Lecture Notes
in Control and Information Sciences 286

Editors: M. Thoma · M. Morari

Springer

Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Engineering ONLINE LIBRARY

http://www.springer.de/engine/

Anders Rantzer, Christopher I. Byrnes (Eds.)

Directions in Mathematical Systems Theory and Optimization

With 39 Figures

Springer

Series Advisory Board

A. Bensoussan · P. Fleming · M.J. Grimble · P. Kokotovic ·
A.B. Kurzhanski · H. Kwakernaak · J.N. Tsitsiklis

Editors

Anders Rantzer
Department of Automatic Control
Lund Institute of Technology
Box 118
SE-221 00 Lund
Sweden

Christopher I. Byrnes
School of Engineering and Applied Science
Washington University
One Brookings Drive
St. Louis, MO 63130
USA

ISSN 0170-8643

ISBN 3-540-00065-8 Springer-Verlag Berlin Heidelberg New York

Cataloging-in-Publication Data applied for
Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in other ways, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution act under German Copyright Law.

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science + Business Media GmbH

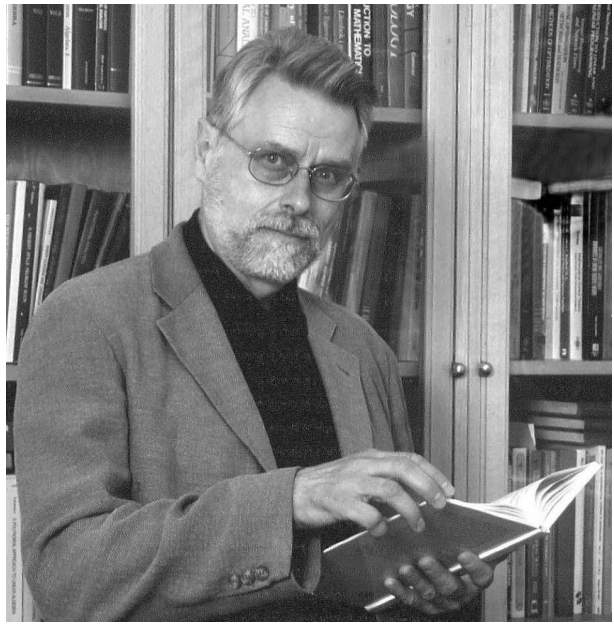
http://www.springer.de

© Springer-Verlag Berlin Heidelberg 2003
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Digital data supplied by author. Data-conversion by PTP-Berlin, Stefan Sossna e.K.
Cover-Design: design & production GmbH, Heidelberg
Printed on acid-free paper SPIN 10897638 62/3020Rw - 5 4 3 2 1 0

To Anders Lindquist
on the occasion of his sixtieth birthday



Preface

For more than three decades, Anders Lindquist has delivered fundamental contributions to the fields of systems, signals and control. Throughout this period, four themes can perhaps characterize his interests: Modeling, estimation and filtering, feedback and robust control.

His contributions to modeling include seminal work on the role of splitting subspaces in stochastic realization theory, on the partial realization problem for both deterministic and stochastic systems, on the solution of the rational covariance extension problem and on system identification. His contributions to filtering and estimation include the development of fast filtering algorithms, leading to a nonlinear dynamical system which computes spectral factors in its steady state, and which provide an alternate, linear in the dimension of the state space, to computing the Kalman gain from a matrix Riccati equation. His further research on the phase portrait of this dynamical system gave a better understanding of when the Kalman filter will converge, answering an open question raised by Kalman.

While still a student he established the separation principle for stochastic function differential equations, including some fundamental work on optimal control for stochastic systems with time lags. He continued his interest in feedback control by deriving optimal and robust control feedback laws for suppressing the effects of harmonic disturbances. Moreover, his recent work on a complete parameterization of all rational solutions to the Nevanlinna-Pick problem is providing a new approach to robust control design.

The editors join with the authors of the research articles in this book in congratulating Anders Lindquist on the occasion of his sixtieth birthday and on the continuation of a prolific career.

Christopher I. Byrnes *Anders Rantzer*

Acknowledgement: The book editors are deeply grateful to the publisher and all contributing authors for their efforts in meeting the tight time schedules of this project. Moreover, Leif Andersson provided outstanding technical support in preparation of the final manuscript. Thank you all!

Contents

1

Systems with Lebesgue Sampling

Karl Johan Åström *Bo Bernhardsson*

Abstract

Sampling is normally done periodically in time. For linear time invariant systems this leads to closed loop systems that linear and periodic. Many properties can be investigated by considering the behavior of the systems at times that are synchronized with the sampling instants. This leads to drastic simplifications because the systems can be described by difference equations with constant coefficients. This is the standard approach used today when designing digital controllers. Using an analog from integration theory, periodic sampling can also be called Riemann sampling . Lebesgue sampling or event based sampling, is an alternative to Riemann sampling, it means that signals are sampled only when measurements pass certain limits. This type of sampling is natural when using many digital sensors such as encoders. Systems with Lebesgue sampling are much harder to analyze than systems with Riemann sampling, because the time varying nature of the closed loop system can not be avoided. In this paper we investigate some systems with Lebesgue sampling. Analysis of simple systems shows that Lebesgue sampling gives better performance than Riemann sampling.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 1-13, 2003.
© Springer-Verlag Berlin Heidelberg 2003

1.1 Introduction

The traditional way to design digital control systems is to sample the signals equidistant in time, see [4]. A nice feature of this approach is that analysis and design becomes very simple. For linear time-invariant processes the closed loop system become linear and periodic. It is often sufficient to describe the behavior of the the closed loop system at times synchronized with the the sampling instants. This can be described by difference equations with constant coefficients.

There are several alternatives to periodic sampling. One possibility is to sample the system when the output has changed with a specified amount. Such a scheme has many conceptual advantages. Control is not executed unless it is required, control by exception, see [14]. This type of sampling is natural when using many digital sensors such as encoders. A disadvantage is that analysis and design are complicated. This type of sampling can be called Lebesgue sampling. Referring to to integration theory in mathematics we can also call conventional sampling Riemann sampling and Lebesgue sampling Lebesgue sampling. Much work on systems of this type was done in the period 1960-1980, but they have not received much attention lately. Some of the early work is reviewed in Section 2. In Section 3 we will analyze a simple example of Lebesgue sampling. The system considered is a first order system with random disturbances. It can be viewed as a simple model for an accelerator with pulse feedback. In this case it is possible to formulate and solve sensible control problems, which makes it possible to compare Riemann and Lebesgue sampling. The control strategy is very simple, it just resets the state with a given control pulse whenever the output exceeds the limits. The analysis indicates clearly that there are situations where it is advantageous with Lebesgue sampling. The mathematics used to deal with the problem is based on classical results on diffusion processes, [9], [10], [11]. An interesting conclusion is that the steady state probability distribution of the control error is non-Gaussian even if the disturbances are Gaussian. There are many interesting extensions of the problem discussed in the paper. Extensions to systems of higher order and output feedback are examples of natural extensions. An interesting property of systems with Lebesgue sampling is that the control strategy is an interesting mix of feedback and feed-forward control that often occurs in biological systems, see [13].

1.2 Examples of Systems with Lebesgue Sampling

Because of their simplicity Lebesgue sampling was used in many of early feedback systems. An accelerometer with pulse feedback is a typical example, see [8]. A pendulum was provided with pulse generators that moved the pendulum towards the center position as soon as a deviation was detected. Since all correcting impulses had the same form the velocity could be obtained simply by adding pulses.

Lebesgue sampling occurs naturally in many context. A common case is in motion control where angles and positions are sensed by encoders that give a pulse whenever a position or an angle has changed by a specific amount. Lebesgue sampling is also a natural approach when actuators with on-off characteristic are used. Satellite control by thrusters is a typical example, [7]. Systems with pulse

frequency modulation, [21], [19], [18], [24], [16], [25], [12], [22] and [23] are other examples. In this case the control signal is restricted to be a positive or negative pulse of given size. The control actions decide when the pulses should be applied and what sign they should have. Other examples are analog or real neurons whose outputs are pulse trains, see [15] and [6].

Systems with relay feedback are yet other examples which can be regarded as special cases of Lebesgue sampling, see [26], [27] and [3]. The sigma delta modulator or the one-bit AD converter, [17], which is commonly used in audio and mobile telephone system is one example. It is interesting to note that in spite of their wide spread there does not exist a good theory for design of systems with sigma delta modulators.

Traditionally systems with Lebesgue sampling were implemented as analog systems. Today they are commonly implemented as digital systems with fast sampling. Apart from their intrinsic interest systems with Lebesgue sampling may also be an alternative way to deal with systems with multi-rate sampling, see [1].

Analysis of systems with Lebesgue sampling are related to general work on discontinuous systems, [28], [29], [27] and to work on impulse control, see [5]. It is also relevant in situations where control complexity has to be weighted against execution time. It also raises other issues such as complexity of control. Control of production processes with buffers is another application area. It is highly desirable to run the processes at constant rates and make as few changes as possible to make sure that buffers are not empty and do not overflow, see [20]. Another example is where limited communication resources put hard restrictions on the number of measurement and control actions that can be transmitted.

Design Issues

All sampled systems run open loop between the sampling instants. For systems with Riemann sampling the inter-sample behavior is very simple. The control signal is typically constant or affine. Systems with Lebesgue sampling can have a much richer behavior which offers interesting possibilities. To illustrate this we will consider a regulation problem where it is desired to keep the state of the system in a given region in the state space that contains the origin. When the state leaves that region a control signal is generated. This open loop control signal is typically such that the state will reach the origin and stay there. The control signal required to do this can be quite complicated. Under ideal circumstances of no disturbances the state will then reach the origin and stay there until new disturbances occur. There are thus two issues in the design, to find a suitable switching boundary and to specify the behavior of the control signal after the switch.

The design of a system with Lebesgue sampling involves selection of an appropriate region and design of the control signal to be used when the state leaves the region. In a regulation problem it is also possible to use several different control regions, one can deal with normal disturbances and another larger region can deal with extreme disturbances.

1.3 A Simple Example

We will first consider a simple case where all calculations can be performed analytically. For this purpose it is assumed that the system to be controlled is described by the equation

$$dx = udt + dv,$$

where the disturbance $v(t)$ is a Wiener process with unit incremental variance and u the control signal. The problem of controlling the system so that the state is close to the origin will be discussed. Conventional periodic sampling will be compared with Lebesgue sampling where control actions are taken only when the output is outside the interval $-d < x < d$. We will compare the distribution of $x(t)$ and the variances of the outputs for both sampling schemes.

Periodic Sampling

First consider the case of periodic sampling with period h. The output variance is then minimized by the minimum variance controller, see [2]. The sampled system becomes

$$x(t+h) = x(t) + u(t) + e(t)$$

and the mean variance over one sampling period is

$$\begin{aligned} V &= \frac{1}{h}\int_0^h Ex^2(t)\,dt = \frac{1}{h}J_e(h) \\ &+ \frac{1}{h}\left(Ex^T Q_1(h)x + 2x^T Q_{12}(h)u + u^T Q_2(h)u\right) \\ &= \frac{1}{h}(R_1(h)S(h) + J_e(h)), \end{aligned} \tag{1.1}$$

where $Q_1(h) = h$, $Q_{12}(h) = h^2/2$, $Q_2(h) = h^3/3$, $R_1(h) = h$ and

$$J_e(h) = \int_0^h Q_{1c}\int_0^t R_{1c}\,d\tau\,dt = h^2/2. \tag{1.2}$$

The Riccati equation for the minimum variance strategy gives $S(h) = \sqrt{3}h/6$, and the control law becomes

$$u = -\frac{1}{h}\frac{3+\sqrt{3}}{2+\sqrt{3}}x$$

and the variance of the output is

$$V_R = \frac{3+\sqrt{3}}{6}h. \tag{1.3}$$

Lebesgue Sampling

In this case it is natural to choose the control region as an interval that contains the region. It is also easy to devise a suitable strategy. One possibility is to apply an impulse that drives the state of the system to the origin, another is to use an pulse of finite width. The case of impulse control is the simplest so we will start with that. Control actions are taken thus only taken when $|x(t_k)| = d$. When this

happens, an impulse control that makes $x(t_k+0) = 0$ is applied to the system. With this control law the closed loop system becomes a Markovian diffusion process of the type investigated in [10].

Let $T_{\pm d}$ denote the exit time i.e. the first time the process reaches the boundary $|x(t_k)| = d$ when it starts from the origin. The mean exit time can be computed from the fact that $t - x_t^2$ is a martingale between two impulses and hence

$$h_L := E(T_{\pm d}) = E(x^2_{T_{\pm d}}) = d^2.$$

The average sampling period thus equals $h_L = d^2$.

The stationary probability distribution of x is given by the stationary solution to the Kolmogorov forward equation for the Markov process, i.e.

$$0 = \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x) - \frac{1}{2}\frac{\partial f}{\partial x}(d)\delta_x + \frac{1}{2}\frac{\partial f}{\partial x}(-d)\delta_x.$$

with $f(-d) = f(d) = 0$ This ordinary differential equation has the solution

$$f(x) = (d - |x|)/d^2 \tag{1.4}$$

The distribution is thus symmetric and triangular with the support $-d \leq x \leq d$. The steady state variance is

$$V_L = \frac{d^2}{6} = \frac{h_L}{6}.$$

Comparison

To compare the results obtained with the different sampling schemes it is natural to assume that the average sampling rates are the same in both cases, i.e. $h_L = h$. This implies that $d^2 = h$ and it follows from equations (1.3) and (1.3) that

$$\frac{V_R}{V_L} = 3 + \sqrt{3} = 4.7.$$

Another way to say this is that one must sample 4.7 times faster with Riemann sampling to get the same mean error variance.

Notice that we have compared Lebesgue sampling with impulse control with periodic sampling with conventional sampling and hold. A natural question is if the improvement is due to the impulse nature of control or to the sampling scheme. To get some insight into this we observe that periodic sampling with impulse control gives and error which is a Wiener process which is periodically reset to zero. The average variance of such a process is

$$V_R' = Ex^2 = \frac{1}{h}E\int_0^h e^2(t)dt = \frac{1}{h}\int_0^h tdt = \frac{h}{2} \tag{1.5}$$

Periodic sampling with impulse control thus gives

$$\frac{V_R'}{V_L} = 3 \tag{1.6}$$

The major part of the improvement is thus due to the sampling scheme.

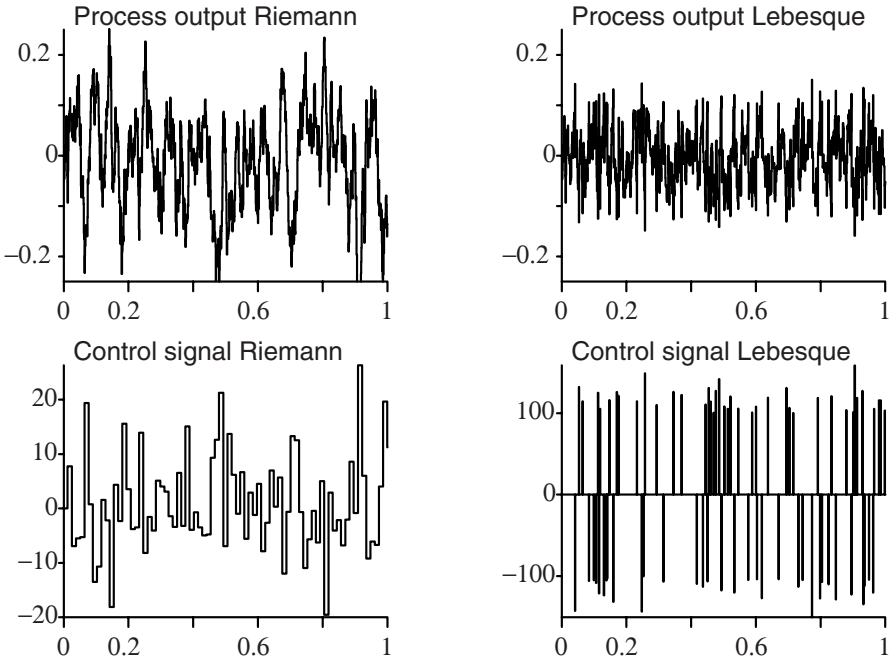


Figure 1.1 Simulation of an integrator with Riemann (left) and Lebesgue sampling (right).

