cone}_\infty(e^{At}) = \mathbb{R}^n_+$ for every $t \ge 0$), we have for every $0 < t_1 < t_2 < +\infty$:

$$\mathbb{R}^n_+ = \mathrm{Cone}(e^{A\cdot 0}) \supsetneqq \mathrm{Cone}(e^{At_1}) \supsetneqq \mathrm{Cone}(e^{At_2}) \supsetneqq \mathrm{Cone}_\infty(e^{At}). \tag{19}$$

Notice, also, that while $\mathrm{Cone}(e^{At})$ is a simplicial cone for every $t \ge 0$, $\mathrm{Cone}_\infty(e^{At})$ is typically not, since it is not solid.

We aim to provide a characterization of the boundary of $\mathrm{Cone}(e^{A\tau})$, as τ varies over the positive real numbers. This allows us to obtain the solution of the Problem Statement, and, finally, a new sufficient condition for reachability.

Lemma 3. *[10] Given an $n \times n$ Metzler matrix A and a strictly positive vector $v \in \mathbb{R}^n_+$, the following facts are equivalent:*

i) there exists $\tau > 0$ such that v belongs to $\partial\mathrm{Cone}(e^{A\tau})$;
ii) $v \notin \mathrm{Cone}_\infty(e^{At})$.

As an immediate corollary of Lemma 3, we get.

Corollary 1. *Given an $n \times n$ Metzler matrix A, the following are equivalent:*

i) $\forall\, v \gg 0$ there exists $\tau > 0$ such that v belongs to $\partial\mathrm{Cone}(e^{A\tau})$;
ii) $\mathrm{Cone}_\infty(e^{At}) \subseteq \partial\mathbb{R}^n_+$;
iii) there exists some index $r \in \langle n\rangle$ such that $r \in \mathrm{ZP}(v_j^\infty)$ for every $j \in \langle n\rangle$.

By resorting to Proposition 6 and Corollary 1, we get the following sufficient condition for reachability.

Proposition 7. *Consider an n-dimensional positive switched system (2), commuting among n single-input subsystems $(A_i, b_i), i = 1, 2, \dots, n$, and suppose that, for every proper subset $\mathcal{S} \subset \langle n\rangle$, $|\mathcal{I}_{\mathcal{S}}| = 1$, namely there exists a unique index $j(\mathcal{S}) \in \langle n\rangle$ such that $\overline{\mathrm{ZP}}(e^{A_{j(\mathcal{S})}}\mathbf{e}_{\mathcal{S}}) = \mathcal{S}$. Then the system is reachable if and only if the following two conditions hold:*

a) the system is monomially reachable;
b) for every $\mathcal{S}$, with $r := |\mathcal{S}| > 1$, either $\overline{\mathrm{ZP}}(b_{j(\mathcal{S})}) \subseteq \mathcal{S}$ or $\mathrm{Cone}_\infty\big(P_{\mathcal{S}}^T\, e^{A_{j(\mathcal{S})}t} P_{\mathcal{S}}\big) \subseteq \partial\mathbb{R}^r_+$, where $P_{\mathcal{S}}$ is the selection matrix which selects all the columns corresponding to the indices belonging to $\mathcal{S}$.

Proof. [Sufficiency] Notice, first, that if $\mathrm{Cone}_\infty\big(P_{\mathcal{S}}^T\, e^{A_{j(\mathcal{S})}t} P_{\mathcal{S}}\big) \subseteq \partial\mathbb{R}^r_+$, then, by Corollary 1, for every strictly positive vector $v_{\mathcal{S}} \in \mathbb{R}^r_+$ there exists $\tau > 0$ such that $v_{\mathcal{S}} \in \partial\mathrm{Cone}\big(P_{\mathcal{S}}^T\, e^{A_{j(\mathcal{S})}\tau} P_{\mathcal{S}}\big)$. So, as a consequence of condition $\overline{\mathrm{ZP}}(e^{A_{j(\mathcal{S})}}\mathbf{e}_{\mathcal{S}}) = \mathcal{S}$, for every positive vector $v \in \mathbb{R}^n_+$, with $\overline{\mathrm{ZP}}(v) = \mathcal{S}$, there exists $\tau > 0$ such that $v = e^{A_{j(\mathcal{S})}\tau} P_{\mathcal{S}}\mathbf{u}_{\mathcal{S}}$, with $\mathrm{ZP}(\mathbf{u}_{\mathcal{S}}) \ne \emptyset$. Consequently, assumptions a) and b) imply conditions iia) and iib) of Proposition 6, and reachability follows.

[Necessity] By comparing the proposition statement with the result of Proposition 6, it remains to prove that if the system is reachable and for every $\mathcal{S} \subset \langle n \rangle$ there is a single index $j(\mathcal{S})$ such that $\overline{\mathrm{ZP}}(e^{A_{j(\mathcal{S})}}\mathbf{e}_{\mathcal{S}}) = \mathcal{S}$, then condition $\emptyset \neq \overline{\mathrm{ZP}}(b_{j(\mathcal{S})}) \not\subseteq \mathcal{S}$ implies $\mathrm{Cone}_\infty\big(P_{\mathcal{S}}^T\, e^{A_{j(\mathcal{S})}t} P_{\mathcal{S}}\big) \subseteq \partial\mathbb{R}^r_+$. Indeed, let $v > 0$ with $\overline{\mathrm{ZP}}(v) = \mathcal{S}$. By referring to the same notation employed in the proof of Proposition 6, we have that $v = e^{A_{j(\mathcal{S})}\tau_k}\mathcal{B}_k$, with $\tau_k > 0$ and

$$\mathcal{B}_k := e^{A_{i_{k-1}}\tau_{k-1}} \dots e^{A_{i_1}\tau_1} e^{A_{i_0}\tau_0} b_{i_0} c_{i_0} + \dots + e^{A_{i_{k-1}}\tau_{k-1}} b_{i_{k-1}} c_{i_{k-1}}, \qquad (20)$$

for suitable indices i_ℓ (with $i_\ell \neq i_{\ell+1}$), positive time intervals τ_ℓ and nonnegative coefficients c_ℓ. From Lemma A.3, it follows that $\mathcal{S}_k := \overline{\mathrm{ZP}}(\mathcal{B}_k) \subseteq \mathcal{S}$, and the uniqueness of $j(\mathcal{S})$ ensures that $\mathcal{S}_k \subsetneq \mathcal{S}$. So, $v = e^{A_{j(\mathcal{S})}\tau_k} P_{\mathcal{S}}\mathbf{u}_{\mathcal{S}}$, $\exists\ \mathbf{u}_{\mathcal{S}} \geq 0$, with $\mathrm{ZP}(\mathbf{u}_{\mathcal{S}}) \neq \emptyset$. But since this must be true for every vector $v \in V_{\mathcal{S}} := \{v : \overline{\mathrm{ZP}}(v) = \mathcal{S}\}$, then every $v_{\mathcal{S}} \in \mathbb{R}^r_+$, with $v_{\mathcal{S}} \gg 0$, must lie on the boundary of $\mathrm{Cone}(P_{\mathcal{S}}^T e^{A_{j(\mathcal{S})}\tau} P_{\mathcal{S}})$ for some $\tau = \tau(v_{\mathcal{S}}) > 0$. By Corollary 1, then, it must be $\mathrm{Cone}_\infty\big(P_{\mathcal{S}}^T\, e^{A_{j(\mathcal{S})}t} P_{\mathcal{S}}\big) \subseteq \partial\mathbb{R}^n_+$.

Of course, at this point, we have derived a sufficient condition for reachability which is based on the structure of the cones $\mathrm{Cone}_\infty\big(P_{\mathcal{S}}^T\, e^{A_{j(\mathcal{S})}t} P_{\mathcal{S}}\big)$. Since Corollary 1 relates the structure of any asymptotic exponential cone to the zero patterns of its generating vectors v_j^∞, we want to further explore this technical issue, thus finally deriving a graph based condition which allows to check, for every set $\mathcal{S}$ and every index $j(\mathcal{S}) \in \mathcal{I}_{\mathcal{S}}$, when $\mathrm{Cone}_\infty\big(P_{\mathcal{S}}^T\, e^{A_{j(\mathcal{S})}t} P_{\mathcal{S}}\big) \subseteq \partial\mathbb{R}^n_+$.

Lemma 4. *[10] Given an $n \times n$ Metzler matrix A, for every $j \in \langle n \rangle$ the jth generating vector v_j^∞ of $\mathrm{Cone}_\infty(e^{At})$ is a positive eigenvector of A. As a consequence, $\mathrm{Cone}_\infty(e^{At})$ is A-invariant and therefore e^{At}-invariant $\forall\, t \geq 0$.*

Note that, by the previous result, the asymptotic exponential cone of a Metzler matrix is a polyhedral cone whose generators are all positive eigenvectors of A. If A is an irreducible matrix, it admits only one positive eigenvector of unitary norm, which is strictly positive and corresponds to the dominant eigenvalue [1]. Therefore $\mathrm{Cone}_\infty(e^{At})$ collapses into a one dimensional cone (a ray) which lies in the interior of the positive orthant. So, condition ii) in Corollary 1 cannot be fulfilled, unless A is a reducible matrix.

At this point, we want to analyze when either one of the equivalent conditions in Corollary 1 is verified. Before proceeding, we recall a result of [10], describing the dominant mode in the expression of every single entry of the exponential e^{At} of a Metzler matrix $A \in \mathbb{R}^{n\times n}$. To this end, we introduce the following notation: given an index $i \in \langle n \rangle$, we let $\mathcal{C}(i)$ be the index of the irreducibility class the vertex i belongs to (w.r.t. the directed graph $\mathcal{G}(A)$). If we assume that A is in Frobenius normal form (1) and that A_{kk} is the diagonal block corresponding to $\mathcal{C}_k$, we set $\lambda_{\max}(\mathcal{C}_k) := \lambda_{\max}(A_{kk})$. Then [10], at every time instant $t > 0$

$$e^{At} =: \mathcal{A}(t) = \begin{bmatrix} \mathcal{A}_{11}(t) & \mathcal{A}_{12}(t) & \dots & \mathcal{A}_{1\ell}(t) \\ & \mathcal{A}_{22}(t) & \dots & \mathcal{A}_{2\ell}(t) \\ & & \ddots & \vdots \\ & & & \mathcal{A}_{\ell\ell}(t) \end{bmatrix}, \tag{21}$$

where $\mathcal{A}_{ii}(t)$ is strictly positive for every i, while for $i \neq j$

$$\mathcal{A}_{ij}(t) = \begin{cases} \gg 0, & \text{if } i \in \mathcal{D}(\mathcal{C}_j) \ (\Leftrightarrow j \in \mathcal{A}(\mathcal{C}_i)); \\ 0, & \text{otherwise.} \end{cases}$$

It is worthwhile noticing that, as a consequence of the previous result, if the rth entry of some asymptotic vector v_j^∞ is zero (namely, $r \in \mathrm{ZP}(v_j^\infty)$), then (1) $\mathcal{C}(r) \subseteq \mathrm{ZP}(v_j^\infty)$; (2)for every other index $i \in \langle n \rangle$ such that $\mathcal{C}(i) = \mathcal{C}(j)$ we have $\mathcal{C}(r) \subseteq \mathrm{ZP}(v_i^\infty)$.

For any positive vector v, we let $\mathrm{block}_k[v]$ denote the kth block of v, where the block-partition of the vector corresponds to the one adopted for the matrix A. We recall the following result from [10].

Theorem 1. *Let $A \in \mathbb{R}^{n\times n}$ be a Metzler matrix in Frobenius normal form (1). For any pair of indices (i, j) in $\langle \ell \rangle \times \langle \ell \rangle$, we have:*

- *if $\mathcal{A}(\mathcal{C}_i) \cap \mathcal{D}(\mathcal{C}_j) = \emptyset$, then $\mathrm{block}_{ij}[e^{At}] = 0$;*
- *if $\mathcal{A}(\mathcal{C}_i) \cap \mathcal{D}(\mathcal{C}_j) \neq \emptyset$, then $\mathrm{block}_{ij}[e^{At}] \sim e^{\lambda^*_{i,j} t} \dfrac{t^{m_{i,j}}}{m_{i,j}!}$, where*

  $$\lambda^*_{i,j} := \max\{\lambda_{\max}(A_{kk}) : \ k \in \mathcal{A}(\mathcal{C}_i) \cap \mathcal{D}(\mathcal{C}_j)\},$$

  and $m_{i,j} + 1$ is the maximum number of classes $\mathcal{C}_k$ with $\lambda_{\max}(A_{kk}) = \lambda^*_{i,j}$ that lie in a single chain from $\mathcal{C}_j$ to $\mathcal{C}_i$ in the reduced graph $\mathcal{R}(A)$.

We finally get the desired technical result.

Proposition 8. *Let A be an $n \times n$ Metzler matrix in Frobenius normal form (1). The following facts are equivalent:*

i) $\mathrm{Cone}_\infty(e^{At}) \subseteq \partial\mathbb{R}^n_+$*;*
ii) $\exists\, k \in \langle \ell \rangle$ *such that*
  a) $\exists\, \hat{k} \in \mathcal{D}(\mathcal{C}_k) \setminus \{k\} : \ \lambda_{\max}(\mathcal{C}_{\hat{k}}) \geq \lambda_{\max}(\mathcal{C}_k)$*;*
  b) $\forall\ h \in \mathcal{A}(\mathcal{C}_k)$, $\lambda^*_h := \max_{p \in \mathcal{D}(\mathcal{C}_h)} \lambda^*_{p,h} \geq \lambda^*_{k,h}$*, where $\lambda^*_{p,h}$ and $\lambda^*_{k,h}$ are defined as in Theorem 1.*

Proof. As a preliminary step, we notice that, by Corollary 1 and the previous remarks, condition i) amounts to saying that there exists $k \in \langle \ell \rangle$ such that $\mathrm{block}_k[v_j^\infty] = 0$ for every index $j \in \langle n \rangle$.

In order to evaluate when this condition is verified, we notice that, for every index $j \in \langle n \rangle$, if $\mathcal{C}_h := \mathcal{C}(j)$ does not access $\mathcal{C}_k$, then $\mathrm{block}_k[e^{At}\mathbf{e}_j] = 0$ at every time instant t, and hence this is true also asymptotically. On the other hand, if the class $\mathcal{C}_h$ accesses $\mathcal{C}_k$, then $\text{block}_k[e^{At}\mathbf{e}_j] \neq 0$ at every time instant t. So, $\text{block}_k[v_j^\infty]$ can be zero if and only if the dominant mode of $e^{At}\mathbf{e}_j$ appears in a block different from $\text{block}_k[e^{At}\mathbf{e}_j]$.

So, assuming that $\mathcal{C}_h = \mathcal{C}(j)$ accesses $\mathcal{C}_k$, two cases may occurr:

• If $h = k$, then $\text{block}_k\left[e^{At}\mathbf{e}_j\right]$ grows as $e^{\lambda_{\max}(\mathcal{C}_k)\,t}$. So, a necessary and sufficient condition for the existence of a block of $e^{At}\mathbf{e}_j$ which exhibits a mode which strictly dominates $e^{\lambda_{\max}(\mathcal{C}_k)\,t}$ is that $\exists\,\hat{k} \in \mathcal{D}(\mathcal{C}_k), \hat{k} \neq k$, such that $\lambda_{\max}(\mathcal{C}_{\hat{k}}) \geq \lambda_{\max}(\mathcal{C}_k)$. Indeed, by resorting to Theorem 1, if $\lambda_{\max}(\mathcal{C}_{\hat{k}}) > \lambda_{\max}(\mathcal{C}_k)$, then $\text{block}_{\hat{k}}\left[e^{At}\mathbf{e}_j\right]$ grows at least as $e^{\lambda_{\max}(\mathcal{C}_{\hat{k}})\,t}$; if $\lambda_{\max}(\mathcal{C}_{\hat{k}}) = \lambda_{\max}(\mathcal{C}_k)$, then $\text{block}_{\hat{k}}\left[e^{At}\mathbf{e}_j\right]$ grows at least as $t \cdot e^{\lambda_{\max}(\mathcal{C}_k)\,t}$.

• If $h \neq k$, then, by referring to Theorem 1 statement and notation, $\text{block}_k\left[e^{At}\mathbf{e}_j\right]$ grows as $e^{\lambda^*_{k,h}t}\dfrac{t^{m_{k,h}}}{m_{k,h}!}$. So, a necessary and sufficient condition for the existence of a block of $e^{At}\mathbf{e}_j$ which exhibits a mode which strictly dominates it is that

- either there exists $p \in \mathcal{D}(\mathcal{C}_h)$ such that $\lambda^*_{p,h} > \lambda^*_{k,h}$, or
- $\lambda^*_{k,h} \leq \lambda^*_{p,h}\ \forall\, p \in \mathcal{D}(\mathcal{C}_h)$. In this case, the existence of the class $\mathcal{C}_{\hat{k}}$, accessible from $\mathcal{C}_k$ and such that $\lambda_{\max}(\mathcal{C}_{\hat{k}}) \geq \lambda_{\max}(\mathcal{C}_k)$, automatically ensures the existence of a mode in $e^{At}\mathbf{e}_j$ strictly dominating $e^{\lambda^*_{k,h}t}\dfrac{t^{m_{k,h}}}{m_{k,h}!}$.