Approximate Lebesgue Sampling

In the analysis it has been assumed that sampling is instantaneous. It is perhaps more realistic to assume that that sampling is made at a high fast rate but that no control action is taken if $x(t) < d$. The variance then becomes

$$V_{AL} = d^2(\frac{1}{6} + \frac{5}{6}\frac{h_a}{h_a + d^2}).$$

The second term is negligible when $h_a << d^2 = h_L$. Approximate Lebesgue sampling is hence good as long as d is relatively large.

The results are illustrated with the simulation in Figure 1.1. The simulation was made by rapid sampling (h=0.001). The parameter values used were $d = 0.1$, $h_R = 0.012$ and $\sigma_e = \sqrt{d}$. In the particular realization shown in the Figure there were 83 switches with Riemann sampling and 73 switches with Lebesgue sampling. Notice also the clearly visible decrease in output variance.

1.4 A First Order System

Consider now the first order system

$$dx = axdt + udt + dv. \tag{1.7}$$

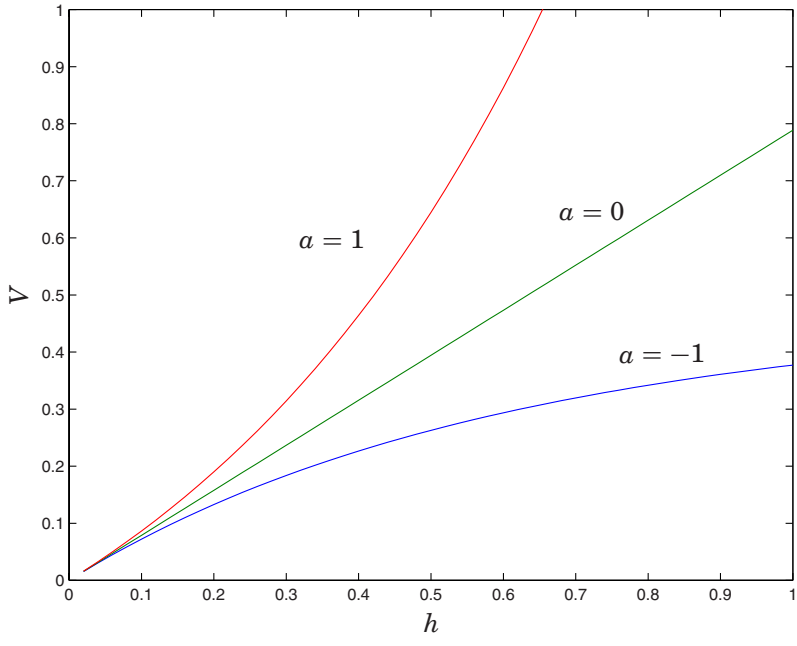


Figure 1.2 Variance $V_R(h)$ as a function of sampling time for $a = -1, 0, 1$ for a system with Riemann sampling.

Periodic Sampling

Sampling the system (1.7) with period h gives

$$x(t+h) = e^{ah}x(t) + \frac{1}{a}(e^{ah} - 1)u(t) + e(t) \tag{1.8}$$

where the variance of e is given by

$$J_e(h) = \int_0^h \int_0^t e^{2a\tau} d\tau dt = \left(\frac{e^{2ah} - 1}{2a}\right)^2 = R_1 \tag{1.9}$$

The sampled loss function is characterized by

$$Q_1 = \frac{e^{2ah} - 1}{2a}$$
$$Q_{12} = \frac{e^{ah}ah - e^{ah} + 1}{a^2}$$
$$Q_2 = \frac{h^3}{3}$$

The minimum variance control law is obtained by solving a Riccati equation for $S(h)$. The formula which is complicated is omitted. The variance of the output is shown in Figure 1.2 for different values of the parameter a. Notice that the increase of the variance with the sampling period increases much faster for unstable systems $a > 0$.

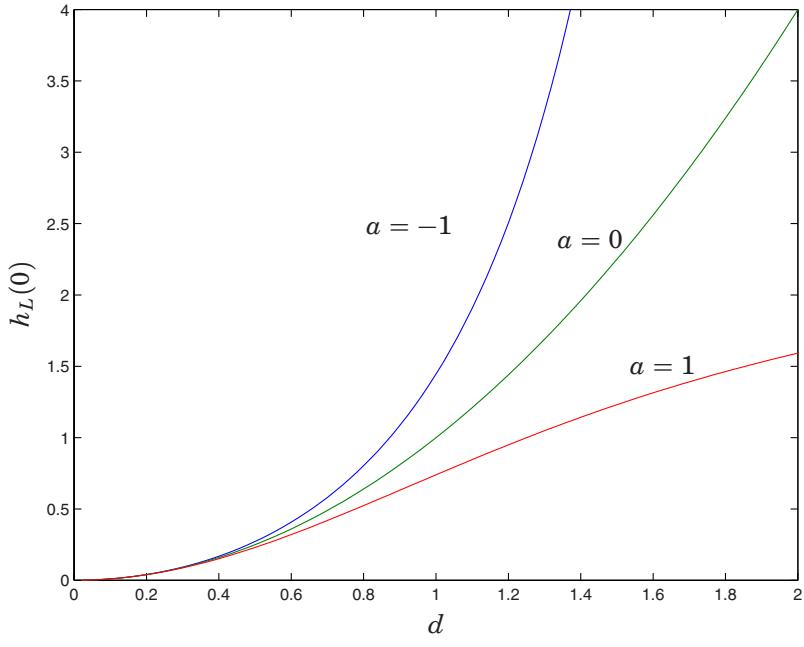


Figure 1.3 Mean exit time $E(T_{\pm d}) = h_L(0)$ as a function of d for $a = -1, 0, 1$ for a system with Lebesgue sampling.

Lebesgue Sampling

For Lebesgue sampling we assume as in Section 2 that the variable x is reset to zero when $|x(t_k)| = d$. The closed loop system obtained is then a diffusion process. The average sampling period is the mean exit time when the process starts at $x = 0$. This can be computed from the following result in [10].

THEOREM 1.1
Consider the differential equation $dx = b(x)dt + \sigma(x)dv$ and introduce the backward Kolmogorov operator

$$(\mathcal{A}h)(x) \triangleq \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}(x)\frac{\partial^2 h(x)}{\partial x_i \partial x_k} + \sum_{i=1}^{n} b_i(x)\frac{\partial h(x)}{\partial x_i}, \tag{1.10}$$

where $h \in C^2(R^n)$ and $a_{ik} = [\sigma\sigma^T]_{ik}$. The mean exit time from $[-d, d]$, starting in x, is given by $h_L(x)$, where

$$\mathcal{A}h_L = -1$$

with $h_L(d) = h_L(-d) = 0$. □

In our case the Kolmogorov backward equation becomes

$$\frac{1}{2}\frac{\partial^2 h_L}{\partial x^2} + ax\frac{\partial h_L}{\partial x} = -1$$

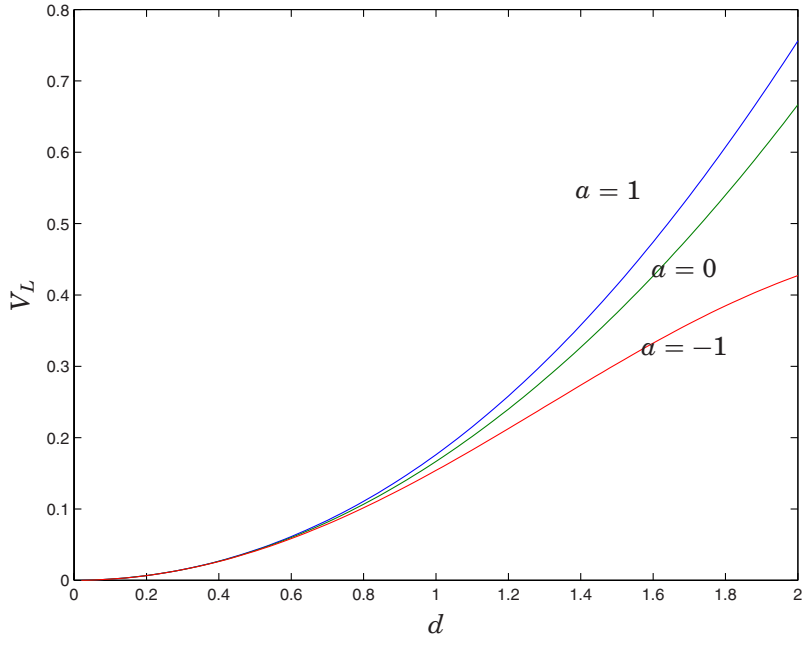


Figure 1.4 Variance as a function of level d for $a = -1, 0, 1$, for a system with Lebesgue sampling.

with $h_L(d) = h_L(-d) = 0$. The solution is given by

$$h_L(x) = 2\int_x^d \int_0^y e^{-a(y^2-t^2)}\, dt\, dy,$$

which gives

$$\begin{aligned} h_L(0) &= \sum_{k=1}^{\infty} 2^{2k-1}(-a)^{k-1}(k-1)!\, d^{2k}/(2k)! \\ &= d^2 - \frac{a}{3}d^4 + \frac{4a^2}{45}d^6 + O\left(d^8\right)). \end{aligned}$$

Figure 1.4 shows $h_L(0)$ for $a = -1, 0, 1$.

The stationary distribution of x is given by the forward Kolmogorov equation

$$\begin{aligned} 0 = \frac{\partial}{\partial x}\left(\frac{1}{2}\frac{\partial f}{\partial x} - axf\right) - \left(\frac{1}{2}\frac{\partial f}{\partial x} - axf\right)_{x=d}\delta_x \\ + \left(\frac{1}{2}\frac{\partial f}{\partial x} - axf\right)_{x=-d}\delta_x. \end{aligned} \tag{1.11}$$

To solve this equation we observe that the equation

$$0 = \frac{\partial}{\partial x}\left(\frac{1}{2}\frac{\partial f}{\partial x} - axf\right) \tag{1.12}$$

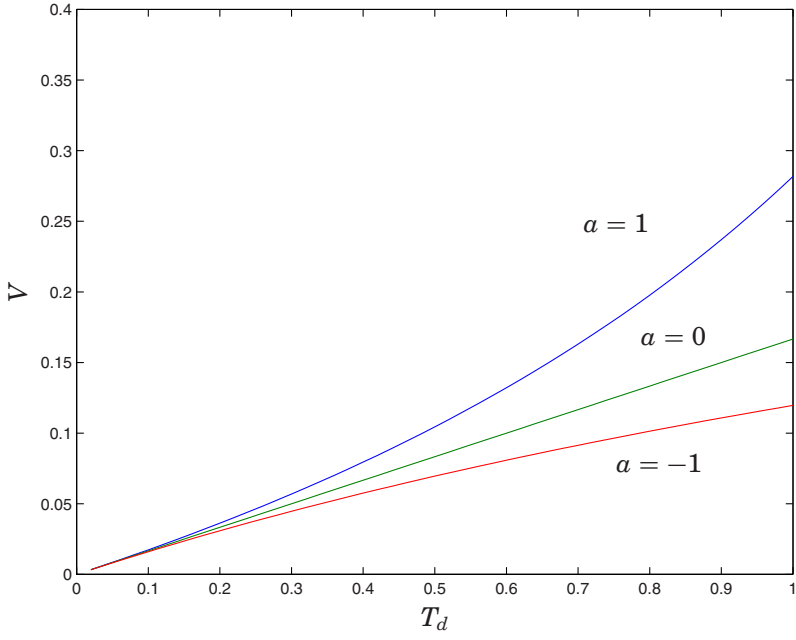


Figure 1.5 Variance as a function of mean exit time $T_{\pm d}$ for $a = -1, 0, 1$, for a system with Lebesgue sampling.

has the solutions

$$f(x) = c_1 e^{ax^2} + c_2 \int_0^x e^{a(x^2-t^2)} dt.$$

The even function

$$f(x) = c_1 e^{ax^2} + c_2 \operatorname{sign}(x) \int_0^x e^{a(x^2-t^2)} dt,$$

then satisfies (1.11) also at $x = 0$. The constants c_1, c_2 are determined by the equations

$$\int_{-d}^{d} f(x)dx = 1, \tag{1.13}$$

$$f(d) = 0, \tag{1.14}$$

which gives a linear equation system to determine c_1, c_2.

Having obtained the stationary distribution of x we can now compute the variance of the output

$$V_L = \int_{-d}^{d} x^2 f(x) dx.$$

The variance V_L is plotted as a function of d in Figure 1.4 for $a = -1, 0, 1$, and as a function of mean exit time h_L in Figure 1.5.

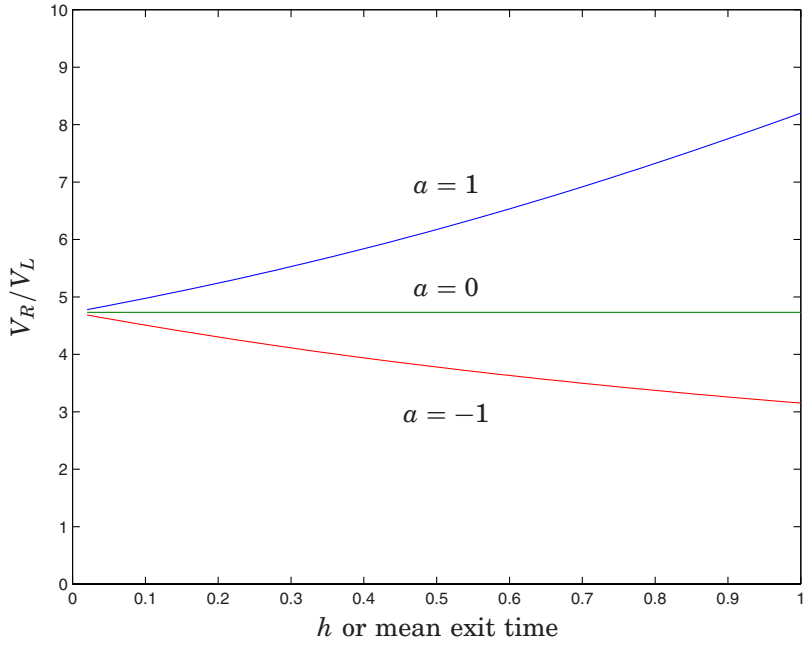


Figure 1.6 Comparison of V_L and V_R for $a = -1, 0, 1$. Note that the performance gain of using Lebesgue sampling is larger for unstable systems with slow sampling.

Comparison

The ratio V_R/V_L as a function of h is plotted in Figure 1.6 for $a = -1, 0, 1$. The figure shows that Lebesgue sampling gives substantially smaller variances for the same average sampling rates. For short sampling periods there are small differences between stable and unstable system as can be expected. The improvement of Lebesgue sampling is larger for unstable systems and large sampling periods.

Note that the results for other a can be obtained from these plots since the transformation $(x, t, a, v) \to (\alpha^{1/2}x, \alpha t, \alpha^{-1}a, \alpha^{1/2}v)$ for $\alpha > 0$ leaves the problem invariant.

1.5 Conclusions

There are issues in Lebesgue sampling that are of interest to explore. The signal representation which is a mixture of analog and discrete is interesting. It would be very attractive to have a system theory similar to the one for periodic sampling. The simple problems solved in this paper indicate that Lebesgue sampling may be worth while to pursue. Particularly since many sensors that are commonly used today have this character. Implementation of controller of the type discussed in this paper can be made using programmable logic arrays without any need for AD and DA converters. There are many generalizations of the specific problems discussed in this paper that are worthy of further studies for example higher order systems and systems with output feedback.