A Technical Lemmas

Lemma A.1 *Let A be a Metzler matrix, then*

i) $e^{At} \geq 0,\ \forall\, t \geq 0$ *and if A is irreducible then* $e^{At} \gg 0$ *for every* $t > 0$*;*
ii) $\text{ZP}(e^{At}) = \text{ZP}(e^A)$ *for every* $t > 0$*;*
iii) if $\overline{\text{ZP}}(v) = \overline{\text{ZP}}(w)$*, then* $\overline{\text{ZP}}(e^{At}v) = \overline{\text{ZP}}(e^{At}w)$ *for every* $t \geq 0$*.*

Lemma A.2 [8] *Given an $n \times n$ Metzler matrix A, for every $\varepsilon > 0$ there exists $t > 0$ such that $\forall\, i,j \in \langle n\rangle$, $(1-\varepsilon)I_n \leq e^{At} \leq I_n + \varepsilon\mathbf{1}_n\mathbf{1}_n^T$, namely*

$$\begin{array}{rcll} (1-\varepsilon) \leq & [e^{At}]_{ii} & \leq (1+\varepsilon), & \\ 0 \leq & [e^{At}]_{ij} & \leq \quad \varepsilon, & \quad \textit{for} \ \ i \neq j. \end{array}$$

Lemma A.3 [9] *Let A be an $n \times n$ Metzler matrix, and let $v \in \mathbb{R}^n_+$ be a positive vector. Then*

$$\overline{\text{ZP}}(e^{A\bar{t}}v) = \mathcal{S},\ \exists\bar{t} > 0 \qquad \Rightarrow \qquad \begin{cases} \overline{\text{ZP}}(e^{At}e_{\mathcal{S}}) = \mathcal{S},\ \forall t > 0 \\ \quad \overline{\text{ZP}}(v) \subseteq \mathcal{S}. \end{cases} \tag{22}$$

Lemma A.4 *Let $v \in \mathbb{R}^n_+$ be strictly positive and set $v_{min} := \min_{i=1,2,\dots,n} v_i > 0$. Let $A \in \mathbb{R}^{n\times n}_+$ be a nonsingular square matrix, with strictly positive diagonal entries, i.e. $[A]_{ii} \gg 0\ \forall i \in \langle n\rangle$, and off-diagonal entries satisfying*

$$[A]_{ij} \leq \frac{v_{min}}{\sqrt{n}\|A^{-1}\|\,\|v\|}, \qquad \forall\, i \neq j, \tag{23}$$

where $\|\cdot\|$ *is the euclidean norm. Then* $A^{-1}v \geq 0$ *or, equivalently, the equation* $Ax = v$ *in the indeterminate* x *has a (uniquely determined) positive solution.*

Proof. Since A is nonsingular, the equation $Ax = v$ necessarily has a unique solution $x = A^{-1}v$. It only remains to show that $x \in \mathbb{R}^n_+$. Note, first, that $\|x\| = \|A^{-1}v\| \leq \|A^{-1}\| \, \|v\|$. Now, $\forall\, j \in \langle n \rangle$, $v_j = \sum_{k=1}^n [A]_{jk}\ x_k$, and hence

$$x_j = \frac{v_j - ([A]_{j1}x_1 + \ldots + [A]_{j\ j-1}x_{j-1} + [A]_{j\ j+1}x_{j+1} + \ldots + [A]_{jn}x_n)}{[A]_{jj}}.$$

Consequently, $x_j \geq 0 \Leftrightarrow (v_j - ([A]_{j1}x_1 + \ldots + [A]_{j\ j-1}x_{j-1} + [A]_{j\ j+1}x_{j+1} + \ldots + [A]_{jn}x_n)) \geq 0$. Upon setting $A_{max} := \max\{[A]_{hk} :\ \ h \neq k\}$, and by resorting to the inequality $\sum_{i=1}^n |x_i| \leq \sqrt{n}\ \|x\|$, after some computations one gets $v_j - ([A]_{j1}x_1 + \ldots + [A]_{j,j-1}x_{j-1} + [A]_{j,j+1}x_{j+1} + \ldots + [A]_{jn}x_n) \geq v_{min} - (A_{max}\sum_{i=1}^n |x_i|) \geq v_{min} - (A_{max}\sqrt{n}\|x\|) \geq v_{min} - (A_{max}\sqrt{n}\|A^{-1}\| \, \|v\|) \geq 0$, and hence the claim is proved.

References

1. A. Berman and R.J. Plemmons. (1979). *Nonnegative matrices in the mathematical sciences.* Academic Press, New York (NY).
2. R.A. Brualdi and H.J. Ryser. (1991). *Combinatorial matrix theory.* Cambridge Univ.Press, Cambridge (GB).
3. L. Farina and S. Rinaldi. (2000). *Positive linear systems: theory and applications.* Wiley-Interscience, Series on Pure and Applied Mathematics, New York.
4. S.S. Ge, Z. Sun, and T.H. Lee. (2001). *IEEE Trans. Aut. Contr.*, 46, no. 9:1437–1441.
5. D. Liberzon, J.P. Hespanha, and A.S. Morse. (1999). *Systems & Control Letters*, 37:117–122.
6. D. Liberzon and A.S. Morse. (1999). *IEEE Contr. Syst. Magazine*, 19:59–70.
7. H. Minc. (1988). *Nonnegative Matrices.* J.Wiley & Sons, New York.
8. P. Santesso and M.E. Valcher. (2006). In C. Commault, editor, *Lecture Notes in Control and Information Sciences, Eds. C.Commault, N. Marchand*, pages 185–192. Springer-Verlag.
9. P. Santesso and M.E. Valcher. (2006). In *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2006)*, File MoP01.3.pdf, Kyoto (Japan).
10. P. Santesso and M.E. Valcher. (2007). *Linear Algebra and its Applications*, 425:283–302.
11. H. Schneider. (1986). *Linear Algebra and Appl.*, 84: 161-189.
12. N.K. Son and D. Hinrichsen. (1996). *Numer. Funct. Anal. and Optimiz.*, 17, no. 5-6: 649-659.
13. Z. Sun and D. Zheng. (2001). *IEEE Trans. Aut. Contr.*, 46, no. 2:291–295.
14. Y. Wang, G. Xie, and L. Wang. (2003). In *Proc. of the 42nd IEEE Conf. on Decision and Control*, pages 5765–5770, Maui, Hawaii.
15. M.A. Wicks, P. Peleties, and R.A. De Carlo. (1998). *European J. of Control*, 4, no. 2:140–147.
16. G. Xie and L. Wang. (2003). *Systems & Control Letters*, 48:135–155.
17. G. Xie and L. Wang. (2003). *J. Math. Anal. Appl.*, 280:209–220.

Stability Tests Revisited

Umberto Viaro

Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica, Università di Udine,
via delle Scienze 208, I-33100 Udine, Italy
viaro@uniud.it

Summary. Stability–test algorithms consist essentially of a recursion that generates a sequence of polynomials of descending degree starting from a polynomial to be tested. The root–locus technique allows us to gain insight into the operation of these procedures. The paper describes with the aid of this graphic tool the stability–test procedure originally suggested by Lepschy *et al.* [9] and shows how the computations implied by this procedure can be organized in a table form.

Keywords. Linear systems, stability–test algorithms, root locus, Lepschy table.

1 Introduction

In consideration of the renewed interest in stability criteria aroused by the publication of Kharitonov's theorem on interval polynomials in the late 1970s [1], Lepschy and coworkers analyzed the structure of *recursive* stability–test algorithms for continuous– and discrete–time systems and presented a unifying frame for the two–term [2], [3], [4], [5] and three–term (or *split* [6]) forms [7], [8] of these procedures. A notable result of this analysis was the suggestion of new algorithms [9], [10] that can be used not only to check the stability of a linear system but also to construct reduced–order models that preserve important features, such as stability, of an original high–order system [11], [12] and to provide alternative model parameterizations [13].

This paper focuses on (a slight variant of) the Lepschy stability–test procedure [9] that can be considered as the s–domain counterpart of the Jury test for linear discrete–time systems [14]. The main objective of these procedures is to determine how many roots of a polynomial lie in a prescribed region of the complex plane by performing elementary algebraic operations on its coefficients. The region considered by the Jury test is the unit disk, whereas the Lepschy test refers to the left half–plane as is the case for the classical

Routh test [15], although the Routh test can be adapted to find the roots in different regions [16].

Section 2 briefly recalls the mechanism of stability–test procedures such as the classical Routh test. Section 3 illustrates the basic step–down recursion of the Lepschy test with the aid of a root locus [17]. Section 4 shows how to determine the number of open–left–half–plane (OLHP) roots of a polynomial. Section 5 provides an alternative perspective on the Lepschy test based on the root–locus invariance property [18]. The implementation of the test with the construction of the Lepschy table is illustrated in Section 6.

2 Stability–test algorithms

Essentially, a stability–test procedure consists of a recursion for generating a sequence of N polynomials of descending degree starting from an original polynomial $p_N(s)$ of degree N to be tested. At each step, the algorithm determines from the current polynomial $p_i(s)$ of degree $i \leq N$ another polynomial $p_{i-1}(s)$ of immediately lower degree $i-1$ together with the value of a characteristic parameter ρ_i. From the entire sequence of n characteristic parameter values ρ_i , $i = N, N-1, N-2, \cdots, 2, 1$, it is possible to evaluate the number of roots of the original polynomial in a given region of the complex plane.

For instance, according to the familiar Routh test, the polynomial to be tested is first decomposed into its even and odd parts. Starting from these parts, the even and odd parts of the polynomials of descending degree are generated recursively by means of a division algorithm. Specifically, denoting by $q_{i,i}(s)$ and $q_{i,i-1}(s)$ the even and odd parts of $p_i(s) = q_{i,i}(s) + q_{i,i-1}(s)$ if i is even, or vice versa if i is odd, the odd and even parts $q_{i-1,i-1}(s)$ and $q_{i-1,1-2}(s)$ of $p_{i-1}(s)$, or vice versa, are

$$q_{i-1,i-1}(s) = q_{i,i-1}(s), \tag{1}$$

$$q_{i-1,i-2}(s) = q_{i,i}(s) \; - \rho_i \, s \, q_{i,i-1}(s), \tag{2}$$

where ρ_i is the ratio between the leading coefficients of $q_{i,i}(s)$ and $q_{i-1,i-1}(s) = q_{i,i-1}(s)$. Relations (1) and (2) can be merged into the single recursion [19] (called *two–term* recursion because it involves *two* consecutive polynomials in the sequence)

$$p_{i-1}(s) = p_i(s) - \rho_i \, s \, q_{i,i-1}(s) \tag{3}$$

on which a simple proof of the Routh test is based [20]. From (1) it follows that the first subscript may be dropped. Therefore the algorithm can be presented in its usual *three–term* form involving one part of *three* consecutive polynomials as

$$q_{i-2}(s) = q_i(s) \; - \rho_i \, s \, q_{i-1}(s). \tag{4}$$

The entries of the Routh table are precisely the coefficients of $q_i(s)$, $i = N, N-1, \ldots, 0$.

The following sections concentrate on the Lepschy test [9] that can be considered as the s–domain counterpart of the Jury test because of the symmetry about the imaginary axis of the root locus in which the basic recursion of the Lepschy test can be embedded (Section 5). The computational complexity of the Lepschy test is the same as that of the Routh test and, therefore, the former represents a viable alternative to the latter at least for testing the stability of a linear system. However, the properties of the polynomials generated by the Lepschy test [12] are different from the properties of the polynomials generated by the Routh test [21].

3 Basic recursion of the Lepschy test

Let us consider a real *monic* polynomial $p_i(s)$ and decompose it into the sum of its even and odd parts as

$$p_i(s) = q_{i,i}(s) + q_{i,i-1}(s), \tag{5}$$

where $q_{i,i}(s)$ contains only the even powers of s and $q_{i,i-1}(s)$ only the odd powers of s if i is even, and vice versa if i is odd. The i roots of (5) belong to the *complete root locus* (that is, the root locus corresponding to positive *and* negative values of the (real) varying parameter [18]) for

$$q_{i,i}(s) + K\, q_{i,i-1}(s) = 0 \tag{6}$$

that can be associated with the loop function

$$L(s) := K\frac{q_{i,i-1}(s)}{q_{i,i}(s)}. \tag{7}$$

of a fictitious feedback system. Figure 1 shows the complete root locus whose departure and arrival points are, respectively, the roots of the even part and those of the odd part of $p_4(s) = s^4 + 7s^3 + 16.99s^2 + 16.95s + 5.94$. Since $p_4(s)$ is a *Hurwitz polynomial*, that is, a polynomial whose roots are all in the OLHP, the roots of its odd part $q_{4,3}(s)$ alternate with those of its even part $q_{4,4}(s)$ along the imaginary axis according to the Hermite–Biehler theorem [22].

The left–hand side of (6) is equal to zero only if both its real and imaginary parts are equal to zero. For $s = \jmath\omega$, these real and imaginary parts coincide with $q_{i,i}(\jmath\,\omega)$ and $K\, q_{i,i-1}(\jmath\,\omega)$, or vice versa. Therefore locus branches may cross the $\jmath\,\omega$-axis:
(i) at the imaginary roots of $q_{i,i}(s)$ for $K = 0$,
(ii) at the imaginary roots of $q_{i,i-1}(s)$ and at infinity for $K = \pm\infty$, and
(iii) for finite nonzero values of K when $q_{i,i}(\jmath\,\omega) = 0$ and $q_{i,i-1}(\jmath\,\omega) = 0$ simultaneously.

The last case can be ruled out if we assume $q_{i,i-1}(s)$ and $q_{i,i}(s)$ *coprime*. It follows that the number of roots with negative real part, as well as the

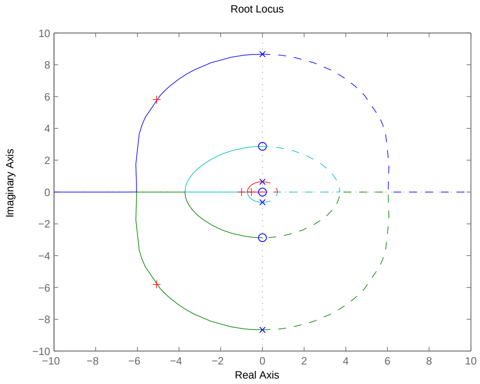


Fig. 1. Complete root locus for $q_{4,4}(s) + K\,q_{4,3}(s) = 0$, where $q_{4,4}(s)$ and $q_{4,3}$ are the even and odd parts of $p_4(s) = s^4 + 15.3s^3 + 75.5s^2 + 126.3s + 31.5 = (s+0.3)(s+3)(s+5)(s+7)$. The complete root locus is the union of the positive root locus ($K > 0$), represented by solid lines, and the negative root locus ($K < 0$), represented by dashed lines. Since $p_4(s)$ is a Hurwitz polynomial, the roots of $q_{4,3}(s) = 15.3s^3 + 126.3s = 15.3s(s+\jmath 2.8731)(s-\jmath 2.8731)$, denoted by small circles, alternate along the imaginary axis with the roots of $q_{4,4}(s) = s^4 + 75.5s^2 + 31.5 = (s+\jmath 0.6477)(s-\jmath 0.6477)(s+\jmath 8.6649)(s-\jmath 8.6649)$, denoted by ×. The symbol + denotes the roots of the Hurwitz polynomial $p_3(s) = s^3 + 10.6696s^2 + 64.8309s + 31.5013 = (s+0.5298)(s+5.0699+\jmath 5.8099)(s+5.0699-\jmath 5.8099)$ obtained from $p_4(s)$ according to (10) with $P = -1$; the point P too is represented by a +. From (11) the value K_{-1} taken by the varying parameter K in (6) at the point $P = -1$ turns out to be $K_{-1} = \rho_4 = -q_{4,4}(-1)/q_{4,3}(-1) = 0.7627$.

number of roots with positive real part, may change only when the sign of K changes.

Any point P of the real axis belongs to the complete root locus (6) for the (real) value

$$K_P := -\frac{q_{i,i}(P)}{q_{i,i-1}(P)} \tag{8}$$

of the varying parameter K because the coefficients of $q_{i,i}(s)$ and $q_{i,i-1}(s)$ are real. In other words, it is always possible to find a linear combination

$$\ell_{P_i}(s) := q_{i,i}(s) + K_P\,q_{i,i-1}(s), \tag{9}$$

with K_P real, that admits $s - P$, with P real, as a factor.

The step–down recursion of the Lepschy test determines $p_{i-1}(s)$ by choosing a real point P and factoring out $s - P$ from $\ell_{P_i}(s)$ according to

$$q_{i,i}(s) + K_P\,q_{i,i-1}(s) = (s - P)\,p_{i-1}(s). \tag{10}$$

From the computational point of view, the most convenient choice is $P = -1$. The roots of the polynomial $p_3(s) = s^3 + 10.6696s^2 + 64.8309s + 31.5013 =$

$(s+0.5298)(s+5.0699+\jmath 5.8099)(s+5.0699-\jmath 5.8099)$ obtained from $p_4(s) = s^4+15.3s^3+75.5s^2+126.3s+31.5 = (s+0.3)(s+3)(s+5)(s+7)$ according to (10) with $P=-1$ are denoted, like the point P itself, by the symbol + in Figure 1.

4 Polynomial root distribution

When the sign of K_P in (8) is equal to the sign of the leading coefficient c of $q_{i,i-1}(s)$, the numbers of OLHP roots of $p_i(s)$ coincides with the number of OLHP roots of $\ell_{P_i}(s)$ in (9) because in this case *both* the roots of $p_i(s)$ *and* the roots of $\ell_{P_i}(s)$ belong to the positive root locus of (6) for $\mathrm{sgn}(K_P) = \mathrm{sgn}(c) = 1$ and to the negative root locus of (6) for $\mathrm{sgn}(K_P) = \mathrm{sgn}(c) = -1$. Instead, when $\mathrm{sgn}(K_P) = -\,\mathrm{sgn}(c)$, the number of OLHP roots of $p_i(s)$ coincides with the number of OLHP roots of $q_i(s) - K_P\, q_{i,i-1}(s)$ and, thus, with the number of OLHP roots of $\ell_{P_i}(-s)$.

If the point P lies on the negative real semi–axis, $\ell_{P_i}(s)$ has the same number of roots in the closed right half–plane as $p_{i-1}(s)$ but one more OLHP root, that is, P itself. For $P=-1$ the parameter K_P in (8) takes the value

$$K_{-1} = \rho_i := -\frac{q_{i,i}(-1)}{q_{i,i-1}(-1)} \qquad (11)$$

which is the ratio between the sum of the even–order coefficients of $p_i(s)$ and the sum of its odd–order coefficients for i even, and vice versa for i odd. If the monic polynomial $p_i(s)$ is Hurwitz, all of the coefficients of $p_i(s)$ are positive so that $\mathrm{sgn}(\rho_i) = \mathrm{sgn}(c) = 1$ and $p_{i-1}(s)$ is Hurwitz too. For instance, the polynomial $p_3(s)$, whose roots are denoted by small squares in Figure 1, is Hurwitz such as $p_4(s)$, from which $p_3(s)$ is obtained according to (10) with $P=-1$.

From the sequence of the N values $\rho_i\,, i=1,2,\cdots,N$, computed from the coefficients of the polynomials $p_i(s)$ successively generated according to (10) with $P=-1$, it is possible to determine how many roots of the original monic polynomial $p_N(s)$ lie in the OLHP. Precisely, the number of OLHP roots of $p_N(s)$ corresponds to the number of positive entries in the sequence [9]

$$\{\, \sigma_i := \prod_{j=i}^{N} \mathrm{sgn}(\rho_j)\ ,\ i=1,2,\cdots,N\,\}, \qquad (12)$$

similar to the sequence used to determine how many roots lie inside the unit circle in the discrete–time case [14].

In particular, $p_N(s)$ is Hurwitz if and only if

$$\sigma_i > 0\,, i=1,2,\cdots,N, \qquad (13)$$

or, equivalently, if and only if

$$\rho_i > 0\,, i = 1, 2, \cdots, N. \tag{14}$$

Two examples are worked out in Section 6. The critical cases arising when either $\rho_i = 0$ or $\rho_i = \infty$ can be dealt with easily as in [9].

5 An alternative perspective

According to the root–locus–invariance property [18], the *complete* root locus for the loop function (7) does not change if the numerator and denominator polynomials of $L(s)$ are replaced by two linear combinations of $q_{i,i}(s)$ and $q_{i,i-1}(s)$. Therefore the step–down recursion of the Lepschy algorithm can be analyzed by referring to the complete root locus departing from the roots of $p_i(s) = q_{i,i}(s) + q_{i,i-1}(s)$ and arriving at the roots of

$$p_i(-s) = (-1)^i q_{i,i}(s) + (-1)^{i-1} q_{i,i-1}(s) \tag{15}$$

since this locus coincides with the complete root locus constructed from the roots of $q_{i,i}(s)$ and $q_{i,i-1}(s)$. This new choice of departure and arrival points helps us understand why the complete root locus is symmetric about the imaginary axis. The equation of the complete root locus (6) can therefore be rewritten as

$$p_i(s) + K'\, p_i(-s) = 0, \tag{16}$$

where K' is the new varying parameter, which is equal to zero at the roots of $p_i(s)$ and to $\pm\infty$ at the roots of $p_i(-s)$. Correspondingly, the basic step–down recursion of the Lepschy algorithm (10) becomes

$$p_i(s) + K'_P\, p_i(-s) = (1 + K'_P)\,(s - P)p_{i-1}(s), \tag{17}$$

where the factor $1 + K'_P$ at the left–hand side is to match the leading coefficients of both sides of (17) and, taking (10), (15), (16) and (17) into account, the value K'_P taken by the new current parameter K' at the point P is related to K_P via

$$K'_P = -\frac{p_i(P)}{p_i(-P)} = -(-1)^i \frac{1 - K_P}{1 + K_P}. \tag{18}$$

If $p_i(s)$ is a Hurwitz polynomial, as K' increases from 0 to ∞, all i branches of the positive locus for (16) leave the OLHP roots of $p_i(s)$, cross the imaginary axis and arrive at the i roots of $p_i(-s)$ that are the negatives of the roots of $p_i(s)$. Conversely, as K' increases from $-\infty$ to 0, all i branches of the negative locus return to the OLHP roots of $p_i(s)$ after crossing the imaginary axis. Due to the symmetry of the departure and arrival points, *the complete root locus is symmetric about the imaginary axis*, where $|K'| = 1$. It follows that the polynomial at the left–hand side of (16) is Hurwitz for $-1 < K' < 1$, because for these values of K' the roots of $p_i(s) + K'\, p_i(-s)$ are located in the same half–plane as the roots of $p_i(s)$. If the point P belongs to the left real semi–axis, the absolute value of K'_P in (18) is less than 1 and $p_{i-1}(s)$ in (17) is Hurwitz such as $p_i(s)$. Applying this argument to every polynomial successively generated from a given Hurwitz polynomial $p_N(s)$ of degree N according to recursion (17), we conclude that all of the N values of K'_P computed in the N steps of the procedure are included in the open interval $(-1, 1)$.

Figure 2 shows the complete root locus for $p_4(s) + K'\,p_4(-s) = 0$, where $p_4(s)$ is the Hurwitz polynomial $p_4(s) = s^4 + 15.3s^3 + 75.5s^2 + 126.3s + 31.5 = (s + 0.3)(s + 3)(s + 5)(s + 7)$ already considered in Figure 1. All of the four locus branches develop in the OLHP for $|K'| < 1$, because these branches pass through the OLHP roots of $p_4(s)$ for $K' = 0$ and may cross the imaginary axis only for $|K'| = 1$. At $P = -1$ the varying parameter K' takes the value $K'_P = 0.1346$. The four roots corresponding to $K'_P = 0.1346$, including P itself, are denoted by the symbol $+$ in Figure 2. The next polynomial $p_3(s)$ in the sequence generated according to (17) with $P = -1$ is obtained by factoring $s + 1$ out of $p_4(s) + 0.1346\,p_4(-s) = (s + 1)p_3(s)$; specifically, $p_3(s) = s^3 + 10.6696s^2 + 64.8309s + 31.5013 = (s + 0.5298)(s + 5.0699 + \jmath 5.8099)(s + 5.0699 - \jmath 5.8099)$, as in the example considered at the end of Section 3.

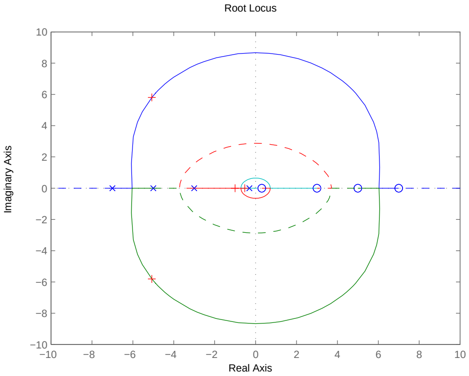


Fig. 2. Complete root locus associated with the basic recursion of the Lepschy test constructed from the roots of $p_i(s)$, denoted by small $\times$, and $p_i(-s)$, denoted by small circles, instead of the roots of $q_{i,i}(s)$ and $q_{i,i-1}(s)$. In particular, this figure depicts the complete root locus of $p_4(s) + K'\,p_4(-s) = 0$, where $p_4(s) = s^4 + 15.3s^3 + 75.5s^2 + 126.3s + 31.5 = (s + 0.3)(s + 3)(s + 5)(s + 7)$ is the polynomial already considered in Figure 1.The arrival points of this locus are the negatives of its departure points. The imaginary axis may be crossed only for $|K'| = 1$. According to (17) (and to (10) as well) with $P = -1$, the next polynomial $p_3(s)$ in the sequence is $p_3(s) = s^3 + 10.6696s^2 + 64.8309s + 31.5013 = (s + 0.5298)(s + 5.0699 + \jmath 5.8099)(s + 5.0699 - \jmath 5.8099)$, whose roots are denoted, like P, by the symbol $+$. From (18) the value K'_P taken by the varying parameter K' at the roots of $p_3(s)$ turns out to be $K'_P = -p_4(-1)/p_4(1) = 0.1346$.

The Lepschy test can be considered as the s–domain counterpart of the Jury test, strictly related to the Levinson algorithm [23], whose basic recursion can be embedded in a locus that is *symmetric with respect to the unit cirle* [24], that is, if a point z belongs to this locus, the point $1/z$ belongs to it too.

6 Implementation of the Lepschy test

Setting $P = -1$ in (10) and recalling the definition (11) of ρ_i, the step–down recursion of the Lepschy test becomes

$$(s+1)\,p_{i-1}(s) = q_{i,i}(s) + \rho_i\,q_{i,i-1}(s). \tag{19}$$

Table 1. Table form of the Lepschy test for a polynomial $p_N(s)$ of even degree N

	s^0	s^1	s^2	s^3	s^4	$\cdot$	ρ_i
$p_N(s)$	$a_{N,0}$	$a_{N,1}$	$a_{N,2}$	$a_{N,3}$	$a_{N,4}$	$\cdot$	ρ_N
	$\downarrow 1$	$\downarrow \rho_N$	$\downarrow 1$	$\downarrow \rho_N$			
$p_{N-1}(s)$	$a_{N-1,0}$	$\overset{-1}{\rightarrow} a_{N-1,1}$	$\overset{-1}{\rightarrow} a_{N-1,2}$	$\overset{-1}{\rightarrow} a_{N-1,3}$	$\cdot$		ρ_{N-1}
	$\downarrow \rho_{N-1}$	$\downarrow 1$	$\downarrow \rho_{N-1}$				
$p_{N-2}(s)$	$a_{N-2,0}$	$\overset{-1}{\rightarrow} a_{N-2,1}$	$\overset{-1}{\rightarrow} a_{N-2,2}$	$\cdot$			ρ_{N-2}
	$\downarrow 1$	$\downarrow \rho_{N-2}$	$\downarrow 1$				
	$\cdot$	$\cdot$	$\cdot$				$\cdot$
	$\downarrow \rho_3$	$\downarrow 1$	$\downarrow \rho_3$				
$p_2(s)$	$a_{2,0}$	$\overset{-1}{\rightarrow} a_{2,1}$	$\overset{-1}{\rightarrow} a_{2,2}$				ρ_2
	$\downarrow 1$	$\downarrow \rho_2$					
$p_1(s)$	$a_{1,0}$	$\overset{-1}{\rightarrow} a_{1,1}$					ρ_1
	$\downarrow \rho_1$						
$p_0(s)$	$a_{0,0}$						

The computation of the coefficients of the polynomials

$$p_i(s) = a_{i,i}s^i + a_{i,i-1}s^{i-1} + a_{i,i-2}s^{i-2} + \cdots + a_{i,1}s + a_{i,0} \tag{20}$$

successively generated by means of (19) can be organized as in Table 1 that refers to a polynomial $p_N(s)$ of even degree N.

Each row of Table 1 is formed from the coefficients of a polynomial. Parameter ρ_i, necessary to compute the coefficients of the row for $p_{i-1}(s)$, is the ratio between the sum of the even–order coefficients of $p_i(s)$ and the sum of its odd–order coefficients if i is even, or vice versa if i is odd, that is,

$$\rho_i = \frac{a_{i,i} + a_{i,i-2} + \cdots}{a_{i,i-1} + a_{i,i-3} + \cdots} \,. \tag{21}$$

Except for $a_{i,0}$, which depends only on the coefficient $a_{i+1,0}$ placed above it, every coefficient $a_{i,j}$ of the row for $p_i(s)$ depends both on the coefficient $a_{i,j-1}$ at its left and on the coefficient $a_{i+1,j}$ placed above $a_{i,j}$. Precisely, coefficient $a_{i,j}$, $j > 0$, is obtained by subtracting $a_{i,j-1}$ from $a_{i+1,j}$ if $i + j$ is odd and by subtracting $a_{i,j-1}$ from $\rho_{i+1}a_{i+1,j}$ if $i + j$ is even.

Even if the Routh table for the same polynomial contains a smaller number of entries, the Routh and Lepschy tests have the same computational complexity since the computation of the entries of the Routh table requires a larger number of operations [9].

Table 2. Lepschy table for the Hurwitz polynomial $p_4(s) = s^4 + 3s^3 + 5s^2 + 3s + 1$

	s^0	s^1	s^2	s^3	s^4	ρ_i
$p_4(s)$	1	3	5	3	1	$\rho_4 = \frac{7}{6}$
	$\downarrow 1$	$\downarrow \rho_4$	$\downarrow 1$	$\downarrow \rho_4$		
$p_3(s)$	1	$\overset{-1}{\rightarrow} \frac{5}{2}$	$\overset{-1}{\rightarrow} \frac{5}{2}$	$\overset{-1}{\rightarrow} 1$		$\rho_3 = 1$
	$\downarrow \rho_3$	$\downarrow 1$	$\downarrow \rho_3$			
$p_2(s)$	1	$\overset{-1}{\rightarrow} \frac{3}{2}$	$\overset{-1}{\rightarrow} 1$			$\rho_2 = \frac{4}{3}$
	$\downarrow 1$	$\downarrow \rho_2$				
$p_1(s)$	1	$\overset{-1}{\rightarrow} 1$				$\rho_1 = 1$
	$\downarrow \rho_1$					
$p_0(s)$	1					

Table 2 shows the array for the Hurwitz polynomial $p_4(s) = s^4 + 3s^3 + 5s^2 + 3s + 1$. According to condition (14), $\rho_i > 0$, $i = 4, 3, 2, 1$.

Table 3 refers instead to the non–Hurwitz polynomial $p_3(s) = s^3 + 2s^2 - 2s + 1$. Since, in this case, $\rho_i < 0$, $i = 3, 2, 1$, the elements of the sequence (12) for $p_3(s)$ are $\sigma_1 = -1$, $\sigma_2 = 1$ and $\sigma_3 = -1$, so that only one root of $p_3(s)$ is in the OLHP.

Table 3. Lepschy table for $p_3(s) = s^3 + 2s^2 - 2s + 1$

	s^0	s^1	s^2	s^3	ρ_i
$p_3(s)$	1	-2	2	1	$\rho_3 = -\frac{1}{3}$
	$\downarrow \rho_3$	$\downarrow 1$	$\downarrow \rho_3$		
$p_2(s)$	$-\frac{1}{3}$	$\overset{-1}{\rightarrow} -\frac{5}{3}$	$\overset{-1}{\rightarrow} 1$		$\rho_2 = -\frac{2}{5}$
	$\downarrow 1$	$\downarrow \rho_2$			
$p_1(s)$	$-\frac{1}{3}$	$\overset{-1}{\rightarrow} 1$			$\rho_1 = -3$
	$\downarrow \rho_1$				
$p_0(s)$	1				

7 Conclusions

The operation of the Lepschy stability–test algorithm has been described by means of the root–locus technique that shows how the roots of every polynomial recursively generated by this algorithm are related to the roots of the polynomial obtained in the preceding step. The computations for determining the resulting sequence of polynomials can be organized in the Lepschy table. The number of computations needed to form this table is practically the same as the number of operations required by the Routh test.

Toni managed the methods and tools of this note in an extremely brilliant manner. This author misses Toni's friendship and advice more than His unparalleled skill.

References

1. Kharitonov VL (1978) *Differenzial'nye Uravneniya*, 14:2086–2088
2. Krajewski W, Lepschy A, Mian GA, Viaro U (1990) *IEEE Trans Circuits Syst*, 37:290–296
3. Lepschy A, Mian GA, Viaro U (1990) *Int J Systems Sci*, 21:739–747
4. Lepschy A, Viaro U (1993) *Int. J. Control*, 58(2):485–493
5. Lepschy A, Viaro U (1994) *Circuits Systems Signal Processing*, 13(5):615–623
6. Delsarte P, Genin Y (1987) *IEEE Trans Acoust Speech Signal Processing*, 35:645–653
7. Lepschy A, Mian GA, Viaro U (1989) *Int J Control*, 50:2237–2247
8. Lepschy A, Mian GA, Viaro U (1991) *J Franklin Inst*, 328:103–121
9. Lepschy A, Mian GA, Viaro U (1988) *Systems Control Lett*, 10:175–179
10. Lepschy A, Mian GA, Viaro U (1989) *Int J Systems Sci*, 20:945–956
11. Lepschy A, Mian GA, Viaro U (1988) *IEEE Trans Automat Contr*, 33:307–310
12. Krajewski W, Lepschy A, Viaro U (1993) *Arch Contr Sci*, 2(XXXVII):73–83
13. Lepschy A, Mian GA, Viaro U (1988) *IEEE Trans Acoust Speech Signal Processing*, 36:1355–1357
14. Jury EI, Blanchard J (1961) *IRE Proc*, 49:1947–1948
15. Routh EJ (1877) *A treatise on the stability of a given state of motion.* Macmillan, London
16. Lepschy A, Policastro M, Raimondi T (1968) *Calcolo*, 5:525–536
17. Evans WR (1948) *Trans AIEE*, 67:547–551
18. Krajewski W, Viaro U (2007) *IEEE Control Systems Magazine*, 27:36–43
19. Lepschy A, Viaro U (1983) *Atti Mem Acc Patavina*, SS LL AA XCIV:5–21
20. Ferrante A, Lepschy A, Viaro U (1999) *IEEE Trans Automat Contr*, 44:1306–1309
21. Beghi A, Lepschy A, Viaro U (1994) *IEEE Trans Automat Contr*, 39:2494–2496
22. Bhattacharyya SP, Chapellat H, Keel LH (1995) *Robust control: the parametric approach.* Prentice Hall, Upper Saddle River (NJ)
23. Bistritz Y, Lev–Ari H, Kailath T (1989) *IEEE Trans Information Theory*, 35:675–682
24. Lepschy A, Mian GA, Viaro U (1990) *Contr Computers*, 18:70–73

An introduction to quantized control

Fabio Fagnani[1] and Sandro Zampieri[2]

[1] Dipartimento di Matematica, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129 Torino, Italy
fabio.fagnani@polito.it

[2] Department of Information Engineering, Università di Padova, Via Gradenigo 6/a, 35131 Padova, Italy
zampi@dei.unipd.it

Summary. The problem of stabilizing a discrete time linear system by a quantized feedback with memory is solved in the literature using various approaches, all yielding very similar results in term of the rata rate required to obtain the stabilization. In fact these techniques are all strictly connected. In this contribution we propose a unified framework which allows to compare various techniques. This framework suggests also an interesting connection between the synthesis of the quantized feedback problem and the design of the tree-structured vector quantizers.

1 Introduction

The literature devoted to the quantized feedback stabilization problem and to stabilization under communication constraints can be divided in two categories. Some literature restricts to memoryless quantized feedback, whose analysis proved to be quite hard [2, 4, 5, 3]. The remaining literature allows instead a memory structure in the control. More precisely, in this case the control consists in two parts, the first which quantizes the state and encodes the information in a suitable way, the second which decodes the information and produces the control action. In fact, even though the results proposed are quite similar, the controller structures proposed in the various papers differ quite apparently. In this contribution we propose a unified framework by which we can compare the various methods proposed in the literature. We use this framework for deriving in an easy way some convergence results already presented in the literature [8, 12]. Moreover we show an interesting connection between quantized control synthesis and the design of tree-structured vector quantizers [6]. All these facts seem to suggest that the proposed framework is the most natural in the analysis of quantized control strategies.

Consider a linear discrete time system

$$x_{t+1} = Ax_t + Bu_t \tag{1}$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}$ and where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. We want to control this system by a feedback strategy. However we assume that the sensor, by which the state x_t can be determined, is remotely located and that a digital noiseless communication channel can be used to transmit data from the sensor to the controller. The noiseless digital channel is modeled by the identity map on a finite alphabet Ω. The overall scheme is illustrated in Figure 1.