1.6 References

[1] M Araki and K Yamamoto. Multivariable multirate sampled-data systems: state space description, transfer characteristics and nyquist criterion. *IEEE Transactions on Automatic Control*, AC-31:145–154, 1986.

[2] Karl Johan Åström. *Introduction to Stochastic Control Theory*. Academic Press, New York, 1970.

[3] Karl Johan Åström. Oscillations in systems with relay feedback. In Karl Johan Åström, G. C. Goodwin, and P. R. Kumar, editors, *Adaptive Control, Filtering, and Signal Processing*, volume 74 of *IMA Volumes in Mathematics and its Applications*, pages 1–25. Springer-Verlag, 1995.

[4] Karl Johan Åström and Björn Wittenmark. *Computer-Controlled Systems*. Prentice Hall, third edition, 1997.

[5] Alain Bensoussan and Jacques-Louis Lions. *Impulse control and quasi-variational inequalities*. Gauthier-Villars, Paris, 1984.

[6] S. DeWeerth, Lars Nielsen, C. Mead, and Karl Johan Åström. A neuron-based pulse servo for motion control. In *IEEE Int. Conference on Robotics and Automation*, Cincinnati, Ohio, 1990.

[7] S. J. Dodds. Adaptive, high precision, satellite attitude control for microprocessor implementation. *Automatica*, 17(4):563–573, 1981.

[8] S. S. Draper, W. Wrigley, and J. Hovorka. *Inertial Guidance*. Pergamon Press, Oxford, 1960.

[9] Wilhelm Feller. The parabolic differential equations and the associated semi-groups of transformations. *Ann. of Math.*, 55:468–519, 1952.

[10] Wilhelm Feller. Diffusion processes in one dimension. *Trans. Am. Math. Soc.*, 55:1–31, 1954.

[11] Wilhelm Feller. The general diffusion operator and positivity preserving semi-groups in one dimension. *Ann. of Math.*, 60:417–436, 1954.

[12] Paul M. Frank. A continuous-time model for a pfm-controller. *IEEE Transactions on Automatic Control*, AC-25(5):782–784, 1979.

[13] Russel K. Hobbie. *Intermediate Physics for Medicine and Biology*. Springer, 1997.

[14] H Kopetz. Should responsive systems be event triggered or time triggered? *IEICE Trans. on Information and Systems*, E76-D(10):1525–1532, 1993.

[15] C. A. Mead. *Analog VLSI and Neural Systems*. Addison-Wesley, Reading, Massachusetts, 1989.

[16] E. Noges and P. M. Frank. *Pulsfrequenzmodulierte Regelungssysteme*. R. Oldenbourg, München, 1975.

[17] S. R. Norsworthy, R. Schreier, and G. Temes. *Delta-Sigma Data Converters*. IEEE Press, New York, 1996.

[18] T. Pavlidis. Optimal control of pulse frequency modulated systems. *IEEE Transactions on Automatic Control*, AC-11(4):35–43, 1966.

[19] T Pavlidis and E. J Jury. Analysis of a new class of pulse frequency modulated control systems. *IEEE Transactions on Automatic Control*, AC-10:35–43, 1965.

[20] Bengt Pettersson. Production control of a complex integrated pulp and paper mill. *Tappi*, 52(11):2155–2159, 1969.

[21] E Polak. Stability and graphical analysis of first order of pulse-width-modulated sampled-data regulator systems. *IRE Trans. Automatic Control*, AC-6(3):276–282, 1968.

[22] Herbertt Sira-Ramirez. A geometric approach to pulse-width modulated control in nonlinear dynamical systems. *IEEE Transactions on Automatic Control*, AC-34(2):184–187, 1989.

[23] Herbertt Sira-Ramirez and Pablo Lischinsky-Arenas. Dynamic discontinuous feedback control of nonlinear systems. *IEEE Transactions on Automatic Control*, AC-35(12):1373–1378, 1990.

[24] Ronald A. Skoog. On the stability of pulse-width-modulated feedback systems. *IEEE Transactions on Automatic Control*, AC-13(5):532–538, 1968.

[25] Ronald A. Skoog and Gilmer L. Blankenship. Generalized pulse-modulated feedback systems: Norms, gains, lipschitz constants and stability. *IEEE Transactions on Automatic Control*, AC-15(3):300–315, 1970.

[26] Ya. Z. Tsypkin. Theory of intermittent control. *Avtomatika i Telemekhanika*, 1949, 1950. Vol. **10** (1949) Part I, pp. 189–224; Part II, pp. 342–361; Vol. **11** (1950) Part III, pp. 300–331.

[27] Ya. Z. Tsypkin. *Relay Control Systems*. Cambridge University Press, Cambridge, UK, 1984.

[28] V. Utkin. *Sliding modes and their applications in variable structure systems*. MIR, Moscow, 1981.

[29] V. I. Utkin. Discontinuous control systems: State of the art in theory and applications. In *Preprints 10th IFAC World Congress*, Munich, Germany, 1987.

2

Acoustic Attenuation Employing Variable Wall Admittance

H. T. Banks *K. M. Furati* *K. Ito* *N. S. Luke*
C. J. Smith

Abstract

In this discussion we present results for sound attenuation in a half plane using arrays of micro-acoustic actuators on a control surface. We explain and use a computational method developed in a previous paper [2]. Specifically, we carry out computations that document the sensitivity of the overall energy dissipation as a function of wall admittance.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 15-26, 2003.
© Springer-Verlag Berlin Heidelberg 2003

2.1 Introduction

A recent topic of much interest focuses on the use of arrays of small acoustic sources, which could be constructed of fluidic devices [1] or piezoelectric materials, for sound absorption. In such an array, each element would be locally reacting by feedback of a pressure measurement via an appropriate compensator to the elemental acoustic source.

In this paper we report on the use of a computational technique developed in [2] that can be employed to analyze the absorption rate of acoustic arrays. We present computational findings that suggest enhanced energy dissipation via locally varying admittances on the boundary. These results are a continuous effort in the development of a conceptual framework for design, analysis, and implementation of *smart* or *adaptive* acoustic arrays to be used in noise control in structural systems.

2.2 Problem Formulation

In this paper we consider a typical physical scenario for sound absorption or reflection on a treated planar wall. The sound pressure p satisfies the acoustic wave equation

$$p_{\tau\tau} - c^2\nabla^2 p = 0, \tag{2.1}$$

and pressure and velocity are related through Euler's equation

$$\rho \mathbf{v}_\tau = -\nabla p, \tag{2.2}$$

where c is the speed of sound in air and ρ is the mass density of air. The vector $\mathbf{v} = (u, v, w)$ is the particle velocity of air, τ denotes unscaled time and $\mathbf{v}_\tau = \partial \mathbf{v}/\partial\tau$; see [3], [5], and [6].

We consider this system in an infinite half plane bounded by an infinite rigid wall located at $x = 0$ treated with an acoustic array that is finite and symmetric about the origin, $y = 0$. We assume that the acoustic array is sufficiently long in the z-direction, making the system essentially independent of z and thus two dimensional, as depicted in Figure 2.1. We define the domain $\Omega = \{(x, y) : x \geq 0, y \in \Re\}$ for our problem.

The interaction of waves with the boundary is described by the boundary condition

$$\mathbf{v}(0, y, \tau) \cdot \vec{\imath} = u(0, y, \tau) = -g(y)p(0, y, \tau), \qquad -\infty < y < \infty, \tau > 0. \tag{2.3}$$

This boundary condition follows from the *normal input admittance*, which is defined as the ratio of the particle velocity normal to the wall to the pressure [4], where g is the interaction admittance. We observe that $g \equiv 0$ in the case of a rigid boundary, while $g \to \infty$ corresponds to a pressure release boundary, $p \equiv 0$.

Let ϕ be the velocity potential, $\mathbf{v} = \nabla\phi$. Then from (2.2) we have $p = -\rho\phi_\tau$ and one can readily argue that ϕ satisfies

$$\phi_{\tau\tau} - c^2\nabla^2\phi = 0 \qquad \text{in } \Omega, \tag{2.4}$$

$$\phi_x(0, y, \tau) = \rho\, g(y)\, \phi_\tau(0, y, \tau), \qquad -\infty < y < \infty, \tau > 0. \tag{2.5}$$

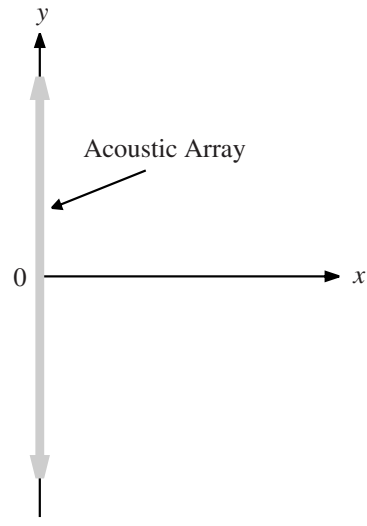


Figure 2.1 An acoustic array in half plane acoustic domain

We scale time by the sound speed, letting $t = c\tau$, and we define $\gamma(y) = \rho c\, g(y)$, so that (2.4) and (2.5) become

$$\phi_{tt} - \nabla^2\phi = 0 \qquad \text{in } \Omega \tag{2.6}$$

$$\phi_x(0, y, t) = \gamma(y)\, \phi_t(0, y, t), \qquad -\infty < y < \infty, t > 0. \tag{2.7}$$

The magnitude of the admittance in a given direction is a function of the wave propagation angle. Let $\mathbf{x}$ and $\mathbf{k}$ denote the position vector (x, y) and the wavenumber vector (k, l), respectively. Then for a plane wave, denoted $\phi(\mathbf{x}, t) = \phi_o\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{x})}$, the normalized wave admittance in the x-direction is

$$\beta = \rho c \frac{u}{p} = \frac{k}{\omega} = \frac{k}{|\mathbf{k}|}, \tag{2.8}$$

since $\omega = |\mathbf{k}|$. We note that $0 \le |\beta| \le 1$ with $|\beta| = 0$ corresponding to a plane wave traveling only in the y-direction and $|\beta| = 1$ corresponding to a plane wave traveling only in the x-direction. In order to absorb sound at the boundary we might choose the normal input admittance γ to match the normal wave admittance β. However, in general, the knowledge of the angle of incidence of the incoming wave is unknown. Moreover, for most fields the angle of incidence is arbitrary. For this reason, in our model we test the absorption characteristics for a given γ by the absorption rate of an ideal reverberant sound field under the assumption that the angle of incidence is uniformly distributed among all directions.

2.3 Frequency Domain and Approximation

To facilitate approximations and numerical computations, we reformulate equations (2.6) and (2.7) in the frequency domain. Since the spatial domain Ω is the

half plane $x > 0$, then without loss of generality, we use the Fourier cosine transform with respect to x and full Fourier transform with respect to y. Let $\hat{\phi}$ be the Fourier cosine/full transform of ϕ,

$$\hat{\phi}(\mathbf{k}, t) = \hat{\phi}(k, l, t) = \int_{-\infty}^{\infty} \int_{0}^{\infty} \phi(x, y, t) \cos(kx)\, e^{-ily}\, dx\, dy. \tag{2.9}$$

Then the inverse transform is given by

$$\phi(\mathbf{x}, t) = \phi(x, y, t) = \frac{1}{\pi^2} \int_{-\infty}^{\infty} \int_{0}^{\infty} \hat{\phi}(k, l, t) \cos(kx)\, e^{ily}\, dk\, dl. \tag{2.10}$$

We can readily derive an equation for $\hat{\phi}(k, l, t)$. As detailed in [2], we obtain

$$\hat{\phi}_{tt}(k, l, t) = -(k^2 + l^2)\hat{\phi}(k, l, t) - \int_{-\infty}^{\infty} \int_{0}^{\infty} \hat{\phi}_t(\alpha, \beta, t)\, \hat{\gamma}(l, \beta)\, d\alpha\, d\beta, \tag{2.11}$$

where

$$\hat{\gamma}(l, \beta) = \frac{1}{\pi^2} \int_{-\infty}^{\infty} \gamma(y) e^{-i(l-\beta)y}\, dy.$$

Since the acoustic array is assumed to be symmetric, $\gamma(y)$ is an even function and thus

$$\hat{\gamma}(l, \beta) = \frac{2}{\pi^2} \int_{0}^{\infty} \gamma(y) \cos[(l - \beta)y]\, dy = \tag{2.12}$$

$$\frac{2}{\pi^2} \int_{0}^{\infty} \gamma(y)[\cos(ly)\cos(\beta y) + \sin(ly)\sin(\beta y)]\, dy.$$

If we further assume that the initial field is an even function of y, then the map $y \to \phi(\cdot, y, \cdot)$ is even, and we only need to work with the nonnegative values of l. Moreover, the integrand $\gamma(y)\sin(ly)\sin(\beta y)$ in (2.12) is an odd function of β and thus does not contribute to the integral term of (2.11) since $\beta \to \hat{\phi}_t(\alpha, \beta, t)$ is even. As a result, equation (2.11) reduces to

$$\hat{\phi}_{tt}(k, l, t) = -(k^2 + l^2)\hat{\phi}(k, l, t) - \int_{0}^{\infty} \int_{0}^{\infty} \hat{\phi}_t(\alpha, \beta, t)\, \Gamma(l, \beta)\, d\alpha\, d\beta, \tag{2.13}$$

where

$$\Gamma(l, \beta) = \frac{4}{\pi^2} \int_{0}^{\infty} \gamma(y) \cos(ly) \cos(\beta y)\, dy. \tag{2.14}$$

We note that this yields a nonstandard integro-partial differential equation for $\hat{\phi}$ for which approximation techniques must be developed. We summarize a semi-discrete finite element approximation based on piecewise constant elements (zero-order splines) that was developed and tested numerically in [2].

For finite positive integers K and L, we consider the truncated finite domain in frequency space

$$\Omega_{k,l} = \{(k, l) : 0 \le k \le K, 0 \le l \le L\}. \tag{2.15}$$

We discretize (2.13) in this finite domain.

Let $0 = k_0 < k_1 < k_2 < \cdots < k_M = K$ and $0 = l_0 < l_1 < l_2 < \cdots < l_N = L$ be partitions along the k- and l-axis in the kl-plane. Let $\Delta k_i = k_i - k_{i-1}$, $\Delta l_j = l_j - l_{j-1}$, and $R_{ij} = (k_{i-1}, k_i] \times (l_{j-1}, l_j]$. Let $(\tilde{k}_i, \tilde{l}_j)$ denote the midpoint of each cell R_{ij}.

In equation (2.13), we apply the (piecewise constant in frequency) approximation

$$\hat{\phi}(k,l,t) \approx \Phi^{MN}(k,l,t) = \sum_{i=1}^{M}\sum_{j=1}^{N} \Phi_{ij}^{MN}(t)\chi_{ij}^{MN}(k,l). \tag{2.16}$$

where

$$\chi_{ij}^{MN}(k,l) = \begin{cases} 1 & (k,l) \in R_{ij}, \\ 0 & \text{otherwise}, \end{cases}$$

and $\Phi_{ij}^{MN}(t)$ is an approximation of $\hat{\phi}(\tilde{k}_i, \tilde{l}_j, t)$. Then, we evaluate (2.16) at the midpoint of each cell R_{ij}. This results in the system of equations of the form

$$\ddot{\Phi}_{ij}^{MN}(t) + \lambda_{ij}^2 \Phi_{ij}^{MN}(t) + \sum_{n=1}^{N}\left(g_{jn}^{MN} \sum_{m=1}^{M} \Delta k_m \dot{\Phi}_{mn}^{MN}(t)\right) = 0, \tag{2.17}$$

where

$$\lambda_{ij} = \sqrt{\tilde{k}_i^2 + \tilde{l}_j^2}, \qquad g_{jn}^{MN} = \int_{l_{n-1}}^{l_n} \Gamma(\tilde{l}_j, \beta) d\beta.$$

We remark that we use the superscript MN with the variables in (2.17) since they depend on the discretization in (k,l). For notational convenience, we shall suppress this superscript in the remainder of our discussions, whenever no confusion will result.