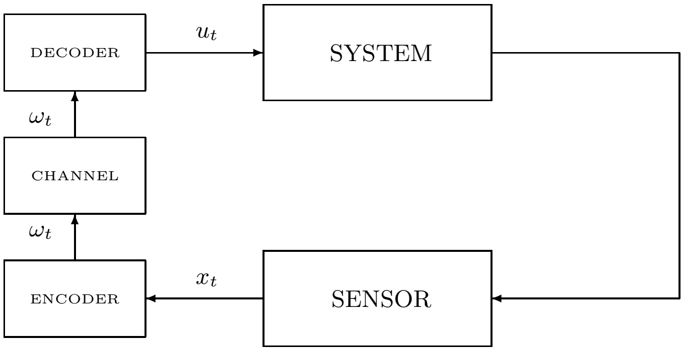


Fig. 1. Scheme representing the control under communication constraints

Notice that $\omega_t \in \Omega$ is the the signal to be transmitted and so, from the communication point of view, the important parameter is the cardinality of the alphabet Ω which is denoted by L.

In the literature it is possible to distinguish two ways for introducing quantized controllers with memory, one based on the concept of hybrid automaton [1, 12], another one based on the concept of tree structured quantizers [8].

A Hybrid automaton quantized controller

A quantized controller in this set up is given by an encoder

$$\begin{cases} s'_{t+1} = f'(s'_t, x_t) \\ \omega_t = h'(s'_t, x_t)\ , \end{cases} \tag{2}$$

where $s'_t \in S$, $\omega_t \in \Omega$, $f' : S \times \mathbb{R}^n \to S$, $h' : S \times \mathbb{R}^n \to \Omega$ and where the sets S and Ω are finite or denumerable. The decoder coincides with the system

$$\begin{cases} s''_{t+1} = f''(s''_t, \omega_t) \\ u_t = h''(s''_t, \omega_t)\ , \end{cases} \tag{3}$$

where $s''_t \in S$, $\omega_t \in \Omega$, $f'' : S \times \mathbb{R}^n \to S$, $h'' : S \times \mathbb{R}^n \to \Omega$. Imposing the compatibility condition

$$\begin{aligned} f'(s,x) &= f''(s,h'(s,x)) \\ h'(s,x) &= h''(s,h'(s,x)) \end{aligned} \qquad \forall x \in \mathbb{R}^n,\ s \in S \tag{4}$$

and the synchronized initialization $s'_0 = s''_0$ we have that $s'_t = s''_t$ for all $t \geq 0$ and so the couple encoder/decoder can be represented by the quantized controller with memory

$$\begin{cases} s_{t+1} = f(s_t, x_t) \\ u_t = k(s_t, x_t) \\ s_0 = s^* \ , \end{cases} \tag{5}$$

where $f(s,x) = f'(s,x)$ and $k(s,x) = h''(s, h'(x,s))$. The cardinality of S is a parameter indicating the computational complexity of the control scheme. This approach covers the quantized control based on the zooming in/zooming out strategy proposed in [1] adapted to control under communication constraints in [4].

B Tree structured quantized controller

Another class of control under communication constraints methods is based on the encoder

$$\omega_t = k'_t(x_t; \omega_0, \omega_1, \ldots, \omega_{t-1}) \tag{6}$$

where $\omega_t \in \Omega$, $k' : \mathbb{R}^n \times \Omega^t \to \Omega$ and where the set Ω is finite. The decoder is given by

$$u_t = k''_t(\omega_t; \omega_0, \omega_1, \ldots, \omega_{t-1}) \tag{7}$$

where $k'' : \Omega^{t+1} \to \mathbb{R}$. Imposing that the map $k''_t(\cdot; \omega_0, \omega_1, \ldots, \omega_{t-1}) : \Omega \to \mathbb{R}$ is invertible, then the couple encoder/decoder can be represented by the quantized controller with memory

$$u_t = k_t(x_t; u_0, u_1, \ldots, u_{t-1}) \tag{8}$$

where $k : \mathbb{R}^n \times \mathbb{R}^t \to \mathbb{R}$. Notice that also in this case ω_t is the the signal to be transmitted and the cardinality L of Ω is the relevant parameter from the communication point of view. This approach covers the control under communication constraints as proposed in [8].

C The relation between these quantized control schemes

In fact, the two schemes modeling the control under communication constraints are essentially equivalent even though they highlight different aspects of the problem.

Starting from an encoder/decoder pair (2,3), the synchronization $s'_t = s''_t$ for all $t \geq 0$ implies that ω_t can be obtained from $x_t, \omega_{t-1}, \omega_{t-2}, \ldots, \omega_{t-k}, s''_{t-k}$ for all k and so ω_t is a function of $x_t, \omega_{t-1}, \omega_{t-2}, \ldots, \omega_0$. In the same way it can be shown that u_t can be obtained from $\omega_t, \omega_{t-1}, \omega_{t-2}, \ldots, \omega_{t-k}, s''_{t-k}$ for all k and so u_t is a function of $\omega_t, \omega_{t-1}, \omega_{t-2}, \ldots, \omega_0$.

Conversely, is we consider an encoder/decoder pair (6,7), in order to pass to an an encoder/decoder pair like (2,3) we need to define $S := \Omega^*$ namely the set of finite strings on the alphabet Ω. In this way it is clear how to construct the maps f', f'', h', h'' from k' and k''. Notice that in this case we need to impose that S is denumerable.

Despite the equivalence between these two approaches, these seems to be convenient for solving different problems related to quantized control. The first approach (5) is more convenient when it is necessary to restrict the controller complexity [4, 1]. The second approach instead seems to be more flexible when determining the best way to use the allowed rate to achieve stabilization [8, 12] or optimal control [7].

D Estimation and control: a separation principle

All the control quantized control strategies proposed in the literature have a supplementary structure, namely they are based on the preliminary estimation of the state of the system. Namely the control in the decoder is computed from the estimate $\hat{x}_t$ of the state

$$u_t = K_t \hat{x}_t$$

Notice that $\hat{x}_t$ will depend on x_0 as well on the sequence of controls $u_0, \dots, u_{t-1}$ so on the feedback matrices K_t. We say that we have a separation principle, if the quantized state estimator is constructed in such a way that $e_t := x_t - \hat{x}_t$ does not depend on the feedback matrices K_t. In this case the closed loop system takes the form

$$x_{t+1} = Ax_t + BK_t\hat{x}_t = (A + BK_t)x_t - BK_t e_t$$

The effects of the control strategy and of the estimation process in the closed loop system is now very easily analyzed. For instance, if e_t converges to 0 exponentially, and K_t is an exponentially stabilizing feedback, than x_t converges to 0 exponentially.

Quantized controller based on the separation principle can be obtained using both in the hybrid automaton strategy and in the tree structured quantization strategy.

In the hybrid automaton strategy the encoder will have the following form

$$\begin{cases} s'_{t+1} = f'(s'_t, x_t, u_t) \\ \omega_t = h'(s'_t, x_t, u_t) \end{cases} \tag{9}$$

while the decoder/estimator will have the following form

$$\begin{cases} s''_{t+1} = f''(s''_t, \omega_t, u_t) \\ \hat{x}_t = g''(s''_t, \omega_t, u_t) \end{cases} \tag{10}$$

Notice that, differently for above, the two systems need the knowledge of u_t. This can be obtained by the knowledge of of the matrices K_t and from the fact that $u_t = K_t\hat{x}_t$, since $\hat{x}_t$ is known on both sides.

In the tree structured quantization strategy instead the encoder has the following form

$$\omega_t = k'_t(x_t; \omega_0, \omega_1, \ldots, \omega_{t-1}, u_0, u_1, \ldots, u_{t-1}) \tag{11}$$

while the decoder/estimator will have the following form

$$\hat{x}_t = k''_t(\omega_t; \omega_0, \omega_1, \ldots, \omega_{t-1}, u_0, u_1, \ldots, u_{t-1}) \tag{12}$$

Also in this case the observation we made above about the knowledge holds true. The overall scheme is illustrated in Figure 2.

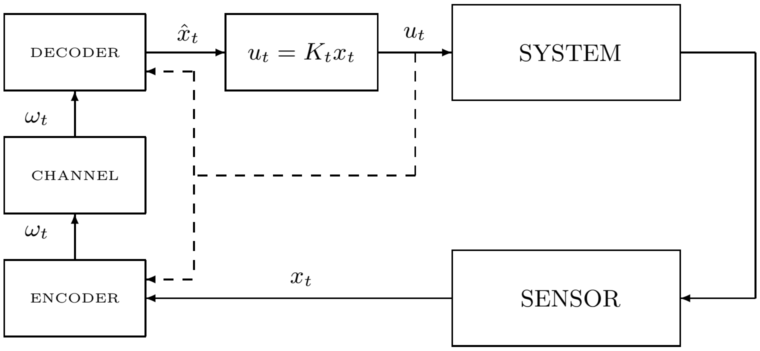


Fig. 2. Scheme representing the control under communication constraints with the estimator

We will show in the sequel two examples of quantized controller satysfying the separation principle.

2 Tree-structured quantized controllers satisfying the separation principle

In this section we will describe a quantized controller with memory in the form of a tree-structures quantizer which can be expressed as the connection of a quantized estimator and a standard state feedback. This approach has been proposed first by Nair and Evans in [8].

Consider again an encoder/decoder-controller of the form (6,7). With it we can associate in a natural way a family of partitions of the space $\mathbb{R}^n$ which contain all the relevant information on the control scheme. Indeed, it is clear that by connecting the system (1) with the quantized controller (6,7) we obtain a closed loop systems in which the state evolution x_t is completely determined by the initial condition x_0. This implies that from the closed loop system we can obtain a family of maps $\phi_t : \mathbb{R}^n \to \Omega^t$ such that

$$(\omega_0, \omega_1, \dots, \omega_{t-1}) = \phi_t(x_0)\,.$$

Clearly, the maps ϕ_t are quantized and if we denote by $\mathcal{I}_t$ the corresponding partition of $\mathbb{R}^n$ into quantization regions, we clearly have that the quantization regions of $\mathcal{I}_{t+1}$ are obtained by partitioning the quantization regions of $\mathcal{I}_t$, equivalently, each quantization region of $\mathcal{I}_t$ is the union of quantization regions of $\mathcal{I}_{t+1}$. A family of partitions of this type is said to be a tree-structured. The number of quantization subsets ν_t in $\mathcal{I}_t$ is called the complexity sequence of the family $\mathcal{I}_t$. Notice that we always have that $\nu_t \leq L^t$ for every t.

Given a partition $\mathcal{I}$ of $\mathbb{R}^n$, a quantized map $q : \mathbb{R}^n \to A$ is said to be adapted to $\mathcal{I}$ if q is constant on the sets of $\mathcal{I}$. A family of vector quantizers $q_t : \mathbb{R}^n \to \mathbb{R}^n$ which are adapted to a tree-structured sequence of partitions, is called tree-structured vector quantizers. Tree-structured vector quantizers are quite classical in the theory of vector quantization and source coding (see [6] and they are also related with some wavelet based quantization methods [10]. There are not many results proposed in the literature on the asymptotic performances of this class of quantizers (see [9]).

We now show how we can construct quantized control schemes satisfying the separation principle using these families of quantizers. Suppose we have tree-structured quantizers q_t. Once we choose the family of controls $u_0, \dots, u_{t-1}$, we have that

$$x_t = A^t x_0 + A^{t-1} B u_0 + \cdots + B u_{t-1}$$

From the quantizers q_t we can construct an estimator of x_t by simply defining

$$\hat{x}_t = A^t \hat{x}_{0|t} + A^{t-1} B u_0 + \cdots + B u_{t-1}$$

where

$$\hat{x}_{0|t} := q_t(A^{-t}(x_t - A^{t-1} B u_0 - \cdots - B u_{t-1})) = q_t(x_0)$$

The estimation error $e_t = x_t - \hat{x}_t$, satisfies

$$e_t = A^t(q_t(x_0) - x_0))\,, \qquad t \in \mathbb{N}$$

and it does not depend on the particular control strategy. If the tree-structured quantizers are chosen in such a way that

$$e_t = A^t(q_t(x_0) - x_0)$$

tends to zero as $t \to +\infty$, and if the controls are chosen according to an exponentially stabilizing feedback law $u_t = K_t \hat{x}_t$, we will obtain that closed loop system is exponentially stable. The only point that remains to be proven is that this control strategy can be represented as in (6,7). This can be seen as follows. Since the quantizers are tree-structured, we can represent a quantization region I of q_t by a length t string on the alphabet $\Omega := \{1, \dots, L\}$ in such a way that any quantization subregion I' of I will be represented by a length $t+1$ string obtained by adding a letter to the string associated with I. More formally, there exists two maps

$$\begin{aligned}\psi_t &: \mathbb{R}^n \to \Omega \\ \phi_t &: \Omega^t \to \mathbb{R}^n\end{aligned}$$

such that

$$q_t(x_0) = \phi_t(\psi_0(x_0), \ldots, \psi_t(x_0)) \tag{13}$$

Now we define

$$\begin{aligned}\omega_t &= \psi_t(A^{-t}(x_t - A^{t-1}Bu_0 - \cdots - Bu_{t-1})) \\ \hat{x}_{0|t} &= \phi_t(\omega_0, \ldots, \omega_t) \\ \hat{x}_t &= A^t \hat{x}_{0|t} + A^{t-1}Bu_0 + \cdots + Bu_{t-1}\end{aligned} \tag{14}$$

Notice that the encoder consists in the first equation, while the decoder/estimator consists of the last two equations.

3 Hybrid automaton quantized controllers satisfying the separation principle

In this section we will describe a quantized controller with memory in the form of hybrid automaton which can be expressed as the connection of a quantized estimator and a standard state feedback. This approach has been proposed first by Tatikonda in [11, 12].

Let $q_L : [-1, 1] \to \mathbb{R}$ be the uniform quantizer defined as follows. When L is odd, then

$$q_L(x) := 2k/L \text{ if } (2k-1)/L \leq x \leq (2k+1)/L$$

while when L is even, then

$$q_L(x) := (2k+1)/L \text{ if } 2k/L \leq x \leq (2k+2)/L$$

It is easy to see that that q_L has L quantization intervals and that $|x - q(x)| \leq 1/L$ for all x such that $|x| \leq 1$.

Let $Q := \{x \in \mathbb{R}^n | \ ||x||_\infty \leq 1\}$. Fix a n-tuple of integers $L_1, \ldots, L_n$ and define the quantizer

$$q : Q \to \mathbb{R}^n : (x_1, \ldots, x_n) \mapsto (q_{L_1}(x_1), \ldots, q_{L_n}(x_n))$$

The state estimator works as follows. Its state is given by the estimates $\hat{x}_t$ and the scaling matrix S_t. It is convenient to split it into two sections. The first section is

$$\begin{cases}\hat{x}_{t+1|t} = A\hat{x}_{t|t} + Bu_t \\ S_{t+1|t} = AS_{t|t}\end{cases} \tag{15}$$

while the second section is

$$\begin{cases} \hat{x}_{t+1|t+1} = S_{t+1|t}q(S_{t+1|t}^{-1}(x_{t+1} - \hat{x}_{t+1|t})) + \hat{x}_{t+1|t} \\ S_{t+1|t+1} = S_{t+1|t}D^{-1} \end{cases} \tag{16}$$

where $D := \text{diag}\{L_1, \dots, L_n\}$, if $S_{t+1|t}^{-1}(x_{t+1} - \hat{x}_{t+1|t}) \in Q$, and

$$\begin{cases} \hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} \\ S_{t+1|t+1} = S_{t+1|t}\varGamma \end{cases} \tag{17}$$

otherwise, where $\varGamma$ is a suitable (expansion) matrix. We impose $x_{0|0} = 0$ and $S_{0|0} = I$. By letting $\hat{x}_t := \hat{x}_{t|t}$ and $S_t = S_{t|t}$ we obtain the equantions

$$\begin{cases} \hat{x}_{t+1} = AS_tq(S_t^{-1}(x_t - \hat{x}_t)) + A\hat{x}_t + Bu_t \\ S_{t+1} = AS_tD^{-1} \end{cases} \tag{18}$$

if $S_t^{-1}(x_t - \hat{x}_t) \in Q$, and

$$\begin{cases} \hat{x}_{t+1} = A\hat{x}_t + Bu_t \\ S_{t+1} = AS_t\varGamma \end{cases} \tag{19}$$

otherwise. Finally we let $u_t = K_t\hat{x}_t$.

Observe that, if we define $e_t = x_t - \hat{x}_t$ then

$$\begin{cases} e_{t+1} = Ae_t - AS_tq(S_t^{-1}e_t) \\ S_{t+1} = AS_tD^{-1} \end{cases}$$

if $S_t^{-1}(e_t) \in Q$, and

$$\begin{cases} e_{t+1} = Ae_t \\ S_{t+1} = AS_t\varGamma \end{cases}$$

otherwise. This shows that e_t is independent of u_t and so that the separation principle holds true. Observe moreover that, if $\varGamma$ is chosen expanding enough, then the dynamics will consist first in a some number of zooming out steps in which $S_{t+1} = \varGamma AS_t$. Then after a certain number of steps we will have that $S_t^{-1}(e_t) \in Q$. It is easy to verify that, if $S_t^{-1}(e_t) \in Q$, then $S_{t+1}^{-1}(e_{t+1}) \in Q$. To obtain convergence it is sufficient to choose $\varGamma, S_0$ and $L_1, \dots, L_n$ in such a way that $S_t \to 0$.

We show finally how to adapt the proposed quantized estimator into the form (9,10). First define $L := \prod L_i$ and $\varOmega := \{0, 1, \dots, L\}$. We can find maps $\eta : \mathbb{R}^n \to \varOmega$ and $\bar{q} : \varOmega \to \mathbb{R}^n$ such that $\eta(x) = 0$ iff $x \notin Q$ and such that $q(x) = \bar{q} \circ \eta(x)$ for all $x \in Q$. The states of both the encoder and of the decoder/estimator is given by $(\hat{x}_t, S_t)$ The encoder state updating map coincides with the maps given by (18,19). The output map is

$$\omega_t = \eta(S_t^{-1}(x_t - \hat{x}_t))$$

The state state updating map of the decoder/estimator is given by

$$\begin{cases} \hat{x}_{t+1} = AS_t\bar{q}(\omega_t) + A\hat{x}_t + Bu_t \\ S_{t+1} = AS_tD^{-1} \end{cases}$$

if $\omega_t \neq 0$, and

$$\begin{cases} \hat{x}_{t+1} = A\hat{x}_t + Bu_t \\ S_{t+1} = AS_t\Gamma \end{cases} \tag{20}$$

if $\omega_t = 0$. The output is trivial since the estimate $\hat{x}_t$ is a part of the decoder/estimator state.

4 Optimal tree-structure vector quantizers and quantized LQ optimal control

In this section we introduce a way of defining an optimal family of tree-structure vector quantizers and we will show that this concept is strictly connected with the quantized linear quadratic optimal control problem.

Given a random vector X in $\mathbb{R}^n$ and two families of matrices $P_0, P_1, \ldots, P_{M-1} \in \mathbb{R}^{d\times n}$ and $R_0, R_1, \ldots, P_{M-1} \in \mathbb{R}^{d\times l}$ we say that a complexity L family of tree-structure vector quantizers $q_0, q_1, \ldots, q_{M-1}$ on $\mathbb{R}^n$ is optimal with respect to X and to $P_0, P_1, \ldots, P_{M-1}$ and $R_0, R_1, \ldots, P_{M-1}$ if they minimize the index

$$E\{\sum_{t=0}^{M-1}(P_tX - R_tq_t(X))^T(P_tX - R_tq_t(X))\} \tag{21}$$

over all possible complexity L families of tree-structure vector quantizers.

Consider now the following control problem. Given the discrete time linear system (1), find the quantized controller

$$u_t = k_t(x_t; u_0, u_1, \ldots, u_{t-1}) \qquad t = 0, 1, \ldots, M-1$$

with L quantization regions minimizing the cost function

$$J_M = E\{\sum_{t=0}^{M-1}(x_t^TQx_t + u_t^TRu_t) + x_M^TSx_M\} \tag{22}$$

where the expected value is done with respect to the probability density of the initial condition x_0. We assume as usual that $Q \in \mathbb{R}^{n\times n}$ is positive semi-definite, while $R \in \mathbb{R}^{m\times m}$ is positive definite.

Using standard Riccati difference equation arguments, it is possible to find $M \in \mathbb{R}^{n\times n}$ and $K_0, K_1, \ldots, K_{M-1} \in \mathbb{R}^{1\times n}$ and $N_0, N_1, \ldots, N_{M-1} \in \mathbb{R}$ such that

$$J_M = E\{x_0^TMx_0 + \sum_{t=0}^{M-1}(N_tu_t - K_tx_t)^T(N_tu_t - K_tx_t)\}$$

where the matrices N_t are invertible. Letting

$$\begin{gathered} k_t(x_t; u_0, u_1, \ldots, u_{t-1}) := N_t^{-1}(A^{t-1}Bu_0 + \cdots + Bu_{t-1}) + \\ +q_t(x_t - A^tBu_0 - \cdots - Bu_{t-1}) \end{gathered}$$

the cost becomes

$$J_M = E\{x_0^T M x_0 + \sum_{t=0}^{M-1} (P_t x_0 - R_t q_t(x_0))^T (P_t x_0 - R_t q_t(x_0))\}$$

where $P_t := K_t A^t$ and $R_t = N_t$. Now choose $q_0, q_1, \dots, q_{M-1}$ as an optimal family of tree-structure vector quantizers with respect to x_0 and to $P_0, P_1, \dots, P_{M-1}$ and $R_0, R_1, \dots, R_{M-1}$.

Observe conversely that, once a family of quantized feedback maps $k_t(x_t; u_0, u_1, \dots, u_{t-1})$, $t = 0, 1, \dots, M-1$ is fixed, the sequence of inputs $u_0, u_1, \dots, u_m$ is uniquely determined from x_0. Hence, we can define the family of vector quantizers

$$\begin{aligned} q_t : \mathbb{R}^n &\to \mathbb{R}^m \\ x_0 &\mapsto q_t(x_0) = u_t - N_t^{-1}(A^{t-1}Bu_0 - \cdots - Bu_{t-1}) \end{aligned}$$

Notice that the family of vector quantizers so defined is tree-structured. Notice moreover that, the feedback maps $k_t(x_t; u_0, u_1, \dots, u_{t-1})$ minimize the cost function (22), if and only if the quantizers $\bar{q}_t$ minimize the cost function (21).

5 Application to the scalar case

In this section we use a sequence of tree-structured quantizers in the construction of a quantized control like (8).

We propose here a simple construction of tree-structured sequence of quantizers on $\mathbb{R}$ whose performance can be analyzed quite simply. The construction of the sequence of quantizers q_t is based of the construction of the sequence of partitions $\mathcal{I}_t$ of $\mathbb{R}$ which of them having L^t subsets and such that $\mathcal{I}_{t+1}$ is a refinement of $\mathcal{I}_t$. The construction is recursive. Fix $D > 0$ and $\rho > 1$. Define

$$\mathcal{I}_0 := \{\mathbb{R}\}$$

Assume now that we are given the partition $\mathcal{I}_t$ and assume by induction that

$$\mathcal{I}_t = \{I_1, \dots, I_N, J', J''\}$$

such that I_i are finite intervals such that $\cup I_i = [-D\rho^t, D\rho^t]$ and $J' = (-\infty, -D\rho^t]$, $J'' = [D\rho^t, +\infty)$. Then $\mathcal{I}_{t+1}$ is obtained by cutting each interval I_i into L equal pieces and J', J'' into $2(L-1)$ equal finite pieces plus the two half straight lines $(-\infty, -D\rho^{t+1}]$ and $[D\rho^{t+1}, +\infty)$.

From partition $\mathcal{I}_t$ we define the quantizer q_t as follows:
For every $x \in I$, with I finite subset in $\mathcal{I}_t$, we let $q_t(x)$ be the center of I, while when I is not finite simply we let $q_t(x) = 0$.

It is clear that these quantizers are very rough and could be improved much from the knowledge of the probability density $p(x)$ of X. In any case it can be obtained the following asymptotic result when $p(x)$ satisfies a certain condition.

Lemma 1. *Assume that the random variable X has a probability density $p(x)$ such that*

$$|p(x)| \leq C|x|^{-\alpha} \tag{23}$$

where C, α are positive constants such that $\alpha > 3$. Let $q_0, q_1, \dots, q_t, \dots$ be the sequence of tree-structured quantizers constructed above. Assume finally that $L^2 \neq \rho^{(3-\alpha)}$. Then

$$E|X - q_t(X)|^2 \leq K_1 L^{-2t} + K_2 \rho^{(3-\alpha)t}$$

where K_1, K_2 are suitable positive constants.

Consider the scalar linear discrete time system

$$x_{t+1} = ax_t + u_t \tag{24}$$

where $x_t, u_t \in \mathbb{R}$. Apply the quantized control proposed in the previous section starting from the tree structured family of quantizers proposed above. Then we have the following result.

Corollary 1. *Assume that probability density of x_0 satisfies condition (23) and consider the quantized estimator (14) with controller $u_t = -a\hat{x}_t$. Then*

$$E(x_t^2) \leq K_1 \left(\frac{a^2}{L^2}\right)^t + K_2 (a^2 \rho^{(3-\alpha)})^t$$

where K_1, K_2 are suitable positive constants.

It is clear from the corollary that if $a < L$ and $a < \rho^{\frac{\alpha-3}{2}}$, then $E(x_t^2)$ converges to zero. If $\rho^{\frac{\alpha-3}{2}} < L$, then the convergence rate is a^2/L^2.

6 Zooming quantized controllers

In this section we analyze the performances of the zooming quantized feedback scheme which is an instance of the quantized control (5). Consider the system (24) and a quantized controller (5) with $S := \{D\rho^h (a/L)^k | h, k \in \mathbb{N}\}$ and

$$\begin{aligned} &f(s,x) = \begin{cases} L^{-1}s & \text{if } -s \leq x \leq s \\ \rho s & \text{otherwise} \end{cases} \\ &k(s,x) = -asq(x/s) \\ &s_0 = D\ , \end{aligned} \tag{25}$$

where $q : \mathbb{R} \to \mathbb{R}$ is a uniform quantized feedback with L quantization intervals such that $|y - q(y)| \leq 1/L$ if $|y| \leq 1$. In this context the following result holds true which shows that this quantizer has the same performance as the one proposed in the previous section in a completely different context. Notice that the proof of this result is much harder than the one given in the previous section [12].

Proposition 1. *Assume that probability density of x_0 satisfies condition (23). Consider the quantized controllers (25). Then*

$$E(x_t^2) \leq K_1 \left(\frac{a^2}{L^2}\right)^t + K_2(a^2\rho^{(3-\alpha)})^t$$

where K_1, K_2 are suitable positive constants.

7 General lower bounds on the stability rate

Consider the system (1) and assume now that the initial state x_0 is described by the probability density $p(x)$. We want to evaluate the evolution of $E[x_t^T x_t]$ and $t \to \infty$. A well-known general inequality of information theory provides a useful lower bound on $E[x_t^T x_t]$ based on properties the differential entropy.

Given a random vector X with values in $\mathbb{R}^n$ and density function $f(x)$, its differential entropy is defined as

$$h(X) := -\int f(x) \log f(x) dx$$

If a random vector U is singular namely it assumes a finite number of values $u_1, \dots, u_N$, then we have to consider its entropy defined as

$$H(U) := -\sum \mathbb{P}[U = u_i] \log \mathbb{P}[U = u_i] dx$$

A joint probability description of X and U is given by the family of unnormalized densities

$$f(x, u_i) := \frac{\partial}{\partial x}\mathbb{P}[X \leq x,\ U = u_i]$$

In fact for any $A \subset \mathbb{R}^n \times \mathbb{R}^n$, we have that

$$\mathbb{P}[(X, U) \in A] = \sum \int_{A_i} f(x, u_i) dx$$

where $A_i := \{x \in \mathbb{R}^n | (x, u_i) \in A\}$. We can now introduce the following hybrid definitions of entropies

$$h(X, U) := -\sum \int f(x, u_i) \log f(x, u_i) dx$$

$$h(X|U) := -\sum \int f_{X|u_i}(x, u_i) \log f_{X|u_i}(x, u_i) dx$$

where

$$f_{X|u_i}(x, u_i) := \frac{\partial}{\partial x}\mathbb{P}[X \leq x | U = u_i] = \frac{f(x, u_i)}{\mathbb{P}[U = u_i]}$$

Through these definitions it is possible to obtain all the classical inequalities on entropies, obtaining in this way the following result.

Lemma 2. *Let X, U be two random vectors in $\mathbb{R}^n$ and assume that U is discrete namely it assumes finitely many values. Then*

$$h(X+U) \geq h(X) - H(U)$$

Proof. By using the extensions of the classical inequalities on differential entropies we have that

$$\begin{aligned} h(X+U) \geq h(X+U|U) = h(X|U) = \\ = h(X,U) - h(U) \geq h(X) - h(U) \end{aligned}$$

Consider now the system (1) controlled by (5) or by (8). Then we have that

$$x_t = A^t x_0 + \sum_{i=0}^{t-1} A^{t-1-i} B u_i$$

which implies that

$$h(x_t) \geq h(A^t x_0) - H(\sum_{i=0}^{t-1} A^{t-1-i} B u_i)$$

Notice that $\sum_{i=0}^{t-1} A^{t-1-i} B u_i$ may assume at most L^t values and so

$$H(\sum_{i=0}^{t-1} A^{t-1-i} B u_i) \leq t \log L$$

Moreover, assuming that A is invertible, we have that

$$h(A^t x_0) = h(x_0) + t \log |A|$$

where $|A|$ is the absolute value of the determinant of A. This yields

$$h(x_t) \geq h(x_0) + t \log \frac{|A|}{L}.$$

Finally, using the fact that for any random vector X we have that

$$E[X^T X] \geq \frac{n}{2\pi e} e^{\frac{2}{n}} h(X)$$

we obtain that

$$E[x_t^T x_t] \geq \frac{n}{2\pi e} e^{\frac{2}{n} h(x_0)} \left(\frac{|A|}{L} \right)^{\frac{2}{n} t}$$

This inequality proves that if $|A| \geq L$ we can not achieve convergence.

8 Conclusions

In this paper we gave a brief introduction to the theory of quantized controller and its relations with the theory of control under communication constraints. We presented two approaches to this problem which are generally used also in all most of the literature devoted to this topic. We showed moreover that both these approaches to quantized control are in fact based on the design of a quantized state estimator plus a linear state feedback controller.

Finally we highlighted the relation between one quantized control approach and the theory of tree structure vector quantizer. Through this relation, we were able to translate the linear quadratic quantized optimal control problem into an optimal tree structure quantization problem. Moreover, by using elementary entropy inequalities from information theory we showed a simple lower bound on the rate which is necessary to stabilize a linear system.

References

1. R.W. Brockett and D. Liberzon. (2000). *IEEE Trans. Automatic Control*, AC-45:1279–1289.
2. D.F. Delchamps. (1990). *IEEE Trans. Automatic Control*, AC-35:916–924.
3. Jean-Charles Delvenne. (2006). *IEEE Trans. Automat. Control*, 51(2):298–303.
4. F. Fagnani and S. Zampieri. (2004). *IEEE Trans. Automatic Control*, AC-49:1534–1548.
5. F. Fagnani and S. Zampieri. (2005). *SIAM Journal on Contr. Optim.*, 44:816–866.
6. A. Gersho and R.M. Gray. (1992). *Vector Quantization and Signal Compression.* Kluwer Academic Publishers.
7. A.S. Matveev and A.V. Savkin. (2002). In *Proc. of CDC Conf.*, pages 4047–4052, Las Vegas.
8. G.N. Nair and R.J. Evans. (2004). *SIAM Journal on Contr. Optim.*, 43:413–436.
9. A.B. Nobel. (1997). *IEEE Trans. on Information Theory*, IT-43:1122–1133.
10. S. Stifter. (2002). *IEEE Trans. on Image Processing*, 12:1337–1348.
11. S. Tatikonda. (2000). *Control under communication constraints.* PhD thesis, MIT, Cambridge.
12. S. Tatikonda and S.K Mitter. (2004). *IEEE Trans. Automatic Control*, AC-49:1056–1068.

Geometric Methods for Output Regulation in Discrete-Time Switching Systems with Preview

Elena Zattoni

Dipartimento di Elettronica, Informatica e Sistemistica, Università di Bologna,
Viale Risorgimento 2, 40136 Bologna, Italy
ezattoni@deis.unibo.it

> *My memory of Antonio Lepschy is indissolubly tied to the years of my doctoral studies, when I enjoyed his lessons and, meanwhile, appreciated the noble character and the highest moral standards that shaped them.*

Summary. This contribution is focused on a geometric methodology devised to achieve optimization, expressed as the minimization of the ℓ_2 norm of the tracking error, of regulation transients caused by instantaneous, wide parameter variations occurring in discrete-time, linear systems. The regulated system switching law is assumed to be completely known a priori in a given time interval. A set of feedback regulators, designed according to the internal model principle, guarantee closed-loop asymptotic stability and asymptotic tracking of the reference signal generated by an exosystem, for each regulated system (i.e., for each pair to-be-controlled system/exosystem). The compensation scheme for optimization of transients consists of feedforward actions on the regulation loop and switching policies for suitably setting the states of the feedback regulators and those of the exosystems at the switching times. The theoretical bases of this approach comprise (i) a geometric interpretation, specifically aimed at discrete-time stabilizable and detectable systems, of the multivariable autonomous regulator problem and (ii) a non-recursive solution, still aimed at discrete-time stabilizable systems, of the finite-horizon optimal control problem with final state weighted by a generic quadratic function, based on a characterization of the structural invariant subspaces of the associated singular Hamiltonian system holding on the sole, fairly general, assumptions that guarantee the existence of the stabilizing solution of the corresponding discrete algebraic Riccati equation.

1 Introduction

A wise handling of the transients caused by severe parameter changes occurring in the regulated systems is crucial in a large number of control applications: e.g., in chemical or petrochemical plants, the processes may be subject to planned modifications; in numerical control of machine tools, the profiles to be tracked may be obtained by a set of appropriately switched exosystems; in flight and underwater vehicle control, the features of the systems involved may be subject to relevant adjustments during the course of the manoeuvres. A feedback regulator designed according to the well-known principles set forth in [1] and [2] attains closed-loop asymptotic stability and asymptotic tracking of the reference signals for sufficiently small parameter variations. However, in the presence of drastic changes of the regulated system parameters, control strategies specifically aimed at cancelling or minimizing the effects of those changes are required.

A set of sufficient conditions for perfect elimination of regulation transients in continuous-time linear systems subject to large parameter jumps was shown in [3]: in particular, that set includes a stabilizability condition which requires the to-be-controlled systems to be minimum-phase systems. In the later contribution [4], a sharper stabilizability condition was proved, where internal stabilizability of a peculiar robust controlled invariant subspace was considered. Nonetheless, the requirement of perfect elimination of regulation transients and the consequent geometric conditions which are sufficient to achieve it turn out to be too restrictive for many applications. Hence, in this work, the specification of exact elimination of regulation transients is replaced by a milder one, namely the minimization of the ℓ_2 norm of the tracking error caused by parameter variations. Moreover, any preview of the switching law available within a certain time interval is exploited, which further enhances the achievable performance. In these aspects, this work extends to switching systems the approach recently developed in other contexts, like signal rejection in particular (see e.g. [5, 6, 7]).

The theoretical bases for ℓ_2-optimization of regulation transients set forth in this work are twofold. First, some results on the geometric interpretation of the multivariable autonomous regulator problem, derived from the earlier works [4] and [8], point out the invariant subspaces (hence, the state trajectories) corresponding to the zero-error, steady-state condition for each autonomous extended system (i.e., for each pair regulated system/feedback regulator) of the set of the switched systems. Then, a technique, whose details were discussed further in [9], provides an analytic, non-recursive solution to a discrete-time finite-horizon optimal control problem with final state weighted by a generic quadratic function. This is crucial in defining a global criterion whose minimization leads to the feedforward control actions steering the state of the regulated system along trajectories that approach, in the optimal sense, the ideal trajectories on the invariant subspaces. Both the abovementioned aspects rely on an extensive use of geometric and structural concepts, in the framework established and shared in some rather recent books like, e.g., [10, 11, 12, 13, 14].