We can write the system (2.17) in matrix form by introducing the $\Re^{MN}$ vector

$$\Phi = [\Phi_{11}, \Phi_{21}, \ldots, \Phi_{M1}, \Phi_{12}, \Phi_{22}, \ldots, \Phi_{M2}, \ldots, \Phi_{1N}, \Phi_{2N}, \ldots, \Phi_{MN}]^T,$$

the $MN \times MN$ diagonal matrix

$$\Lambda = \operatorname{diag}(\lambda_{11}^2, \lambda_{21}^2, \ldots, \lambda_{M1}^2, \lambda_{12}^2, \lambda_{22}^2, \ldots, \lambda_{M2}^2, \ldots, \lambda_{1N}^2, \lambda_{2N}^2, \ldots, \lambda_{MN}^2),$$

and the damping matrix

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1N} \\ g_{21} & g_{22} & \cdots & g_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ g_{N1} & g_{N2} & \cdots & g_{NN} \end{bmatrix}.$$

We define the $M \times M$ matrix

$$\Delta_k = \begin{bmatrix} \Delta k_1 & \Delta k_2 & \cdots & \Delta k_M \\ \Delta k_1 & \Delta k_2 & \cdots & \Delta k_M \\ \vdots & \vdots & \cdots & \vdots \\ \Delta k_1 & \Delta k_2 & \cdots & \Delta k_M \end{bmatrix},$$

and let D be the Kronecker tensor product

$$D = \text{kron}(G, \Delta_k).$$

Then, the second order system (2.17) can be written in the matrix form

$$\ddot{\Phi}(t) + D\dot{\Phi}(t) + \Lambda\Phi(t) = 0, \tag{2.18}$$

or, as a first order system by defining

$$Z(t) = \left[\begin{array}{c} \Phi(t) \\ \hline \dot{\Phi}(t) \end{array}\right]$$

to obtain the equation

$$\dot{Z} = AZ. \tag{2.19}$$

Here, A is the $(2MN) \times (2MN)$ matrix defined by

$$A = \left[\begin{array}{c|c} 0 & I \\ \hline -\Lambda & -D \end{array}\right],$$

and I is the $(MN) \times (MN)$ identity matrix. As in [2], we use this approximate system in the calculations described here.

In all of our investigations, we assume that the initial sound field is random and the spatial autocorrelation is independent of the position and phase, being a function only of the distance between two points in space. Thus, the initial random field $\phi(\mathbf{x}, 0)$ satisfies

$$\mathcal{E}[\phi(\mathbf{x}, 0)\, \phi(\mathbf{x} + \mathbf{r}, 0)] = R(|\mathbf{r}|)$$

for all $\mathbf{x}$, where $\mathcal{E}[\cdot]$ denotes the expectation. In the frequency domain this is equivalent to

$$\mathcal{E}\left[\overline{\hat{\phi}(\mathbf{k}, 0)}\, \hat{\phi}(\mathbf{k}', 0)\right] = (2\pi)^2\, \delta(\mathbf{k} - \mathbf{k}') \int R(|\mathbf{r}|)\, e^{-i\mathbf{k}\cdot\mathbf{r}}\, d\mathbf{r}, \tag{2.20}$$

where δ is the Dirac delta function.

From this it follows that a natural condition to require on the approximating random Fourier components $\{Z_i(0)\}$ for the system (2.19) is

$$\mathcal{E}[Z_i(0)\, Z_j(0)] = 0, \quad \text{if } i \neq j. \tag{2.21}$$

If we then define the corresponding approximate correlation matrix

$$C(t) = \mathcal{E}[Z(t)\, Z^T(t)], \tag{2.22}$$

we see that (2.21) implies that $C(0)$ is diagonal. Thus, we can write

$$C(0) = \sum_{\nu=1}^{2MN} w_\nu\, e^{(\nu)} \left(e^{(\nu)}\right)^T, \tag{2.23}$$

where $e^{(\nu)}$ is the unit vector in $\Re^{2MN}$ and $W = (w_\nu)$ is the vector of the initial field discrete power spectral densities.

We then see that the $2MN \times 2MN$ matrix function $C(t)$ satisfies the matrix Lyapunov differential equation system

$$\dot{C}(t) = A\,C(t) + C(t)\,A^T, \tag{2.24}$$

which is a large system of coupled equations with dimension $(2MN)^2$.

The solution of (2.24) can be obtained in superposition form (see [2] for details)

$$C(t) = \sum_{\nu=1}^{2MN} w_\nu\, F^{(\nu)}(t) \left(F^{(\nu)}(t)\right)^T, \tag{2.25}$$

where the vector $F^{(\nu)}(t)$ for each ν satisfies

$$\dot{F}^{(\nu)}(t) = A\,F^{(\nu)}(t), \qquad F^{(\nu)}(0) = e^{(\nu)}. \tag{2.26}$$

When considering fields with compact support Fourier components (such as those with finite bandwidth), as done here, the superposition formula (2.25) coupled with (2.26) is especially efficient.

2.4 Instantaneous Total Energy and Damping

Since our primary interest is to study attenuation provided by the acoustic array as we change the local admittance (as manifested in the shape of γ in (2.7)) it is most useful to have a measure of the total energy in the field.

We can define an approximation to the total energy for (2.6)-(2.7) by employing the total energy of system (2.18)

$$E^{MN}(t) = E_V^{MN}(t) + E_P^{MN}(t), \tag{2.27}$$

where the "kinetic" energy and "potential" energy are given, respectively, by

$$E_V^{MN}(t) = \frac{1}{2}\int |\nabla \phi^{MN}(\mathbf{x},t)|^2\, d\mathbf{x}, \qquad E_P^{MN}(t) = \frac{1}{2}\int [\phi_t^{MN}(\mathbf{x},t)]^2\, d\mathbf{x},$$

and $\phi^{MN}(\mathbf{x},t)$ is the inverse Fourier transform of $\Phi^{MN}(\mathbf{k},t)$ defined in (2.16). By Parseval's theorem, we have

$$E_V(t) = \frac{1}{2}\Phi^T(t)\,\Lambda\,\Phi(t), \qquad E_P(t) = \frac{1}{2}\dot{\Phi}^T(t)\,\dot{\Phi}(t), \tag{2.28}$$

where again we shall suppress the MN superscripts.

Now, by multiplying (2.18) by $\dot{\Phi}$, defining $E(t) = E_V(t) + E_P(t)$, and using (2.28), we can argue that

$$\dot{E}(t) = -\dot{\Phi}^T(t)\,D\,\dot{\Phi}(t) \le 0. \tag{2.29}$$

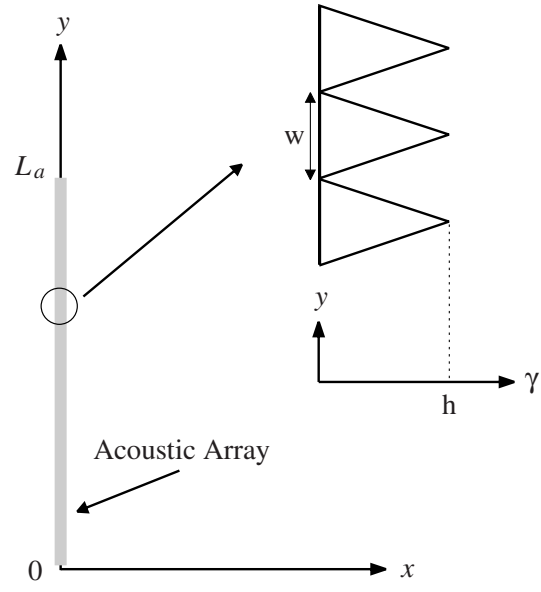


Figure 2.2 Acoustic array and the admittance function $\gamma(y) = \gamma_{(\mathrm{w,h})}(y)$.

It is immediately apparent from (2.29) that the system is dissipative for any positive γ.

One can further argue (see [2] for details) that the expected energy can be approximated by

$$\mathcal{E}[E(t)] = \frac{1}{2}\,\mathcal{E}\left[\Phi^T(t)\,\Lambda\,\Phi(t) + \dot{\Phi}^T(t)\,\dot{\Phi}(t)\right] = \frac{1}{2}\,\mathrm{trace}[\mathcal{L}\,C(t)\,\mathcal{L}]. \tag{2.30}$$

where the matrix $\mathcal{L}$ is defined by

$$\mathcal{L} = \left[\begin{array}{c|c} \sqrt{\Lambda} & 0 \\ \hline 0 & I \end{array}\right].$$

Thus for a given γ, we can integrate (2.26) to obtain $F^{(\nu)}(t)$ and use (2.23), (2.25), and (2.30) to determine the decay rate of a field composed of the group of modes determined by a given bandwidth.

2.5 Computational Results

We turn finally to computations for a family of acoustic arrays using the methodology outlined in the previous sections. We consider a family of periodic acoustic arrays of length $2L_a$ in the y-direction and symmetric about the origin as depicted in Figure 2.2.

The arrays consist of elements each of width w and maximum height h. Due to symmetry, we only need to consider the upper half ($y \geq 0$) of the arrays. The

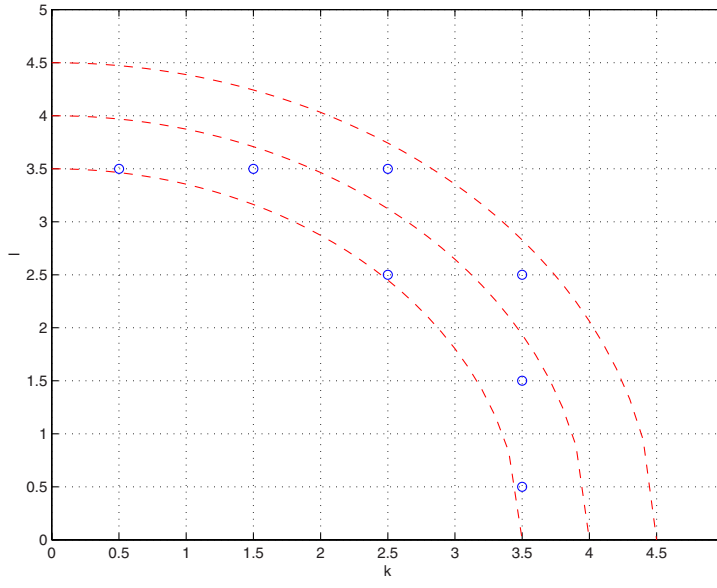


Figure 2.3 The set Υ of the initial discrete modes.

number of elements in a given half array is

$$N_e = L_a/\mathrm{w}.$$

For $n = 1, \ldots, N_e$, each nth element occupies the interval $I_n = [(n-1)\mathrm{w}, n\mathrm{w}]$.

In each element we take the function $\gamma(y)$ to be a symmetric triangle curve of height h as shown in Figure 2.2. Thus we have

$$\gamma(y) = \begin{cases} \gamma^n(y), & y \in I_n, \\ 0 & y > L_a, \end{cases} \tag{2.31}$$

where

$$\gamma^n(y) = \begin{cases} \frac{2\mathrm{h}}{\mathrm{w}} y - 2(n-1)\mathrm{h}, & y \in [(n-1)\mathrm{w}, (n-.5)\mathrm{w}], \\ -\frac{2\mathrm{h}}{\mathrm{w}} y + 2n\mathrm{h}, & y \in [(n-.5)\mathrm{w}, n\mathrm{w}]. \end{cases} \tag{2.32}$$

For our example calculations we used $L_a = 4$ meters along with various admittance widths w and heights h. Thus the family of admittance functions $\gamma = \gamma_{(\mathrm{w},\mathrm{h})}$ are parameterized by (w, h) as in (2.32).

Moreover we chose $K = L = 10$ and uniform partitions $\Delta k_i = \Delta l_j = 1$. This implies that $M = N = 10$ and that A in (2.19) has dimension 200×200. We calculated the entries of the matrix G by using the midpoint approximations

$$g_{jn} \approx \Delta l \, \Gamma(\tilde{l}_j, \tilde{l}_n) \; = \frac{4}{\pi^2} \Delta l \int_0^{L_a} \gamma(y) \cos(\tilde{l}_j y) \cos(\tilde{l}_n y) \, dy, \tag{2.33}$$

and then a quadrature Trapezoidal rule.

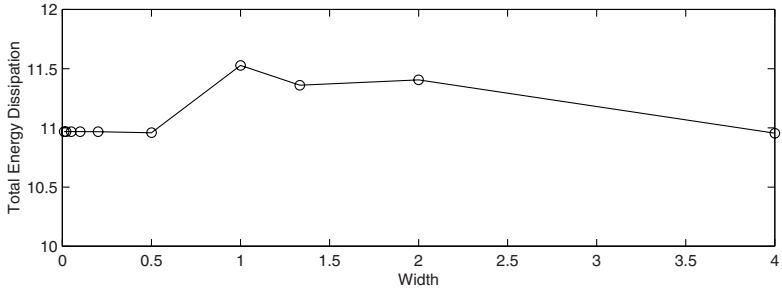


Figure 2.4 Energy dissipation for fixed height 2 and various widths $\in \mathcal{W}$

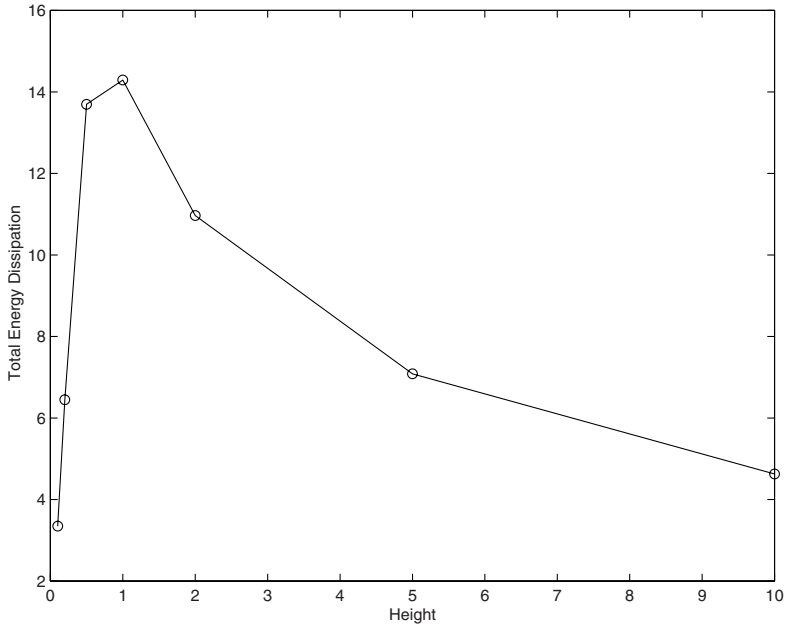


Figure 2.5 Energy dissipation for fixed width .01 and various heights $\in \mathcal{H}$

For our examples, we considered the bandwidth $3.5 \leq |\mathbf{k}| \leq 4.5$. The discrete modes that belong to this bandwidth are shown in Figure 2.3 where the dashed lines represent curves of constant $|\mathbf{k}|$.