As mentioned above, discrete-time, stabilizable systems are specifically addressed. Removing the assumption of controllability which, by converse, was a standing assumption in [8], is natural when dealing with this kind of systems, since systems that switch from one configuration to another are likely to include uncontrollable (stable or pre-stabilized) parts. Moreover, addressing discrete-time systems better conforms to the industrial applications that motivate this work. Further, the discrete-time case is more appealing than the continuous-time case, since the solution, which, as will be shown, includes both feedforward controllers and state switching policies, turns out to be directly implementable as a single digital controller.

Notation: The symbols $\mathbb{Z}_0^+$, $\mathbb{Z}^+$, $\mathbb{R}$ are used for the sets of nonnegative integer numbers, positive integer numbers, real numbers, respectively. The symbols $\mathbb{C}^{\circ}$, $\mathbb{C}^{\odot}$, $\mathbb{C}^{\otimes}$ are used for the unit circle, the open set inside the unit circle, the open set outside the unit circle in the complex plane $\mathbb{C}$. Sets, vector spaces, and subspaces are denoted by capital script letters. The quotient space of a vector space $\mathcal{X}$ over a subspace $\mathcal{V} \subseteq \mathcal{X}$ is denoted by $\mathcal{X}/\mathcal{V}$. Matrices and linear maps are denoted by capital slanted letters. The restriction of a linear map A to an A-invariant subspace $\mathcal{J}$ is denoted by $A|_{\mathcal{J}}$. The spectrum, the image, and the kernel of A are denoted by $\sigma(A)$, $\mathrm{im}\, A$, and $\ker A$, respectively. The symbols A^{-1}, $A^{\dagger}$, and $A^{\top}$ are used for the inverse, the Moore-Penrose inverse, and the transpose of A. The symbols I and O are used for an identity matrix and a zero matrix of appropriate dimensions.

2 Optimization of Regulation Transients: Problem Formulation

The basic structure of the control scheme for ℓ_2-optimization of regulation transients is that of the multivariable autonomous regulator established in [2], with a replica of the exosystem eigenstructure for each regulated output. In addition, feedforward actions on the regulation loop and switching policies for the state of the feedback regulator and for the state of the exosystem are considered (see Fig. 1).

The symbol $\mathcal{L} = \{1, 2, \ldots, n_\ell\}$ denotes a finite index set. $\{\Sigma_1(\ell), \ell \in \mathcal{L}\} = \{(A_1(\ell), B_1(\ell), C_1(\ell)), \ell \in \mathcal{L}\}$, with $A_1(\ell) \in \mathbb{R}^{n_1 \times n_1}$, $B_1(\ell) \in \mathbb{R}^{n_1 \times p}$, and $C_1(\ell) \in \mathbb{R}^{q \times n_1}$, denotes the set of the to-be-controlled systems.

The pairs $(A_1(\ell), B_1(\ell))$ are assumed to be stabilizable for all $\ell \in \mathcal{L}$. $\{\Sigma_2(\ell), \ell \in \mathcal{L}\} = \{(A_2(\ell), E_2(\ell)), \ell \in \mathcal{L}\}$, with $A_2(\ell) \in \mathbb{R}^{n_2 \times n_2}$, $E_2(\ell) \in \mathbb{R}^{q \times n_2}$, and $\sigma(A_2(\ell)) \subset \mathbb{C}^{\otimes} \cup \mathbb{C}^{\circ}$, denotes the set of the exosystems. In particular, for any $\ell \in \mathcal{L}$, the matrices $A_2(\ell)$ and $E_2(\ell)$ should respectively have the structure

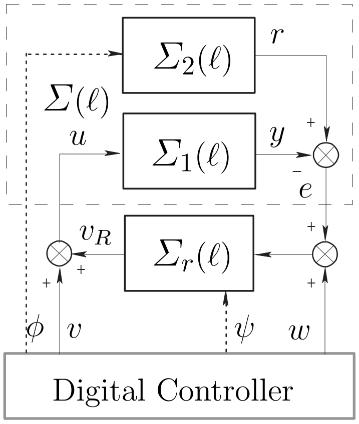


Fig. 1. Block diagram for ℓ_2-optimization of regulation transients.

$$A_2(\ell)=\left[\begin{array}{cccc} J(\ell) & O & \dots & O \\ O & J(\ell) & \ddots & \vdots \\ \vdots & \ddots & \ddots & O \\ O & \dots & O & J(\ell) \end{array}\right]\begin{array}{l} \}\ 1 \\ \}\ 2 \\ \vdots\ \vdots \\ \}\ q \end{array}, \quad E_2(\ell)=\left[\begin{array}{cccc} e_1(\ell)^\top & O & \dots & O \\ O & e_2(\ell)^\top & \ddots & \vdots \\ \vdots & \ddots & \ddots & O \\ O & \dots & O & e_q(\ell)^\top \end{array}\right],$$

where $J(\ell)\in\mathbb{R}^{n_J\times n_J}$, with $\ell\in\mathcal{L}$, denote the elementary Jordan block of the exosystem and $e_j(\ell)$, with $j=1,\dots,q$ and $\ell\in\mathcal{L}$, denote given vectors in $\mathbb{R}^{n_J}$, in order for the exosystem to produce, for any initial state, indipendent reference signals for each regulated output of $\Sigma_1(\ell)$. $\{\Sigma(\ell),\ell\in\mathcal{L}\}=\{(A(\ell),B(\ell),E(\ell)),\ell\in\mathcal{L}\}$ denotes the set of the *regulated systems*, defined as the connection of $\Sigma_1(\ell)$ and $\Sigma_2(\ell)$ such that, with $x=[x_1^\top\ x_2^\top]^\top$, $x_1\in\mathbb{R}^{n_1}$, $x_2\in\mathbb{R}^{n_2}$, $u\in\mathbb{R}^p$, and $e\in\mathbb{R}^q$, the state equations are

$$x_{k+1}=A(\ell)x_k+B(\ell)u_k, \tag{1}$$

$$e_k=E(\ell)x_k, \tag{2}$$

where

$$A(\ell)=\begin{bmatrix} A_1(\ell) & O \\ O & A_2(\ell) \end{bmatrix},\quad B(\ell)=\begin{bmatrix} B_1(\ell) \\ O \end{bmatrix},\quad E(\ell)=\begin{bmatrix} E_1(\ell) & E_2(\ell) \end{bmatrix}, \tag{3}$$

with $E_1(\ell)=-C_1(\ell)$. The pairs $(A(\ell),E(\ell))$ are assumed to be detectable for $\ell\in\mathcal{L}$. $\{\Sigma_r(\ell),\ell\in\mathcal{L}\}=\{(N(\ell),M(\ell),L(\ell),K(\ell)),\ell\in\mathcal{L}\}$, with $N(\ell)\in\mathbb{R}^{m\times m}$, $M(\ell)\in\mathbb{R}^{m\times q}$, $L(\ell)\in\mathbb{R}^{p\times m}$, $K(\ell)\in\mathbb{R}^{p\times q}$, denotes the set of the *feedback regulators*, designed according to the internal model principle, so that, with $z\in\mathbb{R}^m$ and $v_R\in\mathbb{R}^p$, the state equations are

$$z_{k+1}=N(\ell)z_k+M(\ell)e_k, \tag{4}$$

$$v_{R_k}=L(\ell)z_k+K(\ell)e_k. \tag{5}$$

Moreover, $\{\hat{\Sigma}(\ell), \ell \in \mathcal{L}\} = \{(\hat{A}(\ell), \hat{E}(\ell)), \ell \in \mathcal{L}\}$ denotes the set of the *autonomous extended systems*, defined as the connection of $\Sigma(\ell)$ and $\Sigma_r(\ell)$ such that, with $\hat{x} = [x_1^\top \; x_2^\top \; z^\top]^\top$, the state equations are

$$\hat{x}_{k+1} = \hat{A}(\ell)\hat{x}_k, \tag{6}$$

$$e_k = \hat{E}(\ell)\hat{x}_k, \tag{7}$$

where

$$\hat{A}(\ell) = \begin{bmatrix} A_1(\ell) + B_1(\ell)K(\ell)E_1(\ell) & B_1(\ell)K(\ell)E_2(\ell) & B_1(\ell)L(\ell) \\ O & A_2(\ell) & O \\ M(\ell)E_1(\ell) & M(\ell)E_2(\ell) & N(\ell) \end{bmatrix}, \tag{8}$$

and

$$\hat{E}(\ell) = \begin{bmatrix} E_1(\ell) & E_2(\ell) & O \end{bmatrix}. \tag{9}$$

The *control time window* is defined as the discrete-time interval $[k_{pre},\, k_{post})$, where $k_{pre} = 0$ and $k_{post} \in \mathbb{Z}^+$ are assumed. The *regulated-system switching law* is defined as a sequence $\varphi : [0,\, k_{post}) \to \mathcal{L}$, such that the set

$$\mathcal{K} = \{k_i \in (0,\, k_{post}) : \varphi(k_i) \neq \varphi(k_i - 1), \text{ with } k_i < k_{i+1},\ i = 1, 2, \ldots, N - 1\}$$

defines the *finite, ordered set of the switching times* (consistency of the above definitions implies that $N < k_{post}$). The *feedforward control on the to-be-controlled systems* is defined as a sequence $v : [0,\, k_{post}) \to \mathbb{R}^p$. The *feedforward control on the feedback regulators* is defined as a sequence $w : [0,\, k_{post}) \to \mathbb{R}^q$. The *feedback regulator state switching policy* is defined as a sequence $\psi : \mathcal{K} \to \mathbb{R}^m$. The *exosystem state switching policy* is defined as a sequence $\phi : \mathcal{K} \to \mathbb{R}^{n_2}$.

The overall system outlined so far is subject to the following conditions:

$\mathcal{C}$**1**.*known initial zero-error steady-state condition*: i.e., the regulated system is in a zero-error, steady-state condition at the time $k_{pre} = 0$ and its state $x_0 = x_O$ is known;

$\mathcal{C}$**2**.*consistency of the state of the to-be-controlled system*: i.e., in particular, the state of the to-be-controlled system at the switching times $k_i \in \mathcal{K}$ is computable as

$$x_{1_{k_i}} = A_1(\varphi(k_i - 1))\, x_{1_{k_i - 1}} + B_1(\varphi(k_i - 1))\, u_{k_i - 1},$$

for all $k_i \in \mathcal{K}$.

In this context, the problem of ℓ_2-optimization of the regulation transients with preview is stated as follows.

Problem 1. Given the set of regulated systems $\{\Sigma(\ell), \ell \in \mathcal{L}\}$, the set of feedback regulators $\{\Sigma_r(\ell), \ell \in \mathcal{L}\}$, and the regulated system switching law φ, find

(i) a feedforward control v on the to-be-controlled systems,
(ii) a feedforward control w on the feedback regulators,

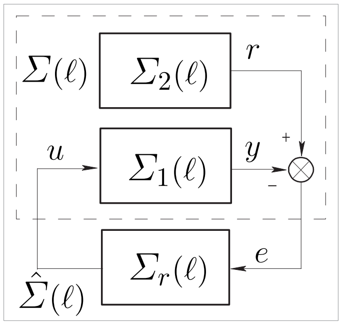


Fig. 2. Block diagram for the multivariable autonomous regulator problem.

(iii) a switching policy ψ of the state of the feedback regulators,
(iv) a switching policy ϕ of the state of the exosystems,

such that

$$\|e\|_{\ell_2} = \left(\sum_{k=0}^{\infty} e_k^\top e_k \right)^{1/2} ,$$

be minimal, on Conditions $\mathcal{C}1$ and $\mathcal{C}2$.

3 Geometric Characterization of the Multivariable Autonomous Regulator Problem

The solution of Problem 1 that will be discussed in Sect. 4 relies on the geometric interpretation of the multivariable autonomous regulator problem first presented for continuous-time, controllable and observable systems in [8] and lately extended to discrete-time, stabilizable and detectable systems in [4]. Some fundamental results shown in the abovementioned works are briefly reviewed herein with reference to the generic, ℓ-th autonomous extended system $\hat{\Sigma}(\ell)$, modeled by (6)–(9) and depicted in Fig. 2.

The multivariable autonomous regulator problem is stated in geometric terms as follows.

Problem 2. Given the regulated system (1)–(3), find a feedback regulator (4)–(5) such that, for the autonomous extended system (6)–(9), a $\hat{A}(\ell)$-invariant subspace $\hat{\mathcal{L}}(\ell)$ exists, which satisfies

(i) $\hat{\mathcal{L}}(\ell) \subseteq \hat{\mathcal{E}}(\ell)$, where $\hat{\mathcal{E}}(\ell) = \ker \hat{E}(\ell)$;
(ii) $\sigma(\hat{A}(\ell)|_{\hat{\mathcal{X}}/\hat{\mathcal{L}}(\ell)}) \subset \mathbb{C}^{\odot}$, where $\hat{\mathcal{X}}$ denotes the state space of (6)–(9).

The following theorem provides a set of necessary and sufficient conditions for solvability of Problem 2 functional to the further developments. In order to express those conditions, the subspace $\hat{\mathcal{P}} \subseteq \hat{\mathcal{X}}$ is introduced through the following definition:

$$\hat{\mathcal{P}} = \operatorname{im} \hat{P} = \operatorname{im} \begin{bmatrix} I & O \\ O & O \\ O & I \end{bmatrix}.$$

Since $\hat{A}(\ell)\hat{P} = \hat{P}\hat{S}(\ell)$ holds with

$$\hat{S}(\ell) = \begin{bmatrix} A_1(\ell) + B_1(\ell)K(\ell)E_1(\ell) & B_1(\ell)L(\ell) \\ M(\ell)E_1(\ell) & N(\ell) \end{bmatrix},$$

the subspace $\hat{\mathcal{P}}$ is $\hat{A}(\ell)$-invariant and $\sigma(\hat{A}(\ell)|_{\hat{\mathcal{P}}}) = \sigma(\hat{S}(\ell))$: i.e., the internal eigenvalues of $\hat{\mathcal{P}}$ match the poles of the regulation loop.

Theorem 1. *Problem 2 is solvable if and only if, for some regulator (4)–(5), a $\hat{A}(\ell)$-invariant subspace $\hat{\mathcal{W}}(\ell)$ exists, such that*

(i) $\hat{\mathcal{W}}(\ell) \subseteq \hat{\mathcal{E}}(\ell)$;
(ii)$\hat{\mathcal{W}}(\ell) \oplus \hat{\mathcal{P}} = \hat{\mathcal{X}}$;
(iii)$\sigma(\hat{A}(\ell)|_{\hat{\mathcal{X}}/\hat{\mathcal{W}}(\ell)}) \subset \mathbb{C}^{\odot}$.

Proof. If. For some regulator (4)–(5), let a $\hat{A}(\ell)$-invariant subspace $\hat{\mathcal{W}}(\ell) \subseteq \hat{\mathcal{X}}$, satisfying conditions (i)–(iii), exist. Then, Problem 2 is solvable, since conditions (i)–(ii) of the problem statement are satisfied with $\hat{\mathcal{L}}(\ell) = \hat{\mathcal{W}}(\ell)$.

Only if. Let Problem 2 be solvable, so that, for some regulator (4)–(5), a $\hat{A}(\ell)$-invariant subspace $\hat{\mathcal{L}}(\ell) \subseteq \hat{\mathcal{X}}$, satisfying conditions (i)–(ii) of Problem 2, exists. Let $\hat{\mathcal{W}}(\ell) \subseteq \hat{\mathcal{X}}$ be the subspace of the non-strictly stable modes of $\hat{A}(\ell)$. Hence, $\hat{\mathcal{W}}(\ell)$ is a $\hat{A}(\ell)$-invariant subspace, $\sigma(\hat{A}(\ell)|_{\hat{\mathcal{W}}(\ell)}) \subset \mathbb{C}^{\otimes} \cup \mathbb{C}^{\circ}$, and $\sigma(\hat{A}(\ell)|_{\hat{\mathcal{X}}/\hat{\mathcal{W}}(\ell)}) \subset \mathbb{C}^{\odot}$. Therefore, condition (iii) directly follows from the definition of $\hat{\mathcal{W}}(\ell)$. By definition of $\hat{\mathcal{P}}$, $\sigma(\hat{A}(\ell)) = \sigma(A_2(\ell)) \cup \sigma(\hat{A}(\ell)|_{\hat{\mathcal{P}}})$, where $\sigma(A_2(\ell)) \subset \mathbb{C}^{\otimes} \cup \mathbb{C}^{\circ}$ by assumption and $\sigma(\hat{A}(\ell)|_{\hat{\mathcal{P}}}) \subset \mathbb{C}^{\odot}$ since Problem 2 is assumed to be solvable for some regulator (4)–(5) and the internal eigenvalues of $\hat{\mathcal{P}}$ match the poles of the regulation loop. Therefore, condition (ii) is implied by $\sigma(\hat{A}(\ell)|_{\hat{\mathcal{W}}(\ell)}) = \sigma(A_2(\ell))$ and $\sigma(\hat{A}(\ell)|_{\hat{\mathcal{X}}/\hat{\mathcal{W}}(\ell)}) = \sigma(\hat{A}(\ell)|_{\hat{\mathcal{P}}})$. Finally, condition (i) is proved by contradiction. In fact, assuming $\hat{\mathcal{W}}(\ell) \not\subseteq \hat{\mathcal{L}}(\ell)$ would imply $\sigma(\hat{A}(\ell)|_{\hat{\mathcal{W}}(\ell)}) \not\subset \mathbb{C}^{\otimes} \cup \mathbb{C}^{\circ}$, due to condition (ii) of Problem 2. Therefore, $\hat{\mathcal{W}}(\ell) \subseteq \hat{\mathcal{L}}(\ell)$, which implies $\hat{\mathcal{W}}(\ell) \subseteq \hat{\mathcal{E}}(\ell)$ due to condition (i) of Problem 2. ■

Remark 1. The subspace $\hat{\mathcal{W}}(\ell)$ satisfying the conditions of Theorem 1 is the minimal $\hat{A}(\ell)$-invariant subspace satisfying the requirements of Problem 2. Moreover, a basis matrix $\hat{W}(\ell)$ of $\hat{\mathcal{W}}(\ell)$ has the structure:

$$\hat{\mathcal{W}}(\ell) = \operatorname{im} \hat{W}(\ell) = \operatorname{im} \begin{bmatrix} X_1(\ell) \\ I \\ Z(\ell) \end{bmatrix}.$$

□

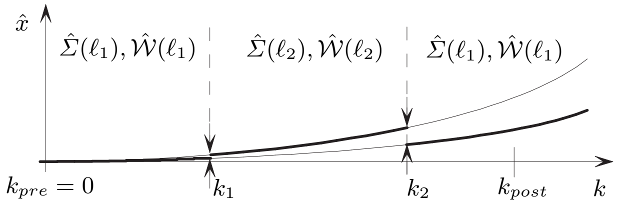


Fig. 3. Extended system state trajectories with switches occurring at k_1 and k_2.

In the light of the previous results, switching without transients would require that the state of the autonomous extended system be instantaneously forced to belong to the invariant subspace $\hat{\mathcal{W}}(\varphi(k_i))$ associated to the autonomous extended system $\hat{\Sigma}(\varphi(k_i))$ at each switching time $k_i \in \mathcal{K}$. However, while the state of the feedback regulator and that of the exosystem are accessible and may arbitrarily be imposed by means of the state switching policies ψ and ϕ, the state of the to-be-controlled system is subject to Condition $\mathcal{C}2$. This is the structural cause of the tracking error during the transients consequent to switches. Hence, the difference between the actual state and the corresponding state in a zero-error, steady-state condition must be steered in such a way that the regulation transients be minimal.

The abovementioned notions are illustrated below with reference to the case of two switches, respectively occurring at the time k_1 and k_2 as is shown in Fig. 3 and Table 1. It is assumed that in $[0,\, k_1)$ the extended autonomous system is $\hat{\Sigma}(\ell_1)$ with the associated invariant subspace $\hat{\mathcal{W}}(\ell_1)$, in $[k_1\;k_2)$ the extended autonomous system is $\hat{\Sigma}(\ell_2)$ with the associated invariant subspace $\hat{\mathcal{W}}(\ell_2)$, and, finally, in $[k_2,\, \infty)$ the extended autonomous system is $\hat{\Sigma}(\ell_1)$, with the associated invariant subspace $\hat{\mathcal{W}}(\ell_1)$. The control action starts at $k_{pre} = 0$ and ends at k_{post}. The curves in Fig. 3 represent the ideal trajectories of the extended state on the invariant subspaces $\hat{\mathcal{W}}(\ell_1)$ and $\hat{\mathcal{W}}(\ell_2)$. Those trajectories are completely known by virtue of Condition $\mathcal{C}1$. The arrows point out the differences between the ideal states belonging to the invariant subspaces $\hat{\mathcal{W}}(\ell_1)$ and $\hat{\mathcal{W}}(\ell_2)$ at the switching times k_1 and k_2. The ideal extended states at the switching times are also listed in the first couple of rows of Table 1. The third and fourth row respectively show the exosystem state and the feedback regulator state that must be imposed at the time k_1 and k_2 through the state switching policies ϕ and ψ. Finally, the fifth row shows the differences in the state of the to-be-controlled system. The synthesis of the control input sequence to be applied to the plant and that of the correction to be applied to the feedback regulator in order to steer these differences along optimal trajectories is discussed in the following section.

4 Optimization of Regulation Transients: Problem Solution

This section is focused on the synthesis of the feedforward actions v and w, which, along with the state switching policies ϕ and ψ considered in Sect. 3, guarantee the minimal ℓ_2 norm of the tracking error for a given control time window $[k_{pre}, k_{post})$ and a given regulated system switching law φ.

The procedure that will be detailed in this section reduces to the solution of a sequence of interconnected, finite-horizon optimal control problems of the type discussed in Appendix. In particular, the non-recursive solution presented therein is crucial in order to formulate the connection of those problems through an appropriate definition of the global cost functional.

Since the generalization to an arbitrary number of switches is trivial, the double switch introduced in Sect. 3 will be considered henceforth. In order to avoid notation clutter, the subscript "1" previously used to denote the state variable and the system matrices of the to-be-controlled systems will henceforth be omitted. Moreover, the variable $x \in \mathbb{R}^{n_1}$ will be used to denote the difference between the actual state, corresponding to the optimal trajectory and the ideal state, corresponding to the ideal, zero-error trajectory.

The following statements define the concatenation of the three subproblems. The propositions providing the expressions for the optimal values of the cost functionals directly follow from Corollary 4 in Appendix.

Problem 3. Consider the system

$$x_{k+1} = A(\ell_1)x_k + B(\ell_1)v_k,$$

and the cost functional

$$J_{[k_2,\,\infty)} = \sum_{k=k_2}^{k_{post}} x_k^\top C(\ell_1)^\top C(\ell_1)x_k + x_{k_{post}}^\top Z(\ell_1)x_{k_{post}},$$

where $Z(\ell_1)$ denotes the solution of the symmetric Stein equation

Table 1. Ideal extended states, state switching policies and differences in the to-be-controlled system states with switches at k_1 and k_2.

	0	k_1	k_2
$\hat{\Sigma}(\ell_1)$	$\hat{x}_{\hat{\Sigma}(\ell_1),0}$	$\hat{x}_{\hat{\Sigma}(\ell_1),k_1}$	$\hat{x}_{\hat{\Sigma}(\ell_1),k_2}$
$\hat{\Sigma}(\ell_2)$	$\hat{x}_{\hat{\Sigma}(\ell_2),0}$	$\hat{x}_{\hat{\Sigma}(\ell_2),k_1}$	$\hat{x}_{\hat{\Sigma}(\ell_2),k_2}$
ϕ	–	${x_2}_{\hat{\Sigma}(\ell_2),k_1}$	${x_2}_{\hat{\Sigma}(\ell_1),k_2}$
ψ	–	$z_{\hat{\Sigma}(\ell_2),k_1}$	$z_{\hat{\Sigma}(\ell_1),k_2}$
dx_1	–	${x_1}_{\hat{\Sigma}(\ell_2),k_1} - {x_1}_{\hat{\Sigma}(\ell_1),k_1}$	${x_1}_{\hat{\Sigma}(\ell_1),k_2} - {x_1}_{\hat{\Sigma}(\ell_2),k_2}$

$$A(\ell_1)^\top Z(\ell_1) A(\ell_1) - Z(\ell_1) + C(\ell_1)^\top C(\ell_1) = 0,$$

and $x_{k_{post}}$ denotes the state at the time k_{post}. Find an admissible control sequence v_k, with $k \in [k_2,\, k_{post})$, such that $J_{[k_2,\,\infty)}$ be minimal.

Proposition 1. *Refer to Problem 3. Let $\bar{x}_{k_2}$ denote the state at the time k_2. The optimal value of $J_{[k_2,\,\infty)}$ is*

$$J^O_{[k_2,\,\infty)} = \bar{x}_{k_2}^\top S_0 \bar{x}_{k_2} + 2\sigma_0^\top \bar{x}_{k_2} + \rho_0,$$

where S_0, σ_0, and ρ_0 are defined according to Corollary 4.

Proposition 2. *Let $\bar{x}_{k_2} = x_{k_2} - dx_{k_2}$, where x_{k_2} denotes the state reached at the end of the time interval $[k_1,\, k_2]$ and dx_{k_2} denotes the difference between the ideal states on the two invariant subspaces at the same time. Then, $J^O_{[k_2,\,\infty)}$ expressed as a function of x_{k_2} is*

$$J^O_{[k_2,\,\infty)} = x_{k_2}^\top S_1 x_{k_2} + 2\sigma_1^\top x_{k_2} + \rho_1,$$

where

$$\begin{aligned} S_1 &= S_0, \\ \sigma_1 &= -S_0 dx_{k_2} + \sigma_0, \\ \rho_1 &= dx_{k_2}^\top S_0 dx_{k_2} - 2\sigma_0^\top dx_{k_2} + \rho_0. \end{aligned}$$

Problem 4. Consider the system

$$x_{k+1} = A(\ell_2) x_k + B(\ell_2) v_k,$$

and the cost functional

$$J_{[k_1,\,\infty)} = \sum_{k=k_1}^{k_2-1} x_k^\top C(\ell_2)^\top C(\ell_2) x_k + x_{k_2}^\top S_1 x_{k_2} + 2\sigma_1^\top x_{k_2} + \rho_1.$$

Find an admissible control sequence v_k, with $k \in [k_1,\, k_2)$, such that $J_{[k_1,\,\infty)}$ be minimal.

Proposition 3. *Refer to Problem 4. Let $\bar{x}_{k_1}$ denote the state at the time k_1. The optimal value of $J_{[k_1,\,\infty)}$ is*

$$J^O_{[k_1,\,\infty)} = \bar{x}_{k_1}^\top S_2 \bar{x}_{k_1} + 2\sigma_2^\top \bar{x}_{k_1} + \rho_2,$$

where S_2, σ_2, and ρ_2 are defined according to Corollary 4.

Proposition 4. *Let $\bar{x}_{k_1} = x_{k_1} - dx_{k_1}$, where x_{k_1} denotes the state reached at the end of the time interval $[k_{pre},\, k_1]$ and dx_{k_1} denotes the difference between the ideal states on the two invariant subspaces at the same time. Then, $J^O_{[k_1,\,\infty)}$ expressed as a function of x_{k_1} is*

$$J^O_{[k_1,\,\infty)} = x_{k_1}^\top S_3 x_{k_1} + 2\sigma_3^\top x_{k_1} + \rho_3,$$

where

$$\begin{aligned}S_3 &= S_2,\\ \sigma_3 &= -S_2 dx_{k_1} + \sigma_2,\\ \rho_3 &= dx_{k_1}^{\top} S_2 dx_{k_1} - 2\sigma_2^{\top} dx_{k_1} + \rho_2.\end{aligned}$$

Problem 5. Consider the system

$$x_{k+1} = A(\ell_1)x_k + B(\ell_1)v_k,$$

and the cost functional

$$J_{[k_{pre},\,\infty)} = \sum_{k=k_{pre}}^{k_1-1} x_k^{\top} C(\ell_1)^{\top} C(\ell_1)x_k + x_{k_1}^{\top} S_3 x_{k_1} + 2\sigma_3^{\top} x_{k_1} + \rho_3.$$

Find an admissible control sequence v_k, with $k \in [k_{pre},\, k_1)$, such that $J_{[k_{pre},\,\infty)}$ be minimal.

Proposition 5. *Refer to Problem 5. Let $x_{k_{pre}}$ denote the state at the time k_{pre}. The optimal value of $J_{[k_{pre},\,\infty)}$ is*

$$J^{O}_{[k_{pre},\,\infty)} = x_{k_{pre}}^{\top} S_4 x_{k_{pre}} + 2\sigma_4^{\top} x_{k_{pre}} + \rho_4,$$

where S_4, σ_4, and ρ_4 are defined according to Corollary 4. Moreover, due to Condition $\mathcal{C}1$, $x_{k_{pre}} = 0$. Hence,

$$J^{O}_{[k_{pre},\,\infty)} = \rho_4,$$

known.

Consequently, Corollary 2 and Corollary 3 in Appendix provide the state trajectory and the optimal feedforward action v. The correction on the feedback regulator is then derived in order to compensate the effect of feedback when the input sequence defined to minimize the regulation transients is applied to the to-be-controlled system: namely, $w_k = C(\ell_1)x_k$ with $k \in [k_{pre},\, k_1)$ and $k \in [k_2,\, k_{post})$, $w_k = C(\ell_2)x_k$ with $k \in [k_1,\, k_2)$.

5 Conclusions

The methodology, discussed in this contribution, for ℓ_2-optimization of the tracking error caused by a finite sequence of switches of the regulated system within a given control time window has been derived by combining results of the geometric approach applied to the multivariable autonomous regulator problem and results of the same approach applied to generalized Riccati theory. The structural properties of the autonomous regulator point out the invariant subspaces containing the ideal state trajectories which would correspond to the ideal, zero-error, steady-state conditions. The structural invariant subspaces of the singular Hamiltonian system associated to the finite-horizon optimal control problems with final state weighted by a generic quadratic function leads to the characterization of the actual state trajectories corresponding to the minimum of a global cost functional appropriately defined. Hence, both the state switching policies for the exosystems and the feedback regulators and the feedforward actions on the regulation loop are completely determined.

A Appendix

In this section, the finite-horizon optimal control problem with final state weighted by a generic quadratic function is solved through a non-recursive procedure based on the original characterization of a pair of structural invariant subspaces of the associated singular Hamiltonian system. The results hold on fairly general assumptions: namely, those that guarantee solvability of an appropriate discrete algebraic Riccati equation as well as solvability of a corresponding symmetric Stein equation. An earlier version of these arguments was presented in [9].

The problem is defined by the discrete, time-invariant, linear system

$$x_{k+1} = Ax_k + Bu_k, \tag{10}$$
$$e_k = Cx_k + Du_k, \tag{11}$$

with state $x \in \mathbb{R}^n$, input $u \in \mathbb{R}^p$, output $e \in \mathbb{R}^q$, and the cost functional

$$J = \sum_{k=0}^{k_f-1} e_k^\top e_k + \Phi(x_{k_f}) \quad \text{with} \quad k_f \geq 1 \tag{12}$$

and

$$\Phi(x_{k_f}) = x_{k_f}^\top S_f x_{k_f} + 2\sigma_f^\top x_{k_f} + \rho_f, \tag{13}$$

or, respectively, by

$$x_{k+1} = Ax_k + Bu_k, \tag{14}$$

and

$$J = \sum_{k=0}^{k_f-1} \left(x_k^\top Q x_k + 2x_k^\top S u_k + u_k^\top R u_k \right) + \Phi(x_{k_f}) \quad \text{with} \quad k_f \geq 1 \tag{15}$$

and the same definition of $\Phi(x_{k_f})$.

The initial state is known: i.e., $x_0 = x_O$, with x_O given. The constant ρ_f in (13) is irrelevant to the problem solution. However, it eases the application of the results presented in this Appendix to the set of problems considered in Sect. 4.

The following assumptions are introduced:

$\boldsymbol{\mathcal{A}}\mathbf{1}$. (A,B) stabilizable;

$\boldsymbol{\mathcal{A}}\mathbf{2}$. (A,B,C,D) left invertible;

$\boldsymbol{\mathcal{A}}\mathbf{3}$. $\mathcal{Z}(A,B,C,D)\cap\mathbb{C}^{\circ}=\emptyset$, with $\mathcal{Z}(A,B,C,D)$ denoting the set of the invariant zeros of (A,B,C,D);

$\boldsymbol{\mathcal{A}}\mathbf{4}$. $\begin{bmatrix} C^\top \\ D^\top \end{bmatrix}\begin{bmatrix} C & D \end{bmatrix} = \begin{bmatrix} Q & S \\ S^\top & R \end{bmatrix} \geq 0$;

$\boldsymbol{\mathcal{A}}\mathbf{5}$. $S_f = S_f^\top \geq 0$;

The Lagrange multiplier approach (see e.g. [15]) leads to a two-point boundary value problem defined by the state equations, the costate equations, the stationarity conditions and the boundary conditions. The results proved in the remainder of this section define an alternative, non-recursive procedure to solve this problem, holding on the general assumptions $\mathcal{A}1$–$\mathcal{A}5$.

The difference equations of the abovementioned, two-point boundary-value problem can be written as the state-space generalized system

$$\begin{bmatrix} I & O & O \\ O & -A^\top & O \\ O & -B^\top & O \end{bmatrix}\begin{bmatrix} x_{k+1} \\ p_{k+1} \\ u_{k+1} \end{bmatrix} = \begin{bmatrix} A & O & B \\ Q & -I & S \\ S^\top & O & R \end{bmatrix}\begin{bmatrix} x_k \\ p_k \\ u_k \end{bmatrix}, \tag{16}$$

also called the *singular Hamiltonian system*. The matrix on the left-hand side of (16) will be denoted by M, that on the right-hand side will be denoted by N. The matrix pencil $\lambda M - N$ is assumed to have non-vanishing determinant, i.e. $\det(\lambda M - N) \not\equiv 0$.

The boundary conditions are

$$x_0 = x_O, \tag{17}$$

$$p_{k_f} = S_f x_{k_f} + \sigma_f. \tag{18}$$

Assumptions $\mathcal{A}1$–$\mathcal{A}4$ are sufficient to guarantee the existence and uniqueness of the stabilizing solution of the discrete algebraic Riccati equation

$$P = -(A^\top PB + S)(R + B^\top PB)^{-1}(B^\top PA + S^\top) + A^\top PA + Q, \tag{19}$$

$$R + B^\top PB > 0, \tag{20}$$

see e.g. [11]. The stabilizing solution, henceforth denoted as P_+, is also positive semi-definite and is the largest real symmetric solution of (19)–(20). Let

$$K_+ = (R + B^\top P_+ B)^{-1}(B^\top P_+ A + S^\top), \tag{21}$$

$$A_+ = A - BK_+. \tag{22}$$

The condition $\sigma(A_+)\subset\mathbb{C}^{\odot}$, implied by (21)–(22), is sufficient to guarantee the existence and uniqueness of the solution W of the symmetric Stein equation

$$A_+ W A_+^\top - W + B(R + B^\top P_+ B)^{-1}B^\top = 0, \tag{23}$$

see e.g. [12]. Let

$$\bar{K}_+ = (R + B^\top P_+ B)^{-1}(B^\top - B^\top P_+ AWA_+^\top - S^\top WA_+^\top). \qquad (24)$$

The following lemmas introduce the geometric characterization of the pair of structural invariant subspaces related to the singular Hamiltonian system.

Lemma 1. *The subspace*

$$\mathcal{V}_1 = \operatorname{im} V_1 = \operatorname{im} \begin{bmatrix} I \\ P_+ \\ -K_+ \end{bmatrix}, \tag{25}$$

is a deflating subspace of the matrix pencil $\lambda M - N$. The spectrum of the pencil restricted to the subspace $\mathcal{V}_1$, denoted by $(\lambda M - N)|_{\mathcal{V}_1}$ is equivalent to $\lambda I - A_+$.

Proof. It is implied by

$$MV_1 S = NV_1 \text{ with } S = A_+.$$

In fact, the previous equation, written according to the partition introduced in (16) and (25), produces the identities listed below, where (19)–(22) have been taken into account. The row blocks, from the first to the third, are considered in order.