The set of these modes is

$$\Upsilon = \{(\tilde{k}_i, \tilde{l}_j) = (\, i - .5, j - .5\,) : (i,j) = (4,1), (4,2), (3,3), (4,3), (1,4), (2,4), (3,4)\}.$$

We assume that the initial field consists only of the velocity field. Thus for the modes in Υ the corresponding set of $\nu = i + 10(j-1)$ values is

$$\mathcal{N} = \{4, 14, 23, 24, 31, 32, 33\},$$

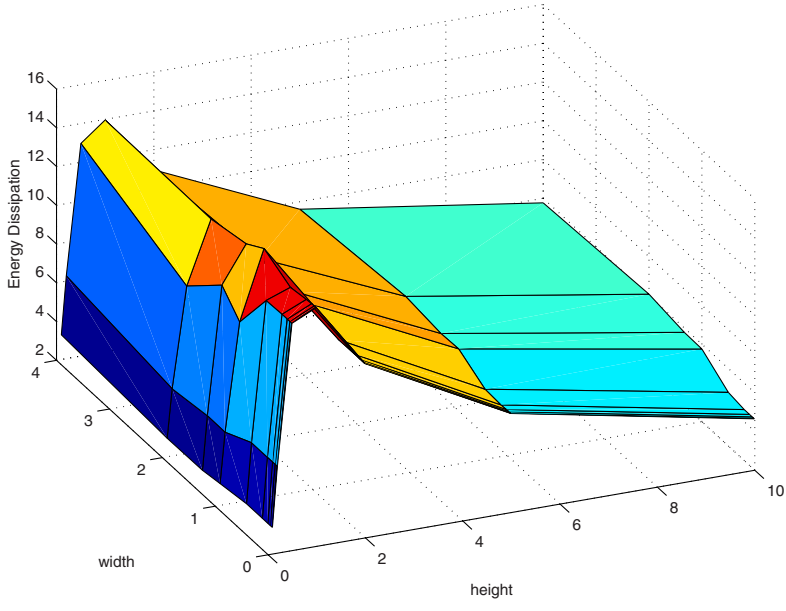


Figure 2.6 3-D plot of energy dissipation with various widths and heights.

and for the coefficients w_ν we have

$$w_\nu = 0, \qquad \nu \notin \mathcal{N},$$

while w_ν, $\nu \in \mathcal{N}$, are given magnitudes that describe the initial field. For simplicity we considered a uniform initial energy distribution of magnitude one for each mode $\nu \in \mathcal{N}$. This leads to the instantaneous correlation matrix given by (see (2.25) and (2.26))

$$C^{\mathcal{N}}(t) = \sum_{\nu \in \mathcal{N}} w_\nu F^{(\nu)}(t) \left(F^{(\nu)}(t)\right)^T,$$

and the corresponding expected instantaneous total energy $E^{\mathcal{N}}$ given by

$$\mathcal{E}\left[E^{\mathcal{N}}(t)\right] = \frac{1}{2} \operatorname{trace}[\mathcal{L} C^{\mathcal{N}}(t) \mathcal{L}]. \tag{2.34}$$

Following the procedures given above and in [2], we computed the energy dissipation for a given admittance $\gamma_{(\mathrm{w,h})}$. We did this for a range of values of width $\mathrm{w} \in \mathcal{W} = \{.01, .02, .05, .1, .2, .5, 1, 1.333, 2, 4\}$ with fixed values of h and also for a range of values of height $\mathrm{h} \in \mathcal{H} = \{.1, .2, .5, 1, 2, 5, 10\}$ with fixed values of w, and plotted the normalized energy dissipation E_{20} at scaled time $t = 20$ given by

$$E_{20} = 10 \log \left[\frac{\mathcal{E}\left[E^{\mathcal{N}}(t)\right]}{\mathcal{E}\left[E^{\mathcal{N}}(0)\right]} \right] \Bigg|_{t=20}.$$

Examples of our findings are plotted in Figure 2.4, where we depict dissipation for fixed height $h = 2$ and various widths $w \in \mathcal{W}$, and in Figure 2.5, where width was fixed at $w = .01$ and $h \in \mathcal{H}$.

These results are quite typical of our findings in which the energy dissipation exhibited much more sensitivity to changes in height than changes in width. A plot of the energy dissipation vs. (w, h) summarizes these findings in Figure 2.6.

2.6 Concluding Remarks

The above results offer promise of the effective use of an array of micro-acoustic actuators as an absorbing surface for enclosed acoustic cavities.

We believe the computational findings discussed here provide a strong motivation for further investigations in the context of *smart* or *active control* in which the design parameter γ in (2.7) is allowed to vary in time and one seeks optimal strategies.

Acknowledgments

This research was supported in part by the US Air Force Office of Scientific Research under grant AFOSR F49620-01-1-0026, in part by the Fulbright Scholar Program, in part by King Fahd University of Petroleum and Minerals through a Sabbatical Leave Program, in part by the David and Lucile Packard Foundation, and in part by the Thomas Lord Research Center, Cary, N.C.

2.7 References

[1] A. O. Andersson. Fluidic element noise and vibration control constructs and methods. European Patent Office, EP0829848, www.european-patent-office.org, 1998.

[2] H. T. Banks, D. G. Cole, K. M. Furati, K. Ito, and G. A. Pinter. A computational model for sound field absorption by acoustic arrays. CRSC Tech Report TR01-19, July, 2001, NCSU; *J. Intel. Material Systems and Structures*, to appear

[3] D. G. Crighton, A. P. Dowling, J. E. F. Williams, M. Heckl, and F. G. Leppington. *Modern Methods in Analytical Acoustics*. Springer-Verlag, London, 1992.

[4] K. U. Ingard. *Notes on Sound Absorption Technology*. Noise Control Foundation, Poughkeepsie, NY, 1994.

[5] P. M. Morse and K. U. Ingard. *Theoretical Acoustics*. Princeton University Press, Princeton, NJ, 1968.

[6] E. G. Williams. *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, San Diego, California, 1999.

3

Some Remarks on Linear Filtering Theory for Infinite Dimensional Systems

Alain Bensoussan

Abstract

In this article, we complete the theory of estimation of Linear Random Functionals, introduced by the Author, in order to extend the Kalman filter to infinite dimensional linear systems. The objective is to show that all properties for the finite dimensional case remain valid in the framework of Linear Random Functionals (this is thanks to linearity of course).

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 27-39, 2003.
© Springer-Verlag Berlin Heidelberg 2003

3.1 Introduction

The Kalman filter for Linear infinite dimensional systems has been developed for many years, in particular in the paper of P.L. Falb [3], and the book the author [2]. In general, this is done once the Wiener process in a Hilbert space is defined. This unfortunately faces the following difficulty. Suppose we have defined a Wiener process in a Hilbert space K, namely $\eta(t;\omega)$, if $\varphi \in L^2(0,T,K)$ then we can define the stochastic integral

$$\int_0^T (\varphi(t), d\eta(t)).$$

We expect the property

$$E\int_0^T (\varphi(t), d\eta(t)) \int_0^T (\psi(t), d\eta(t)) = \int_0^T (\Lambda(t)\varphi(t), \psi(t))dt.$$

It can be proved that this makes sense only whenever the operator $\Lambda(t)$ is nuclear, which means:

$$\sum_{i=1}^{\infty} (\Lambda(t)k_i, k_i) < \infty,$$

for any orthonormal basis $k_1, k_2 \ldots$ of K. This property is incompatible with the invertibility of $\Lambda(t)$ which is desirable to define the analogue of the white noise.

The author has used the concept of Linear Random Functional, which will be recalled later in this paper, to cope with this difficulty. But then, the Wiener process cannot be defined. Thanks to the linearity, and adapting concepts like Least square estimate and Maximum likelihood estimate, it has been possible to give a meaning to the Kalman filter.

The purpose of this paper is to show that the Kalman filter defined in this way has even more natural optimality properties than expected, in particular nothing is lost with respect to the situation of finite dimensional linear systems.

To some extent, linearity allows for full extension of the theory from finite to infinite dimensional systems.

3.2 Linear Random Functionals

Let Φ be a separable Hilbert space, and Φ' its dual space. Let $(\Omega, \mathcal{A}, P)$ be a probability space. A Linear Random Functional (L.R.F.) on Φ' is a family $\zeta_{\varphi_*}(\omega)$ of real random variables, such that:

$$\varphi_* \to \zeta_{\varphi_*}(\omega) \text{ is a.s. linear from } \Phi' \text{ to } R. \tag{3.1}$$

Suppose that

$$\text{the map } \varphi_* \to \zeta_{\varphi_*}(\cdot) \in \mathcal{L}(\Phi'; L^2(\Omega, \mathcal{A}, P)) \tag{3.2}$$

then the quantities $E\zeta_{\varphi_*}$ and $E\zeta_{\varphi_*}\zeta_{\tilde{\varphi}_*}$ have a meaning and moreover we can write:

$$E\zeta_{\varphi_*} = < m, \varphi_* > \qquad m \in \Phi \tag{3.3}$$

$$E\zeta_{\varphi_*}\zeta_{\tilde{\varphi}_*} - E\zeta_{\varphi_*}E\zeta_{\tilde{\varphi}_*} = < \Gamma\varphi_*, \tilde{\varphi}_* > \tag{3.4}$$

With $\Gamma \in \mathcal{L}(\Phi'; \Phi)$, self adjoint and positive.

We call m the mathematical expectation and Γ the covariance operator of the L.R.F. $\zeta_{\varphi_*}(\omega)$.

Of course if $\zeta_{\varphi_*}(\omega)$ is a.s. linear and continuous from Φ' to R then there will exist a random variable $\zeta(\omega)$ with values in Φ, such that:

$$\zeta_{\varphi_*}(\omega) =< \varphi_*, \zeta(\omega) > \tag{3.5}$$

In that case m, Γ will appear as the mathematical expectation and the covariance operator of the random variable $\zeta(\omega)$.

An important concept is the image of a L.R.F. on Φ' by an affine map from Φ to Ψ (an other separable Hilbert space)

$$u(\varphi) = \bar{\psi} + B\varphi \quad , \quad \bar{\psi} \in \Psi \tag{3.6}$$

and $B \in \mathcal{L}(\Phi; \Psi)$. The image is defined by

$$(u\zeta)_{\psi_*}(\omega) =< \bar{\psi}, \psi_* > +\zeta_{B^*\psi_*}(\omega). \tag{3.7}$$

It is easy to check that the mathematical expectation and the covariance operator of the image are given by:

$$E(u\zeta)_{\varphi_*} =< \bar{\psi} + Bm, \psi_* > \tag{3.8}$$

$$E(u\zeta)_{\psi_*}(u\zeta)_{\tilde{\psi}_*} - E(u\zeta)_{\psi_*}E(u\zeta)_{\tilde{\psi}_*} =< B\Gamma B^*\psi_*, \tilde{\psi}_* > \tag{3.9}$$

An interesting case in the sequel is whenever

$$\Phi = L^2(0, T; K), K \text{ separable Hilbert space,} \tag{3.10}$$

with a covariance operator given by:

$$< \Gamma\varphi_*, \tilde{\varphi}_* >= \int_0^T < \Lambda(t)\varphi_*(t), \tilde{\varphi}_*(t) > dt \tag{3.11}$$

where

$$\varphi_* \equiv \varphi_*(\cdot), \tilde{\varphi}_* \equiv \tilde{\varphi}_*(\cdot) \in L^2(0, T; K')$$

and

$$\Lambda(\cdot) \in L^\infty(0, T; \mathcal{L}(K'; K))$$

with $\Lambda(t)$, a.e. a self adjoint non negative operator from K' to K. The operator $\Lambda(t)$ can be invertible. Suppose $K = K' = R^n$, and suppose we have a Wiener process $\eta(t; \omega)$ with values in R^n, such that

$$E\eta(t)\eta^*(s) = \int_0^{min(t,s)} \Lambda(\tau)d\tau$$

where η^* denotes the line vector transposed of η, and Λ is the covariance matrix, then the stochastic integral.

$$\int_0^T k(t) \cdot d\eta(t) \quad , \text{ with } k(\cdot) \in L^2(0, T; R^n)$$

is a L.R.F. on $L^2(0,T;R^n)$. If $\Lambda(\tau) =$ identity, then we get the standard Wiener processor R^n.

For general Hilbert spaces, L.R.F. do not reduce to stochastic integrals.

We shall say that the L.R.F. $\zeta_{\varphi_*}(\omega)$ is Gaussian, whenever the random variable ζ_{φ_*} is gaussian with mean $< m, \varphi_* >$ and variance $< \Gamma\varphi_*, \varphi_* >$. Moreover the covariance of the pair $(\zeta_{\varphi_*}, \zeta_{\tilde{\varphi}_*})$ is $< \Gamma_{\varphi_*}, \tilde{\varphi}_* >$. It is easy to check that if u is an affine map from Φ to Ψ, then the image L.R.F. $(u\zeta)_{\psi_*}(\omega)$ is also a Gaussian L.R.F. on Ψ'.

3.3 Description of the Model and Statement of the Problem

Notation. Assumptions.

Let V, H be two separable Hilbert spaces, with

$$V \subset H \quad , \text{ with continuous embedding.} \tag{3.12}$$

We identifiy H to its dual, thus denoting by V' the dual of V we have the property

$$V \subset H \subset V' \tag{3.13}$$

each space being continuously injected in the next one. We consider a family of operators $A(t)$ such that

$$A(\cdot) \in L^\infty(0,T;\mathcal{L}(V;V')) \tag{3.14}$$

Consider next Hilbert spaces E, F called respectively the input and the output space. We define a linear system as follows:

$$\begin{aligned} &\frac{dy}{dt} + A(t)y = f(t) + B(t)\xi(t) \\ &y(0) = \zeta \end{aligned} \tag{3.15}$$

$$z(t) = C(t)y(t) + \eta(t) \tag{3.16}$$

with

$$\begin{aligned} &f \in L^2(0,T;V'); B \in L^\infty(0,T;\mathcal{L}(E;V')) \\ &C \in L^\infty(0,T;\mathcal{L}(V;F)) \end{aligned} \tag{3.17}$$

The space of data is defined by

$$\Phi = H \times L^2(0,T;E) \times L^2(,T;F) \tag{3.18}$$

For given data $\varphi = (\zeta, \xi(\cdot), \eta(\cdot))$, there exists a unique solution $y(\cdot) \in L^2(0,T;V)$, $y' \in L^2(0,T;V')$ solution of (3.4)

We next consider a L.R.F. on Φ', denoted by $\chi_{\varphi_*}(\omega)$ which is gaussian with mean $(\bar{\zeta}, \bar{\xi}(\cdot), 0)$ and covariance operator:

$$\Gamma = \begin{pmatrix} P_0 & 0 & 0 \\ 0 & Q(t) & 0 \\ 0 & 0 & R(t) \end{pmatrix}$$

in which $\bar{\zeta} \in H, \bar{\xi}(\cdot) \in L^2(0,T;E), P_0 \in \mathcal{L}(H;H)$ positive self adjoint, $Q(\cdot) \in L^\infty(0,T;\mathcal{L}(E';E)), R(\cdot) \in L^\infty(0,T;\mathcal{L}(F';F))$.

Since the solution of (3.4) $\in C^0([0,T];H)$ we can define the map from Φ to H, by:

$$\varphi = (\zeta, \xi(\cdot), \eta(\cdot)) \to y(T) = u_T(\varphi). \tag{3.19}$$

Of course, $\eta(\cdot)$ does not play any role in (3.8) and is left for convenience. We next define the map from Φ to $L^2(0,T;F)$ by:

$$\varphi = (\zeta, \xi(\cdot), \eta(\cdot)) \to z(\cdot) = v(\varphi). \tag{3.20}$$

The maps u_T and v are affine continuous. Following the notation (2.7), they can be written as follows:

$$u_T(\varphi) = \bar{y}(T) + \tilde{u}_T(\varphi) \tag{3.21}$$

$$v(\varphi) = \bar{z}(\cdot) + \tilde{v}(\varphi) \tag{3.22}$$

in which $\bar{y}(\cdot)$ is the solution of:

$$\begin{aligned} &\frac{d\bar{y}}{dt} + A(t)\bar{y} = f(t) \\ &\bar{y}(0) = 0 \end{aligned} \tag{3.23}$$

and $\bar{z}$ is given by

$$\bar{z}(t) = C(t)\bar{y}(t) \tag{3.24}$$

Clearly $\tilde{u}_T, \tilde{v}$ are linear and more precisely

$$\tilde{u}_T(\varphi) = \tilde{y}(T) \quad \tilde{v}(\varphi) = \tilde{z} \tag{3.25}$$

where $\tilde{y}(\cdot)$ is the solution of

$$\begin{aligned} &\frac{d\tilde{y}}{dt} + A(t)\tilde{y} = B(t)\xi(t) \\ &\tilde{y}(0) = \zeta \\ &\tilde{z}(t) = C(t)\tilde{y}(t) + \eta(t) \end{aligned} \tag{3.26}$$

The maps u_T and v induce L.R.F.s on H ans $L^2(0,T;F')$ respectively, denoted by $y_h(T)$ and Z_{f_*} where $h \in H, f_* \in L^2(0,T;F')$. More explicitly we have:

$$y_h(T)(\omega) = (u_T\chi)_h(\omega) \tag{3.27}$$

$$Z_{f_*}(\omega) = (v\chi)_{f_*}(\omega) \tag{3.28}$$

Clearly, also following (2.7)

$$y_h(T)(\omega) = (h, \bar{y}(T)) + \chi_{\tilde{u}_T^* h}(\omega) \tag{3.29}$$

$$Z_{f_*}(\omega) = \int_0^T < \bar{z}(t), f_*(t) > dt + \chi_{\tilde{v}^* f_*}(\omega). \tag{3.30}$$

The mathematical expectation of respectively $y_h(T), Z_{f_*}$ are given by

$$Ey_h(T) = (h, \bar{\bar{y}}(T)) \tag{3.31}$$

$$EZ_{f_*} = \int_0^T < \bar{\bar{z}}(t), f_*(t) > dt \tag{3.32}$$

where

$$\begin{aligned} &\frac{d\bar{\bar{y}}}{dt} + A(t)\bar{\bar{y}} = f + B\bar{\bar{\xi}} \\ &\bar{\bar{y}}(0) = \bar{\zeta} \end{aligned} \tag{3.33}$$

$$\bar{\bar{z}}(t) = C(t)\bar{\bar{y}}(t). \tag{3.34}$$

The covariance operators of $y_h(T), Z_{f_*}$ are the following

$$Ey_h(T)y_{h'}(T) - Ey_h(T)Ey_{h'}(T) = (\tilde{u}_T \Gamma \tilde{u}_T^* h', h) \tag{3.35}$$

$$EZ_{f_*}Z_{f_*'} - EZ_{f_*}EZ_{f_*'} = < \tilde{v}\Gamma\tilde{v}^* f_*', f_* > . \tag{3.36}$$

More explicit expressions will be given later.

Let

$$\mathcal{B} = \sigma(Z_{f_*}, \forall f_* \in L^2(0, T; F')). \tag{3.37}$$

The estimation (filtering) problem we are interested in is to compute:

$$\widehat{y_h(T)} = E[y_h(T)|\mathcal{B}] \tag{3.38}$$

the conditional expectation of the random variable $y_h(T)$, given the σ-algebra $\mathcal{B}$.

The interpretation is clear. Considering the input output relationship (3.5), (3.6) $y(t)$ represents the state of a dynamic system at time t, and $z(t)$ is the output, or observation process. The random character of the input $(\zeta, \xi(\cdot))$ and the measurement error $\eta(\cdot)$ is modelled through the L.R.F. χ_{φ_*}, which induces a random state modelled by the L.R.F. $y_h(T)$ and a random observation process modelled by the L.R.F. Z_{f_*}. Thus (3.27) is the direct analogue of the usual filtering problem, in this context.

Note that $\widehat{y_h(T)}$ is a L.R.F. on H, and our objective is to characterize it in a convenient way (in particular recovering the recursivity property of the filter).

Preliminaries

We shall need the

LEMMA 3.1
The σ-algebra $\mathcal{B}$ is generated by the family of random variables $Z_{g_i 1\!\!1_s}$, where $g_1, g_2, \ldots, g_i, \ldots$ is a orthonormal basis of $F', \forall s \leq T$, with the notation

$$g_i 1\!\!1_s \equiv g_i 1\!\!1_s(t) \equiv \begin{cases} g_i & t \leq s \\ 0 & t > s \end{cases}$$

□

Proof To simplify the notation, we set $G = F'$ and we denote by g a generic element of G. Call

$$\tilde{\mathcal{B}} = \sigma(Z_{g_i 1\!\!1_s}, \forall i, \forall s \leq T)$$

then $\tilde{\mathcal{B}} \subset \mathcal{B}$. It is sufficient to show that $Z_{g(\cdot)}$ is $\tilde{\mathcal{B}}$ measurable $\forall g(\cdot) \in L^2(0,T;G)$. First we show that $Z_{g 1\!\!1_s}$ is $\tilde{\mathcal{B}}$ measurable, $\forall g \in G$. Indeed

$$g = \lim_{m \to \infty} \sum_{i=1}^{m} ((g, g_i)) g_i = lim_{n \to \infty} g^m$$

and $Z_{g^m 1\!\!1_s}$ is clearly $\varphi\mathcal{B}$ measurable. On the other hand $g_m 1\!\!1_s(\cdot) \to g 1\!\!1_s(\cdot)$ in $L^2(0,T;G)$. Therefore for a subsequence

$$Z_{g^{m'} 1\!\!1_s} \to Z_{g 1\!\!1_s} a.s.$$

since $Z_{g^m 1\!\!1_s} \to Z_{g 1\!\!1_s}$ in $L^2(\Omega, \mathcal{A}, P)$.

Therefore $Z_{g 1\!\!1_s}$ is $\tilde{\mathcal{B}}$ measurable, $\forall g \in G, \forall s \leq T$. Consider now any $g(\cdot) \in L^2(0,T;G)$. Define the approximation:

$$g^n(t) = g^{n,j} = \frac{n}{T} \int_{j\frac{T}{n}}^{(j+1)\frac{T}{n}} g(s)ds \qquad , \quad \frac{jT}{n} < t < (j+1)\frac{T}{N}, \quad j = 0, \ldots (n-1)$$

As well known

$$g^n(\cdot) \to g(\cdot) \text{ in } L^2(0,T;G).$$

Therefore

$$Z_{g^n(\cdot)} \to Z_{g\cdot)} \text{ in } {}^2(\Omega, \mathcal{A}, P)$$

and for a subsequence

$$Z_{g^{n'}(\cdot)} \to Z_{g(\cdot)} \text{ a.s.} \tag{3.39}$$

On the other hand

$$Z_{g^n(\cdot)} = \sum_{j=0}^{n-1} \left(Z_{g^{n,j} 1\!\!1_{(j+1)\frac{T}{n}}} - Z_{g^{n,j} 1\!\!1_{j\frac{T}{N}}} \right)$$

which implies that $Z_{g^n(\cdot)}$ is $\tilde{\mathcal{B}}$ measurable. Thus, thanks to (3.28) $Z_{g(\cdot)}$ is also $\tilde{\mathcal{B}}$ measurable. □

3.4 Obtaining the Best Estimate

Linear filter

Let $S \in \mathcal{L}(L^2(0,T;F);H)$. We associate to S the affine map from Φ to H, defined as follows:

$$\mathcal{F}_S(\varphi) = \bar{\bar{y}}(T) + S(z(\cdot) - \bar{\bar{z}}(\cdot)) \tag{3.40}$$

which we call a linear filter with kernel S. To such a map we associate the L.R.F. on H, image of χ_{φ_*} by $\mathcal{F}_S$, namely:

$$y_h^S(T)(\omega) = (\mathcal{F}_S \chi)_h(\omega). \tag{3.41}$$

It is easy to check also that:

$$y_h^S(T)(\omega) = (\bar{\bar{y}}(T), h) + S(\bar{z}(\cdot - \bar{\bar{z}}(\cdot)) + \chi_{\tilde{v}^* S^* h}(\omega).$$

We see from (3.19) that $y_h^S(T)$ is $\mathcal{B}$ measurable for any $S(\cdot)$. We next define the linear estimate error by:

$$\begin{aligned} \varepsilon_h^S(T)(\omega) &= y_h(T)(\omega) - y_h^S(T)(\omega) \\ &= (\bar{y}(T) - \bar{\bar{y}}(T), h) - S(\bar{z}(\cdot) - \bar{\bar{z}}(\cdot)) + \chi_{(\tilde{u}_T^* - \tilde{v}^* S^*)(h)}(\omega) \end{aligned} \tag{3.42}$$

We can easily verify that

$$E\varepsilon_h^S(T) = 0.$$

We shall say that $\hat{S}$ is the best linear filter if we have the optimality property

$$E|\varepsilon_h^{\hat{S}}(T)|^2 \leq E|\varepsilon_h^S(T)|^2, \forall S, \forall h. \tag{3.43}$$

We shall prove that such $\hat{S}$ exists and is unique.

The best linear filter is the solution of the filtering problem

We want here to prove the following important result:

THEOREM 3.1

If $\hat{S}$ satisfies the property (4.5) (the best linear filter exists), then one has:

$$\widehat{y_h(T)} = y_h^{\hat{S}}(T) \tag{3.44}$$

□

Proof Consider, from (4.4)

$$\begin{aligned} \varepsilon_h^{\hat{S}+S}(T) &= \varepsilon_h^{\hat{S}}(T) - S(\bar{z}(\cdot) - \bar{\bar{z}}(\cdot)) - \chi_{\tilde{v}^* S^* h} \\ &= \varepsilon_h^{\hat{S}}(T) + S(\bar{\bar{z}}(\cdot)) - Z_{S^* h} \end{aligned}$$

where we have used (3.19). Taking in (4.5) $S = \hat{S} + S$, we deduce easily the inequality

$$\begin{aligned} 0 \leq E|S(\bar{\bar{z}}(\cdot)) - Z_{S^*h}|^2 \\ -2E\varepsilon_h^{\hat{S}}(T)Z_{S^*h} \end{aligned} \tag{3.45}$$

Since S is arbitrary, this implies necessarily

$$E\varepsilon_h^{\hat{S}}(T)Z_{S^*h} = 0 \qquad \forall S, \forall h \tag{3.46}$$

Take g_i one of the vectors of the orthonormal basis of F' (see Lemma 3.1) and let $s \leq T$. For any h define:

$$S_{h,g_i,s}f(\cdot) = \int_0^t < g_i, f(t) > dt \frac{h}{|h|^2}, \forall f(\cdot) \in L^2(0,T;F).$$

So $S_{h,g_i,s} \in \mathcal{L}(L^2(0,T;F);H)$. Moreover

$$S^*_{h,g_i,s}h = g_i 1\!\!1_s(\cdot)$$

as easily seen. Applying (4.8) with $S \equiv S_{h,g_i,s}$ we thus obtain:

$$E\varepsilon_h^{\hat{S}}(T)Z_{g_i 1\!\!1_s(\cdot)} = 0 \qquad \forall i, \forall s. \tag{3.47}$$

Since $\varepsilon_h^{\hat{S}}(T)$ and $Z_{g_i 1\!\!1_s}$ are gaussian, the non correlation implies the independence, hence $\varepsilon_h^{\hat{S}}(T)$ is independent from $Z_{g_i 1\!\!1_s} \forall i, \forall s$. Therefore $\varepsilon_h^{\hat{S}}(T)$ is independent from $\mathcal{B}$. This implies (4.6) which completes the proof. □

Computation of the best linear filter

It remains to prove that $\hat{S}$ exists and is unique. First for a given h and any S, introduce the solution $\beta(\cdot) \in L^2(0,T;V), \beta' \in L^2(0,T;V')$ to be the solution of the backward parabolic equation

$$\begin{aligned} &\tfrac{-d\beta}{dt} + A^*(t)\beta = -C^*(t) \cdot S^*h(t) \\ &\beta(T) = h \end{aligned} \tag{3.48}$$

We shall check the property:

$$(\tilde{u}_T^* - \tilde{v}^*S^*)(h) = \begin{pmatrix} \beta(0) \\ B^*(\cdot)\beta(\cdot) \\ -S^*h \end{pmatrix} \tag{3.49}$$

Indeed let $\varphi = (\zeta, \xi(\cdot), \eta(\cdot)) \in \Phi$, we need to check the relation

$$\begin{aligned} (h, (\tilde{u}_T - S\tilde{v})(\varphi)) = \\ = (\beta(0), \zeta) + \int_0^T < \beta(t), B(t)\xi(t) > dt - (h, S\eta(\cdot)) \end{aligned} \tag{3.50}$$

Recalling (3.14), this amounts to

$$\begin{aligned}(h,(\tilde{y}(T))-(h,S\tilde{z}(\cdot))=\\ =(\beta(0),\zeta)+\int_0^T<\beta(t),B(t)\xi(t)>dt-(h,S\eta(\cdot))\end{aligned}\tag{3.51}$$

of from (3.15)

$$\begin{aligned}(h,(\tilde{y}(T))-(h,S(C(\cdot)\tilde{y}(\cdot)))=\\ =(\beta(0),\zeta)+\int_0^T<\beta(t),B(t)\xi(t)>dt\end{aligned}\tag{3.52}$$

which is clear from integration by part properties between the equations (3.15) and (4.10).

From (4.4) and (4.11) we can assert that

$$\begin{aligned}E|\varepsilon_h^S(T)|^2=(P_0\beta(0),\beta(0))+\int_0^T<Q(t)^*(t)\beta(t),B^*(t)\beta(t)>dt\\ +\int_0^T<R(t)S^*h(t),S^*h(t)>dt\end{aligned}\tag{3.53}$$

To the problem of minimizing the right hand side of (4.15) with respect to S we associate the optimal control problem

$$\begin{aligned}&\frac{-d\gamma}{dt}+A^*(t)\gamma=-C^*(t)v(t)\\ &\gamma(T)=h\end{aligned}\tag{3.54}$$

$$\begin{aligned}&J(v(\cdot))=(P_0\gamma(0),\gamma(0))+\int_0^T<B(t)Q(t)B^*(t)\gamma(t),\gamma(t)>dt\\ &+\int_0^T<R(t)v(t),v(t)>dt\end{aligned}\tag{3.55}$$

where the control $v(\cdot)\in L^2(0,T;F')$.

Assume from now on the coercivity property:

$$<R(t)f_*,f_*>\geq r\|f_*\|^2,r>0,\forall f_*\in F'.\tag{3.56}$$

Then the optimal control problem (4.16), (4.17) has a unique solution u. Writing the Euler equation, and introducing the adjoint system, it is standard to show that the coupled linear system:

$$\begin{aligned}&\frac{-d\hat{\alpha}}{dt}+A(t)\hat{\alpha}+B(t)Q(t)B^*(t)\hat{\gamma}(t)=0\\ &\hat{\alpha}(0)+P_0\hat{\gamma}(0)=0\\ &\frac{-d\hat{\alpha}}{dt}+A^*(t)\hat{\gamma}+C^*(t)R^{-1}(t)C(t)\hat{\alpha}=0\\ &\hat{\gamma}(T)=h\end{aligned}\tag{3.57}$$

has a unique solution. The optimal control $u(\cdot)$ is given by

$$u(t) = -R^{-1}(t)C(t)\hat{\alpha}(t) \tag{3.58}$$

We can write

$$\hat{S}^*(t) = -R^{-1}(t)C(t)\hat{\alpha}(t) \tag{3.59}$$

which defines an element $\hat{S}$ of $\mathcal{L}(L^2(0,T;F),h)$, by the formula

$$(\hat{S}f(\cdot),h) = -\int_0^T < f(t), R^{-1}(t)C(t)\hat{\alpha}(t) > dt, \forall f(\cdot), \forall h. \tag{3.60}$$

Obviously $\hat{S}$ satisfies the property (4.5). Moreover

$$\inf_{v(\cdot)} J(v(\cdot)) = \inf_S E|\varepsilon_h^S(T)|^2 \tag{3.61}$$

and since the optimal control of the problem at the left hand side of (4.23) is unique, $\hat{S}$ is also the unique element minimizing the right hand side of (4.23).