- 1st row block:
$$A_+ = A - BK_+.$$
- 2nd row block:
$$\begin{aligned}&-A^\top P_+ A_+ = Q - P_+ - SK_+,\\ &-A^\top P_+ A + A^\top P_+ BK_+ = Q - P_+ - SK_+,\\ &P_+ = -(A^\top P_+ B + S)(R + B^\top P_+ B)^{-1}(B^\top P_+ A + S^\top) + A^\top P_+ A + Q.\end{aligned}$$
- 3rd row block:
$$\begin{aligned}&-B^\top P_+ A_+ = S^\top - RK_+,\\ &-B^\top P_+ A + B^\top P_+ BK_+ = S^\top - RK_+,\\ &(B^\top P_+ B + R)(R + B^\top P_+ B)^{-1}(B^\top P_+ A + S^\top) = B^\top P_+ A + S^\top,\\ &B^\top P_+ A + S^\top = B^\top P_+ A + S^\top.\end{aligned}$$

■

Lemma 2. *The subspace*

$$\mathcal{V}_2 = \operatorname{im} V_2 = \operatorname{im} \begin{bmatrix} WA_+^\top \\ (P_+ W - I)A_+^\top \\ \bar{K}_+ \end{bmatrix}, \tag{26}$$

is a deflating subspace of the matrix pencil $\lambda N - M$. The spectrum of the pencil restricted to the subspace $\mathcal{V}_2$, denoted by $(\lambda N - M)|_{\mathcal{V}_2}$, is equivalent to $\lambda I - A_+^\top$.

Proof. It is implied by

$$NV_2S = MV_2 \text{ with } S = A_+^\top.$$

In fact, the previous equation, written according to the partition introduced in (16) and (26), produces the identities listed below, where (19)–(24) have been taken into account. Again, the row blocks are considered in order.

- 1st row block:

$$\begin{aligned}
&AWA_+^\top + B\bar{K}_+ = W,\\
&AWA_+^\top + B(R + B^\top P_+ B)^{-1}B^\top - BK_+WA_+^\top = W,\\
&A_+WA_+^\top - W + B(R + B^\top P_+ B)^{-1}B^\top = 0.
\end{aligned}$$

- 2nd row block:

$$\begin{aligned}
&QWA_+^\top - (P_+W - I)A_+^\top + S\bar{K}_+ = -A^\top(P_+W - I),\\
&QWA_+^\top - P_+WA_+^\top - K_+^\top B^\top + S\bar{K}_+ = -A^\top P_+W,\\
&QWA_+^\top - P_+WA_+^\top - A^\top P_+B(R + B^\top P_+B)^{-1}B^\top\\
&\quad -S(R + B^\top P_+B)^{-1}(B^\top P_+A + S^\top)WA_+^\top = -A^\top P_+W,\\
&Q - P_+ - S(R + B^\top P_+B)^{-1}(B^\top P_+A + S^\top) = -A^\top P_+A_+,\\
&P_+ = -(A^\top P_+B + S)(R + B^\top P_+B)^{-1}(B^\top P_+A + S^\top) + A^\top P_+A + Q.
\end{aligned}$$

- 3rd row block:

$$\begin{aligned}
&S^\top WA_+^\top + R\bar{K}_+ = -B^\top(P_+W - I),\\
&S^\top WA_+^\top + R(R + B^\top P_+B)^{-1}B^\top - R(R + B^\top P_+B)^{-1}(B^\top P_+A + S^\top)\cdot\\
&\quad WA_+^\top = -B^\top P_+A_+WA_+^\top + B^\top - B^\top P_+B(R + B^\top P_+B)^{-1}B^\top,\\
&S^\top - R(R + B^\top P_+B)^{-1}(B^\top P_+A + S^\top) = -B^\top P_+A_+,\\
&B^\top P_+A + S^\top = B^\top P_+A + S^\top.
\end{aligned}$$

■

The following theorem provides the geometric characterization of all the admissible trajectories of the singular Hamiltonian system.

Theorem 2. *A trajectory* $\xi_k = \begin{bmatrix} x_k^\top & p_k^\top & u_k^\top \end{bmatrix}^\top$*, with* $k \in [0,\, k_f)$*, is admissible for the singular Hamiltonian system (16) if and only if*

$$\xi_k = V_1 A_+^k \alpha + V_2 (A_+^\top)^{k_f - k - 1}\beta, \quad \textit{with} \quad k \in [0,\, k_f), \tag{27}$$

where V_1*,* V_2 *are respectively defined as in (25),(26), and* $\alpha, \beta \in \mathbb{R}^n$ *are parameters.*

Proof. If. The trajectory ξ_k, with $k \in [0,\, k_f)$, satisfies

$$M\xi_{k+1} = N\xi_k, \quad \text{with} \quad k \in [0,\, k_f),$$

for any $\alpha, \beta \in \mathbb{R}^n$. In fact, by virtue of (16), (25), (26), (27), the previous equation can also be written as

$$\begin{bmatrix} I \\ -A^\top P_+ \\ -B^\top P_+ \end{bmatrix} A_+^{k+1}\alpha + \begin{bmatrix} WA_+^\top \\ -A^\top (P_+W - I)A_+^\top \\ -B^\top (P_+W - I)A_+^\top \end{bmatrix} (A_+^\top)^{k_f-k-2}\beta =$$

$$\begin{bmatrix} A - BK_+ \\ Q - P_+ - SK_+ \\ S^\top - RK_+ \end{bmatrix} A_+^k\alpha + \begin{bmatrix} AWA_+^\top + B\bar{K}_+ \\ QWA_+^\top - (P_+W - I)A_+^\top + S\bar{K}_+ \\ S^\top WA_+^\top + R\bar{K}_+ \end{bmatrix} (A_+^\top)^{k_f-k-1}\beta,$$

or, equivalently, as

$$\begin{bmatrix} A_+ - A + BK_+ \\ -A^\top P_+A_+ - Q + P_+ + SK_+ \\ -B^\top P_+A_+ - S^\top + RK_+ \end{bmatrix} A_+^k\alpha =$$

$$\begin{bmatrix} -W + AWA_+^\top + B\bar{K}_+ \\ A^\top (P_+W - I) + QWA_+^\top - (P_+W - I)A_+^\top + S\bar{K}_+ \\ B^\top (P_+W - I) + S^\top WA_+^\top + R\bar{K}_+ \end{bmatrix} (A_+^\top)^{k_f-k-1}\beta.$$

The previous equalities hold for any $\alpha, \beta \in \mathbb{R}^n$, since each row block of the matrix on the left is equal to zero due to the respective identities shown in the proof of Lemma 1 and each row block of the matrix on the right is equal to zero due to the respective identities shown in the proof of Lemma 2.

Only if. It follows from the structure of (16) (which derives from the Lagrange-multiplier approach), Lemma 1, and Lemma 2. ■

The following corollaries provide the solution of the finite-horizon optimal control problem defined by (10)–(13) through the geometric characterization of the solutions of the singular Hamiltonian system introduced in Theorem 2.

In the light of Theorem 2, the state and costate trajectories can be written as

$$\begin{bmatrix} x_k \\ p_k \end{bmatrix} = \begin{bmatrix} I \\ P_+ \end{bmatrix} A_+^k\alpha + \begin{bmatrix} W \\ P_+W - I \end{bmatrix} (A_+^\top)^{k_f-k}\beta \quad \text{with} \quad k \in [0,\, k_f]. \tag{28}$$

Hence, Corollary 1 and Remark 2, which follow, provide the criterion to select the trajectories of the singular Hamiltonian system solving the original, two-point boundary-value problem.

Corollary 1. *Let* $\begin{bmatrix} x_O^\top & \sigma_f^\top \end{bmatrix}^\top \in \operatorname{im} \varPsi$, *where*

$$\varPsi = \begin{bmatrix} I & W(A_+^\top)^{k_f} \\ (P_+ - S_f)A_+^{k_f} & (P_+ - S_f)W - I \end{bmatrix}. \tag{29}$$

A trajectory $\xi_k = \begin{bmatrix} x_k^\top & p_k^\top & u_k^\top \end{bmatrix}^\top$, *with* $k \in [0,\, k_f)$, *of the singular Hamiltonian system (16), satisfying the boundary conditions (17),(18), is determined by*

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \varPsi^\dagger \begin{bmatrix} x_O \\ \sigma_f \end{bmatrix}. \tag{30}$$

Proof. From (28) it follows that

$$x_0 = \alpha + W(A_+^\top)^{k_f}\beta, \quad x_{k_f} = A_+^{k_f}\alpha + W\beta, \quad p_{k_f} = P_+A_+^{k_f}\alpha + (P_+W - I)\beta.$$

Hence, (17), (18) can also be written as

$$\alpha + W(A_+^\top)^{k_f}\beta = x_O, \quad (P_+ - S_f)A_+^{k_f}\alpha + ((P_+ - S_f)W - I)\beta = \sigma_f,$$

or, in a more compact form, as

$$\begin{bmatrix} I & W(A_+^\top)^{k_f} \\ (P_+ - S_f)A_+^{k_f} & (P_+ - S_f)W - I \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} x_O \\ \sigma_f \end{bmatrix},$$

which, with definition (29), completes the proof. ■

Remark 2. If $\begin{bmatrix} x_O^\top & \sigma_f^\top \end{bmatrix}^\top \notin \operatorname{im}\Psi$, the two-point boundary-value problem is not solvable: i.e., the solution is unbounded and at infinity. Otherwise, the following two cases are discriminated. If Ψ is invertible, the trajectory ξ_k, $k \in [0,\, k_f)$, satisfying the assigned boundary conditions is unique. The Moore-Penrose inverse of Ψ matches the inverse and, according to the well-known formula for the inverse of a partitioned matrix,

$$\Psi^{-1} = \begin{bmatrix} I + W(A_+^\top)^{k_f}\Delta^{-1}(P_+ - S_f)A_+^{k_f} & -W(A_+^\top)^{k_f}\Delta^{-1} \\ -\Delta^{-1}(P_+ - S_f)A_+^{k_f} & \Delta^{-1} \end{bmatrix},$$

where $\Delta = (P_+ - S_f)(W - A_+^{k_f}W(A_+^\top)^{k_f}) - I$. Conversely, if Ψ is not invertible, the trajectory ξ_k, $k \in [0,\, k_f)$, satisfying the boundary conditions may be not unique and the set of all admissible values of $\alpha, \beta \in \mathbb{R}^n$ is characterized by

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \Psi^\dagger \begin{bmatrix} x_O \\ \sigma_f \end{bmatrix} + \Omega\gamma,$$

where Ω denotes a basis matrix of $\ker\Psi$ and $\gamma \in \mathbb{R}^\nu$, with $\nu = \dim(\ker\Psi)$, denotes a free parameter vector. □

Let $\begin{bmatrix} x_O^\top & \sigma_f^\top \end{bmatrix}^\top \in \operatorname{im}\Psi$ and let $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}^{n\times n}$ be such that

$$\begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{bmatrix} = \Psi^\dagger,$$

where $\Psi^\dagger$ is assumed to be partitioned according to (30). Then, $\alpha, \beta \in \mathbb{R}^n$, determined according to Corollary 1, can be expressed as

$$\alpha = \alpha_1 x_O + \alpha_2\sigma_f, \qquad \beta = \beta_1 x_O + \beta_2\sigma_f. \tag{31}$$

Consequently, the following statements provide analytic expressions for state trajectories, control laws, and optimal cost solving the finite-horizon optimal control problem with final state weighted by a generic quadratic function.

Corollary 2. *An optimal state trajectory x_k, with $k \in [0,\, k_f]$, for the finite-horizon optimal control problem defined by (10)–(13), with known initial state x_O, under assumptions $\mathcal{A}1$–$\mathcal{A}5$, is*

$$x_k = \mathrm{X}_k x_O + \overline{\mathrm{x}}_k,$$

where

$$\begin{aligned}\mathrm{X}_k &= A_+^k \alpha_1 + W (A_+^\top)^{k_f - k} \beta_1,\\ \overline{\mathrm{x}}_k &= \left(A_+^k \alpha_2 + W (A_+^\top)^{k_f - k} \beta_2\right) \sigma_f,\end{aligned}$$

with $k \in [0,\, k_f]$.

Proof. It follows from (28) and (31). ∎

Corollary 3. *An optimal control law u_k, with $k \in [0,\, k_f)$, for the finite-horizon optimal control problem defined by (10)–(13), with known initial state x_O, under assumptions $\mathcal{A}1$–$\mathcal{A}5$, is*

$$u_k = \mathrm{U}_k x_O + \overline{\mathrm{u}}_k,$$

where

$$\begin{aligned}\mathrm{U}_k &= -K_+ A_+^k \alpha_1 + \bar{K}_+ (A_+^\top)^{k_f - k - 1} \beta_1,\\ \overline{\mathrm{u}}_k &= \left(-K_+ A_+^k \alpha_2 + \bar{K}_+ (A_+^\top)^{k_f - k - 1} \beta_2\right) \sigma_f,\end{aligned}$$

with $k \in [0,\, k_f)$.

Proof. In light of Theorem 2, the parametric form of the control law satisfying (16) is

$$u_k = -K_+ A_+^k \alpha + \bar{K}_+ (A_+^\top)^{k_f - k - 1} \beta, \quad \text{with} \quad k \in [0,\, k_f).$$

Hence, the thesis follows by virtue of (31). ∎

Corollary 4. *The optimal cost for the finite-horizon optimal control problem defined by (10)–(13), with known initial state x_O, under assumptions $\mathcal{A}1$–$\mathcal{A}5$, is*

$$J^O = x_O^\top S_O x_O + 2 \sigma_O^\top x_O + \rho_O,$$

where

$$\begin{aligned}S_O &= P_+ \alpha_1 + (P_+ W - I)(A_+^\top)^{k_f} \beta_1,\\ \sigma_O &= \frac{1}{2}\left(X_{k_f}^\top + P_+ \alpha_2 + (P_+ W - I)(A_+^\top)^{k_f} \beta_2\right) \sigma_f,\\ \rho_O &= \sigma_f^\top \overline{\mathrm{x}}_{k_f} + \rho_f.\end{aligned}$$

with X_{k_f} and $\overline{\mathrm{x}}_{k_f}$ defined according to Corollary 2.

Proof. First, note that

$$\begin{aligned}
J^O &= \sum_{k=0}^{k_f-1} \left(x_k^\top Q x_k + 2x_k^\top S u_k + u_k^\top R u_k\right) + \varPhi(x_{k_f}) \\
&= \sum_{k=0}^{k_f-1} \left(x_k^\top p_k - x_{k+1}^\top p_{k+1}\right) + x_{k_f}^\top S_f x_{k_f} + 2\sigma_f^\top x_{k_f} + \rho_f \\
&= x_O^\top p_0 - x_{k_f}^\top p_{k_f} + x_{k_f}^\top (S_f x_{k_f} + \sigma_f) + \sigma_f^\top x_{k_f} + \rho_f \\
&= x_O^\top p_0 + \sigma_f^\top x_{k_f} + \rho_f .
\end{aligned}$$

Moreover, from (28) and (31) it follows that

$$\begin{aligned}
p_0 &= P_+ \alpha + (P_+ W - I)(A_+^\top)^{k_f} \beta \\
&= P_0 x_O + \overline{p}_0 ,
\end{aligned}$$

with

$$\begin{aligned}
P_0 &= P_+ \alpha_1 + (P_+ W - I)(A_+^\top)^{k_f} \beta_1 , \\
\overline{p}_0 &= \left(P_+ \alpha_2 + (P_+ W - I)(A_+^\top)^{k_f} \beta_2\right) \sigma_f .
\end{aligned}$$

Furthermore, from Corollary 2, it follows that

$$x_{k_f} = X_{k_f} x_O + \overline{x}_{k_f} ,$$

where

$$X_{k_f} = A_+^{k_f} \alpha_1 + W \beta_1 , \quad \overline{x}_{k_f} = \left(A_+^{k_f} \alpha_2 + W \beta_2\right) \sigma_f .$$

Hence, the optimal cost can be expressed as

$$\begin{aligned}
J^O &= x_O^\top (P_0 x_O + \overline{p}_0) + \sigma_f^\top (X_{k_f} x_O + \overline{x}_{k_f}) + \rho_f \\
&= x_O^\top P_0 x_O + \left(\sigma_f^\top X_{k_f} + \overline{p}_0^\top\right) x_O + \sigma_f^\top \overline{x}_{k_f} + \rho_f \\
&= x_O^\top S_O x_O + 2\sigma_O^\top x_O + \rho_O ,
\end{aligned}$$

where the definitions of S_O, σ_O, ρ_O have been taken into account. ■

References

1. B. A. Francis and W. M. Wonham. (1976). *Automatica*, 12:457–463.
2. B. A. Francis. (1977) *SIAM Journal on Control and Optimization*, 15(3):486–505.
3. G. Marro and A. Piazzi. (1993) In *Proceedings of the 12th Triennial World Congress of the International Federation of Automatic Control*, volume 4, pages 23–26, Sydney, Australia, July 18–23.

4. G. Marro and E. Zattoni. (2007) In *Proceedings of the 2007 American Control Conference*, pages 5170–5175, New York, NY, July 11–13.
5. E. Zattoni. (2007) *IEEE Transactions on Automatic Control*, 52(1):140–143.
6. G. Marro, D. Prattichizzo, and E. Zattoni. (2006) *IEEE Transactions on Automatic Control*, 51(5):809–813.
7. G. Marro and E. Zattoni. (2005) *Automatica*, 41(5):815–821.
8. G. Marro. (1996) In C. Bonivento, G. Marro, and R. Zanasi, editors, *Colloquium on Automatic Control*, volume 215 of *Lecture Notes in Control and Information Sciences*, pages 77–138. Springer, Berlin / Heidelberg.
9. G. Marro and E. Zattoni. (2007) In *Proceedings of the 2007 American Control Conference*, New York, NY, July 5153–5157.
10. G. Basile and G. Marro. (1992) *Controlled and Conditioned Invariants in Linear System Theory.* Prentice Hall, Englewood Cliffs, New Jersey.
11. A. Saberi, P. Sannuti, and B. M. Chen. (1995) *H_2 Optimal Control.* Systems and Control Engineering. Prentice Hall International, London.
12. V. Ionescu, C. Oară, and M. Weiss. (1999) *Generalized Riccati Theory and Robust Control.* John Wiley and Sons, Chichester, England.
13. H. L. Trentelman, A. A. Stoorvogel, and M. Hautus. (2001) *Control Theory for Linear Systems.* Communications and Control Engineering. Springer-Verlag, London.
14. B. M. Chen, Z. Lin, and Y. Shamash. (2004) *Linear Systems Theory - A Structural Decomposition Approach.* Control Engineering. Birkhäuser, Boston.
15. F. L. Lewis and V. L. Syrmos. (1995) *Optimal Control.* John Wiley & Sons, New York, 2 edition.