So we have proved the following

THEOREM 3.2
Assume (4.18). Then there exists a unique best linear filter $\hat{S}$ defined by duality by the formula (4.21). □

To make $\hat{S}$ more explicit, we introduce the coupled system

$$\begin{aligned}
&\frac{-d\hat{y}}{dt} + A(t)\hat{y} + B(t)Q(t)B^*(t)\hat{p} = 0\\
&\frac{-d\hat{p}}{dt} + A^*(t)\hat{p} - C^*(t)R^{-1}(t)C(t)\hat{y}(t) = -C^*(t)R^{-1}(t)z(t)\\
&\hat{y}(0) + P_0\hat{p}(0) = 0\\
&\hat{p}(T) = 0
\end{aligned} \tag{3.62}$$

for $z(\cdot) \in L^2(0,T;F)$ given. The system (4.24) defines in a unique way the pair $\hat{y},\hat{p}$ since it corresponds to the Euler equation of the control problem.

$$\begin{aligned}
&\frac{-dp}{dt} + A^*(t)p = -C^*(t)v(t) - C^*(t)R^{-1}(t)z(t)\\
&p(T) = 0
\end{aligned} \tag{3.63}$$

$$\begin{aligned}
K(v(\cdot)) = (P_0p(0), (p(0) + \int_0^T < B(t)Q(t)B^*(t)p(t), p(t) > dt\\
+ \int_0^T < R(t)v(t), v(t) > dt
\end{aligned} \tag{3.64}$$

(see (4.16), (4.17)).

It is easy to check that

$$\hat{S}z(\cdot) = \hat{y}(T) \tag{3.65}$$

3.5 Final Form. Decoupling Theory

Riccati equation

$$C(\cdot) \in L^{\infty}(0,T;\mathcal{L}(F;H)) \tag{3.66}$$

then we rely on the theory of Riccati equations (see J.L. Lions [4], A. Bensoussan [1]) to assert that there exists a unique $P(t)$ satisfying

$$P(\cdot) \in L^{\infty}(0,T;\mathcal{L}(H;H)), P(t) \geq 0, \text{ self adjoint} \tag{3.67}$$

$$\text{If } \theta \in L^2(0,T;V)), \theta' \in L^2(0,T;V')), \frac{-d\theta}{dt} + A^*\theta \in L^2(0,T;H) \tag{3.68}$$

then

$$P\theta \in L^2(0,T;V)), \frac{d}{dt}P\theta \in L^2(0,T;V')$$

$$\begin{aligned} &\frac{d}{dt}(P\theta) + P(-\frac{d\theta}{dt} + A^*\theta) + AP\theta + PC^*R^{-1}CP\theta = QB^*\theta \\ &P(0) = P_0 \end{aligned} \tag{3.69}$$

Applying (5.4) with $\theta = \hat{p}$ (see (4.24)), and setting

$$e(t) = \hat{y}(t) + P(t)\hat{p}(t) \tag{3.70}$$

it easily follows from (4.23), (5.4) that e satisfies

$$e \in L^2(0,T;V)), \frac{de}{dt} \in L^2(0,T;V') \tag{3.71}$$

$$\frac{de}{dt} + A(t)e = P(t)C^*(t)R^{-1}(t)(z(t) - C(t)e(t)), \quad e(0) = 0 \tag{3.72}$$

Note that from (4.26) we also have:

$$\hat{S}z(\cdot) = e(T) \tag{3.73}$$

Coming back to (4.1), applying (5.7) with $z(\cdot)$ changed into $z(\cdot) - \bar{\bar{z}}(\cdot)$ and setting:

$$r(t) = e(t) + \bar{\bar{y}}(t) \tag{3.74}$$

We deduce that r is the solution of

$$\begin{aligned} &\tfrac{dr}{dt} + A(t)r = f(t) + B(t)\bar{\bar{\xi}}(t) + P(t)C^*(t)R^{-1}(t)(z(t) - C(t)r(t)) \\ &r(0) = \bar{\bar{\xi}} \end{aligned} \tag{3.75}$$

and we have:

$$\mathcal{F}_{\hat{S}}(\varphi) = r(T) \tag{3.76}$$

which is the usual form of the Kalman filter.

Interpretation of $P(T)$

Applying (5.4) with $\theta = \hat{\gamma}$ (see (4.18)) it is easy to check that

$$\hat{\alpha}(t) + P(t)\hat{\gamma}(t) = 0 \tag{3.77}$$

and thus also

$$\hat{\alpha}(T) = -P(T)h \tag{3.78}$$

together with

$$E|\varepsilon_h^{\hat{S}}(T)|^2 = (P(T)h, h) \tag{3.79}$$

which means that $P(T)$ is the covariance operator of the L.R.F. $\varepsilon_h^{\hat{S}}(T)$.

3.6 References

[1] A. BENSOUSSAN, *Identification et filtrage*, Cahiers de l'IRIA nř1,(1969).

[2] A. BENSOUSSAN, *Filtrage optimal des systèmes linéaires* Dunod, Paris, (1971).

[3] P.L. FALB, *Infinite dimensional filtering. The Kalman Bucy filter in a Hilbert space*, Information and Control, Vol 11, pp. 102-137 (1967).

[4] J.L. LIONS, Contrôle optimal de Systèmes aux dérivées partielles, Dunod, Paris (1968).

4

A Note on Stochastic Dissipativeness

Vivek S. Borkar *Sanjoy K. Mitter*

Abstract

In this paper we present a stochastic version of Willems' ideas on Dissipativity and generalize the dissipation inequality to Markov Diffusion Processes.We show the relevance of these ideas by examining the problem of Ergodic Control of partially observed diffusions.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 41-49, 2003.
© Springer-Verlag Berlin Heidelberg 2003

4.1 Introduction

In [8, 9], Willems introduced the notion of a dissipative dynamical system with associated 'supply rate' and 'storage function,' with a view to building a Lyapunov-like theory of input-output stability for deterministic control systems. In this article, we extend these notions to stochastic systems, specifically to controlled diffusions. This makes contact with the ergodic control problem for controlled diffusions ([2], Chapter VI) and offers additional insight into the latter. In particular, it allows us to obtain a "martingale dynamic programming principle" for ergodic control under partial observations in the spirit of Davis and Varaiya [6]. So far this has been done only in special cases using a vanishing discount limit, see Borkar [3, 4, 5].

The next section introduces the notation and key definitions. Section 3 considers the links with ergodic control with complete or partial observations.

4.2 Notation and Definitions

Our controlled diffusion will be a $d \geq 1$ dimensional process

$$X(\cdot) = [X_1(\cdot), \ldots, X_d(\cdot)]^T$$

satisfying the stochastic differential equation

$$X(t) = X_0 + \int_0^t m\Big(X(s), u(s)\Big)\mathrm{d}s + \int_0^t \sigma\Big(X(s)\Big)\mathrm{d}W(s) \quad , \quad t \geq 0 \quad . \tag{4.1}$$

Here,

(i) for a prescribed compact metric 'control' space U,

$$m(\cdot, \cdot) = [m_1(\cdot, \cdot), \ldots, m_d(\cdot, \cdot)]^T : R^d \times U \to R^d$$

is continuous and Lipschitz in its first argument uniformly with regards to the second argument,

(ii) $\sigma(\cdot) = [[\sigma_{ij}(\cdot)]]_{1 \leq i,j \leq d} : R^d \to R^{d \times d}$ is Lipschitz,

(iii) X_0 is an R^d-valued random variable with a prescribed law π_0,

(iv) $W(\cdot) = [W_1(\cdot), \ldots, W_d(\cdot)]^T$ is a d-dimensional standard Brownian motion independent of X_0 and

(v) $u(\cdot) : [0, \infty) \to U$ is a control process with measurable sample paths satisfying the nonanticipativity condition: for $t \geq s$, $W(t) - W(s)$ is independent of $u(\tau)$, $W(\tau)$, $\tau \leq s$ and X_0.

We shall consider a weak formulation of (4.1), i.e., we look for $(X(\cdot), u(\cdot), W(\cdot), X_0)$ on *some* probability space so that (4.1) holds. See [2], Chapter 1, for an exposition of the weak formulation. In particular, letting $\mathcal{F}_t$ denote the right-continuous completion of $\sigma(X(s), s \leq t)$ for $t \geq 0$, it is shown in Theorem 2.2, pp. 18-19, [2] that it suffices to consider $u(\cdot)$ adapted to $\langle \mathcal{F}_t \rangle$, i.e., $u(\cdot)$ of the

form $u(t) = f_t\Big(X([0,t])\Big)$ where $X([0,t])$ denotes the restriction of $X(\cdot)$ to $[0,t]$ and $f_t : C([0,t]; R^d) \to U$ are measurable maps. If in addition $u(t) = v(X(t))$, $t \geq 0$, for a measurable $v : R^d \to U$, we call $u(\cdot)$ (or, by abuse of terminology, the map $v(\cdot)$ itself) a Markov control.

We shall also be interested in the partially observed case ([2], Chapter V). Here one has an associated observation process $Y(\cdot)$ taking values in $R^m (m \geq 1)$, given by:

$$Y(t) = \int_0^t h\Big(X(s)\Big)\mathrm{d}s + W'(t) \quad , \quad t \geq 0 \quad . \tag{4.2}$$

where $h : R^d \to R^m$ is continuous and $W'(\cdot)$ an m-dimensional standard Brownian motion independent of $W(\cdot), X_0$. Let $\langle \mathcal{G}_t \rangle$ denote the right-continuous completion of $\sigma(Y(s), s \leq t)$ for $t \geq 0$. We say that $u(\cdot)$ is strict sense admissible if it is adapted to $\langle \mathcal{G}_t \rangle$.

For the completely observed control problem where we observe $X(\cdot)$ directly and $u(t) = f_t\Big(X([0,t])\Big)$ for $t \geq 0$, we define

DEFINITION 4.1
A measurable function $V : R^d \to R$ is said to be a storage function associated with a supply rate function $g \in C(R^d \times U)$ if it is bounded from below and $V\Big(X(t)\Big) + \int_0^t g\Big(X(s), u(s)\Big)\mathrm{d}s$, $t \geq 0$, is an $\langle \mathcal{F}_t \rangle$-super martingale for all $\Big(X(\cdot), u(\cdot)\Big)$ satisfying (4.1) as above. □

The storage function need not be unique. For example, we get another by adding a constant. For g, $(X(\cdot), u(\cdot))$ as above, let

$$V_c(x) = \sup_{u(\cdot)} \sup_{\tau} E\Big[\int_0^\tau g\Big(X(s), u(s)\Big)\mathrm{d}s / X_0 = x\Big] \quad , \quad x \in R^d \quad ,$$

where the first supremum is over all bounded $\langle \mathcal{F}_t \rangle$-stopping times and the second supremum is over all $\langle \mathcal{F}_t \rangle$-adapted $u(\cdot)$. Since $\tau = 0$ is a stopping time, $V_c(\cdot) \geq 0$.

LEMMA 4.1
If $V_c(x) < \infty$ for all x, it is the least nonnegative storage function associated with g. □

Proof In the following, τ denotes an $\langle \mathcal{F}_t \rangle$-stopping time. For $t \geq 0$,

$$\begin{aligned} V_c(x) &\geq \sup_{u(\cdot)} \sup_{\tau \geq t} E\left[\int_0^\tau g\Big(X(s), u(s)\Big)\mathrm{d}s / X_0 = x\right] \\ &= \sup_{u([0,t])} E\left[\int_0^t g\Big(X(s), u(s)\Big)\mathrm{d}s \right. \\ &\qquad \left. + \sup_{u(t+\cdot)} \sup_{\tau \geq t} E\Big[\int_t^\tau g\Big(X(s), u(s)\Big)\mathrm{d}s / X(t)\Big] / X_0 = x\right] \quad , \end{aligned}$$

where the equality follows by a standard dynamic programming argument. Thus

$$V_c(x) \geq \sup_{u(\cdot)} E\Big[\int_0^t g\Big(X(s),u(s)\Big)\mathrm{d}s + V_c\Big(X(t)\Big)/X_0 = x\Big] \quad . \tag{4.3}$$

Now suppose $s \leq \tau \leq T < \infty$, where $T > 0$ is deterministic and s, τ are $\langle \mathcal{F}_t \rangle$-stopping times. Then

$$E\Big[\int_0^\tau g\Big(X(s),u(s)\Big)\mathrm{d}s + V_c\Big(X(\tau)\Big)/\mathcal{F}_s\Big]$$

$$= \int_0^s g\Big(X(s),u(s)\Big)\mathrm{d}s + E\Big[\int_s^\tau g\Big(X(s),u(s)\Big)\mathrm{d}s + V_c\Big(X(\tau)\Big)/\mathcal{F}_s\Big] \quad .$$

By Theorem 1.6, p. 13 of [2], the regular condition law of $(X_{s+\cdot})$, $u(s+\cdot)$ given $\mathcal{F}_s$ is again the law of a pair $(\tilde{X}(\cdot),\tilde{u}(\cdot))$ satisfying (4.1) with initial condition $X(s)$, a.s. Therefore by (4.3), the above is less than or equal to

$$\int_0^s g\Big(X(s),u(s)\Big)\mathrm{d}s + E\Big[V_c\Big(X(s)\Big)/\mathcal{F}_s\Big] = \int_0^s g\Big(X(s),u(s)\Big)\mathrm{d}s + V_c\Big(X(s)\Big) \quad .$$

It follows that

$$\int_0^t g\Big(X(s),u(s)\Big)\mathrm{d}s + V_c\Big(X(t)\Big) \quad , \quad t \geq 0 \quad ,$$

is an $\langle \mathcal{F}_t \rangle$-supermartingale. Thus $V_c(\cdot)$ is a storage function. If $F(\cdot)$ is another nonnegative storage function, we have, by the optional sampling theorem

$$\begin{aligned} F(x) &\geq E\Big[\int_0^\tau g\Big(X(s),u(s)\Big)\mathrm{d}s + F\Big(X(\tau)\Big)/X_0 = x\Big] \\ &\geq E\Big[\int_0^\tau g\Big(X(s),u(s)\Big)\mathrm{d}s/X_0 = x\Big] \quad . \end{aligned}$$

for any bounded $\langle \mathcal{F}_t \rangle$-stopping time τ. Therefore

$$F(x) \geq \sup_{u(\cdot)} \sup_{\tau} E\Big[\int_0^\tau g\Big(X(s),u(s)\Big)\mathrm{d}s/X_0 = x\Big] = V_c(x) \quad .$$

This completes the proof. □

LEMMA 4.2
If $V_c(\cdot) < \infty$, it can also be defined by

$$V_c(x) = \sup_{u(\cdot)} \sup_{t \geq 0} E\Big[\int_0^t g\Big(X(s),u(s)\Big)\mathrm{d}s/X_0 = x\Big] \quad .$$

□

Proof Let $\bar{V}_c(x)$ denote the R.H.S. above. Then clearly $V_c(x) \geq \bar{V}_c(x)$. On the other hand, an argument similar to that of the preceding lemma shows that for $t \geq r \geq 0$,

$$V\Big(X(r)\Big) + \int_0^r g\Big(X(s), u(s)\Big)\mathrm{d}s \geq E\Big[V\Big(X(t)\Big) + \int_0^t g\Big(X(s), u(s)\Big)\mathrm{d}s / \mathcal{F}_r\Big] \quad ,$$

implying that

$$\int_0^t g\Big(X(s), u(s)\Big)\mathrm{d}s + V\Big(X(t)\Big) \quad , \quad t \geq 0 \quad ,$$

is an $\langle \mathcal{F}_t \rangle$-supermartingale. Thus $\bar{V}_c(\,\cdot\,)$ is a storage function. Clearly, $\bar{V}_c(\,\cdot\,) \geq 0$. Thus by the preceding lemma, $\bar{V}_c(\,\cdot\,) \geq V_c(\,\cdot\,)$ and hence $\bar{V}_c(\,\cdot\,) = V_c(\,\cdot\,)$. □

For the partially observed control problem, the correct 'state' is $\pi_t \stackrel{\Delta}{=}$ the regular conditional law of $X(t)$ given $\zeta_t \stackrel{\Delta}{=}$ the right-continuous completion of $\sigma\Big(Y(s), u(s), s \leq t\Big)$, $t \geq 0$. Note that $\zeta_t = \mathcal{G}_t$ for strict sense admissible $u(\,\cdot\,)$, but we shall allow the so called wide-sense admissible $u(\,\cdot\,)$ of [7]. (See, also, [2], Chapter V.) Thus, in general, $\mathcal{G}_t \subset \zeta_t$. Let $\mathcal{P}(R^d) =$ the Polish space of probability measures on R^d with Prohorov topology. Viewing $\langle \pi_t \rangle$ as a $\mathcal{P}(R^d)$-valued process, its evolution is given by the nonlinear filter, defined as follows: For $f : R^d \to R$ that are twice continuously differentiable with compact supports,

$$\pi_t(f) \quad = \quad \pi_0(f) + \int_0^t \pi_s\Big(Lf\Big(\,\cdot\,, u(s)\Big)\Big)\mathrm{d}s + \int_0^t \langle \pi_s(hf) - \pi_s(h)\pi_s(f), \mathrm{d}\tilde{Y}(s)\rangle \,,$$

$$t \geq 0 \quad , \qquad (4.4)$$

where:

(i) $\nu(f) \stackrel{\Delta}{=} \int f \mathrm{d}\nu$ for $f \in C_b(R^d)$, $\nu \in \mathcal{P}(R^d)$,

(ii) $Lf(x,u) = \dfrac{1}{2}\displaystyle\sum_{ijk} \sigma_{ik}(x)\sigma_{jk}(x)\frac{\partial^2 f}{\partial x_i \partial x_j}(x) + \sum_i m_i(x,u)\frac{\partial f}{\partial x_i}(x)$ for twice continuously differentiable $f : R^d \to R$,

(iii) $\tilde{Y}(t) = Y(t) - \int_0^t \pi_s(h)\mathrm{d}s$, $t \geq 0$, is an m-dimensional standard Brownian motion.

See [2], Chapter V, for a discussion of wellposedness and related issues for (4.4). In particular, for a given pair $(\tilde{Y}(\,\cdot\,), u(\,\cdot\,))$ of a Brownian motion and a wide sense admissible control, (4.4) has a unique solution if

(i) $\sigma(\,\cdot\,)$ is nondegenerate, i.e., the least eigenvalue of $\sigma(\,\cdot\,)\sigma(\,\cdot\,)^T$ is uniformly bounded away from zero, and

(ii) $h(\,\cdot\,)$ is twice continuously differentiable, bounded, with bounded first and second partial derivatives. (This can be relaxed — see [7].)

The partially observed control problem then is equivalent to the completely observed control problem of controlling the $\mathcal{P}(R^d)$-valued process $\{\pi_t\}$ governed by (4.4), with wide sense admissible $u(\,\cdot\,)$. By analogy with Definition 4.1 above, we have

DEFINITION 4.2
A measurable function $\bar{V} : \mathcal{P}(R^d) \to R$ is said to be a storage function associated with the supply rate function $g \in C(\mathcal{P}(R^d) \times U)$ if it is bounded from below and

$$\bar{V}(\pi_t) + \int_0^t g(\pi_s, u(s))\mathrm{d}s \quad , \quad t \geq 0$$

is a $\langle \zeta_t \rangle$-supermartingale for all $\{\pi_t, u(t)\}_{t \geq 0}$ as above. □

Define $V_p : \mathcal{P}(R^d) \to R$ by

$$V_p(\pi) = \sup_{u(\cdot)} \sup_{\tau} E\Big[\int_0^\tau g\Big(\pi_s, u(s)\Big)\mathrm{d}s / \pi_0 = \pi \Big] \quad ,$$

where the first supremum is over all bounded $\langle \zeta_t \rangle$-stopping times τ and the second supremum is over all wide sense admissible $u(\cdot)$. Then the following can be proved exactly as in Lemmas 4.1 and 4.2.

LEMMA 4.3
If $V_p(\cdot) < \infty$, it is the least nonnegative storage function associated with supply rate g and permits the alternative definition:

$$V_p(\pi) = \sup_{u(\cdot)} \sup_{t \geq 0} E\Big[\int_0^t g\Big(\pi_s, u(s)\Big)\mathrm{d}s / \pi_0 = \pi \Big]$$

where the outer supremum is over all wide sense admissible controls. □

4.3 Connections to Ergodic Control

Let $k \in C_b(R^d \times U)$. The ergodic control problem seeks to maximize over admissible $u(\cdot)$ the reward

$$\limsup_{t \to \infty} \frac{1}{t} \int_0^t E\Big[k\Big(X(s), u(s)\Big)\Big]\mathrm{d}s. \tag{4.5}$$

Likewise, the ergodic control problem under partial observations is to maximize over all wide sense admissible $u(\cdot)$ the above reward, rewritten as

$$\limsup_{t \to \infty} \frac{1}{t} \int_0^t E\Big[\hat{k}\Big(\pi_s, u(s)\Big)\Big]\mathrm{d}s,$$

where $\hat{k}(\mu, u) = \mu\Big(k(\cdot, u)\Big)$, $\mu \in \mathcal{P}(R^d)$, $u \in U$. Under suitable conditions, this problem can be shown to have an optimal stationary solution ([2], Chapter VI, [1]) with $u(\cdot)$ a Markov control. If $\sigma(\cdot)$ is nondegenerate (i.e., the least eigenvalue of $\sigma(\cdot)\sigma(\cdot)^T$ is uniformly bounded away from zero), then one can in fact have, under suitable hypotheses, a Markov control that is optimal for any initial law ([2], Chapter VI). Here our interest is in the 'martingale dynamic programming principle' elucidated in [6], [2], Chapter III, among other places, albeit for cost criteria other than ergodic. Let β (resp. $\hat{\beta}$) denote the optimal costs for the completely observed (resp., partially observed) ergodic control problem.

DEFINITION 4.3
A measurable map $\psi : R^d \to R$ is said to be a value function for the completely observed ergodic control problem if for all $(X(\cdot), u(\cdot))$ satisfying (4.1), the process

$$\psi\Big(X(t)\Big) + \int_0^t \Big[k\Big(X(s), u(s)\Big) - \beta\Big]\mathrm{d}s \quad , \quad t \geq 0 \quad ,$$

is an $\langle \mathcal{F}_t \rangle$-supermartingale and is a martingale if and only if $(X(\cdot), u(\cdot))$ is an optimal pair. □

DEFINITION 4.4
A measurable map $\bar{\psi} : \mathcal{P}(R^d) \to R$ is said to be a value function for the partially observed ergodic control problem if for all $(\pi_t, u(t))$, $t \geq 0$, as in (4.4), the process

$$\bar{\psi}(\pi_t) + \int_0^t \Big[\pi_s\Big(k(\cdot), u(s)\Big) - \hat{\beta}\Big]\mathrm{d}s \quad , \quad t \geq 0 \quad ,$$

is a $\langle \zeta_t \rangle$-supermartingale, and is a $\langle \zeta_t \rangle$-martingale if and only if $\{\pi_t, u(t)\}$, $t \geq 0$, is an optimal pair. □

Proving a martingale dynamic programming principle in either case amounts to exhibiting a ψ (resp. $\bar{\psi}$) satisfying the above. The following lemmas establish a link with the developments of the preceding section.

LEMMA 4.4
(a) If $\psi \geq 0$ is as in Definition (4.3), then ψ is a storage function for $g(\cdot, \cdot) = k(\cdot, \cdot) - \beta$.

(b) If $\bar{\psi} \geq 0$ is as in Definition (4.4), then $\bar{\psi}$ is a storage function for $g(\cdot, \cdot) = \hat{k}(\cdot, \cdot) - \hat{\beta}$.

□

This is immediate from the definitions. Note that if ψ is a value function, so is $\psi + c$ for any scalar c. Thus, in particular, it follows that there is a nonnegative value function whenever there is one that is bounded from below. This is the case for nondegenerate diffusions with 'near-monotone' $k(\cdot, \cdot)$, i.e., $k(\cdot, \cdot)$ satisfying

$$\liminf_{\|x\| \to \infty} \inf_u k(u, u) > \beta \quad .$$

See [2],Chapter VI for details.

Going in the other direction, we have

LEMMA 4.5

(a) If $V_c(\cdot) < \infty$ for $g(\cdot, \cdot) = k(\cdot, \cdot) - \beta$, then $V_c(X(t)) + \int_0^t (k(X(s), u(s)) - \beta)\mathrm{d}s$, $t \geq 0$, is an $\langle \mathcal{F}_t \rangle$-supermartingale for all $(X(\cdot), u(\cdot))$ as in (4.1). Furthermore, if $(X(\cdot), u(\cdot))$ is a stationary optimal solution and $V_c(X(t))$ is integrable under this stationary law, then the above process is in fact a martingale.

(b) If $V_p(\cdot) < \infty$ for $g(\cdot,\cdot) = \hat{k}(\cdot,\cdot) - \hat{\beta}$, then $V_p(\pi_t) + \int_0^t (\hat{k}(\pi_s, u(s)) - \hat{\beta})\mathrm{d}s$, $t \geq 0$, is a $\langle \zeta_t \rangle$-supermartingale for all $(\pi_t, u(t))$, $t \geq 0$, as in (4.4). Furthermore, if $(\pi_t, u(t))$, $t \geq 0$, is a stationary optimal solution and $V_p(\pi_t)$ is integrable under this stationary law, then the above process is in fact a martingale.

□

Proof We prove only (a), the proof of (b) being similar. The first claim is immediate. For stationary optimal $(X(\cdot), u(\cdot))$,

$$E\Big[V\Big(X(0)\Big)\Big] \geq \int_0^t \Big[E\Big[k\Big(X(s), u(s)\Big)\Big] - \beta\Big]\mathrm{d}s + E\Big[V\Big(X(t)\Big)\Big] \quad .$$

Hence

$$0 \geq E\Big[k\Big(X(t), u(t)\Big)\Big] - \beta \quad .$$

But since $(X(\cdot), u(\cdot))$ are stationary, the corresponding reward (4.5) in fact equals $E[k(X(t), u(t))]$. Since it is optimal, this equals β, so equality must hold throughout, which is possible only if

$$V\Big(X(t)\Big) + \int_0^t \Big[k\Big(X(s), u(s)\Big) - \beta\Big]\mathrm{d}s \quad , \quad t \geq 0$$

is in fact an $\langle \mathcal{F}_t \rangle$-martingale. □

What we have established is the fact that storage functions are candidate value functions and vice versa, at least for the situations where the latter are known to be bounded from below. In cases where this is possible, we thus have an explicit stochastic representation for the value function of ergodic control. While an explicit stochastic representation, albeit a different one, was available for the completely observed control problem (see [2], p. 161), its counterpart for partial observations was not available. In the foregoing, however, there is little difference in the way we handle complete or partial observations.

Recall also that the usual approach for arriving at value functions for ergodic control is to consider the vanishing discount limit of suitably renormalized value functions for the associated infinite horizon discounted cost control problems. This limit is often difficult to justify and has been done under suitable hypotheses for completely observed ergodic control in [2], Chapter VI, and under rather restrictive conditions for partially observed ergodic control in [3, 4, 5]. Use of the storage function approach allows us to directly define a candidate value function V_c or V_p as above. The task then is to show that they are finite for the problem at hand. For linear stochastic differential equations describing the controlled process and noisy linear observations this can be done and a theory analogous to that of Willems [8, 9] can be developed. It is worth noting that V_c defined above for $g(\cdot,\cdot) = k(\cdot,\cdot) - \gamma$ would certainly be finite for $\gamma > \beta$ and $+\infty$ for $\gamma < \beta$, thus $\gamma = \beta$ is the 'critical' case.

It would be interesting to investigate the relationship of these ideas to that of Rantzer [10].

Acknowledgments

Vivek S. Borkar's research has been partially supported by a grant for 'Nonlinear Studies' from Indian Space Research Organization and Defence Research and Development Organization, Govt. of India, administered through Indian Institute of Science, Bangalore.

Sanjoy K. Mitter's research has been supported by the NSF-KDI Grant ECS-9873451 and by the Army Research Office under the MURI Grant: Data Fusion in Large Arrays of Microsensors DAAD19-00-1-0466.

4.4 References

[1] A. G. Bhatt, V. S.Borkar: Occupation measures for controlled Markov processes: characterization and optimality. The Annals of Probability, vol. 24, 1996, p. 1531-1562.

[2] V. S. Borkar: Optimal Control of Diffusion Processes. Pitman Research Notes in Mathematics No. 203, Longman Scientific and Technical, Harlow, England.

[3] V. S. Borkar: The value function in ergodic control of diffusion processes with partial observations. Stochastics and Stochastic Reports, vol. 67, 1999, p. 255-266.

[4] V. S. Borkar: The value function in ergodic control of diffusion processes with partial observations II. Applicationes Mathematicae, vol. 27, 2000, p. 455-464 (Correction note in *ibid.*, vol. 28, 2001, p. 245-246).

[5] V. S. Borkar: Dynamic programming for ergodic control with partial observations. To appear in 'Stochastic Processes and Their Applications'.

[6] M. H. A. Davis, P. P. Varaiya: Dynamic programming conditions for partially observable stochastic systems. SIAM Journal of Control and Optimization, Vol. 11, 1973, p. 226-261.

[7] W. H. Fleming, E. Pardoux: Optimal control of partially observed diffusions. SIAM Journal of Control and Optimization, vol. 20, 1982, p. 261-285.

[8] J.C. Willems: Dissipative Dynamical Systems, Part I: General Theory. Archive for Rational Mechanics and Analysis, Vol. 45, No. 5, 1972, p. 321-351 (Springer-Verlag, 1972).

[9] J.C. Willems: Dissipative Dynamical Systems, Part II: Linear Systems with Quadratic Supply Rates. Archive for Rational Mechanics and Analysis, Vol. 45, No. 5, 1972, p. 352-393 (Springer-Verlag, 1972).

[10] A. Rantzer: A Dual to Lyapunov's Stability Theorem, Systems and Control Letters, Vol. 42, No. 3, 2001, p. 161-168.

5

Internal Model Based Design for the Suppression of Harmonic Disturbances

Christopher I. Byrnes *David S. Gilliam* *Alberto Isidori*
Yutaka Ikeda *Lorenzo Marconi*

Abstract

Our interest in suppression of harmonic disturbances arose in the development of feedback control strategies for next generation aircraft. The control objective is to track a prescribed trajectory while suppressing the disturbance produced by a harmonic exogenous system. This is a slight modification of the standard problem of output regulation, in which the reference trajectory itself is also assumed to be generated by an exosystem. As part of an on going research effort, we are developing a solution to the problem for a nonlinear system which incorporates both the rigid body dynamics and certain aerodynamic states. In this paper, we illustrate our use of the internal model principle to solve this problem for continuous-time linear systems. Interestingly, the internal model based controller design leads to a Linear Matrix Inequlaity (LMI) constraint on the design parameters, yielding a convex problem which is easily solved.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 51-70, 2003.
© Springer-Verlag Berlin Heidelberg 2003

5.1 Introduction

Anders Lindquist has had tremendous impact on a variety of research fields in systems, control and signal processing. This paper dovetails with one of these contributions, his joint work with Vladimir Yacubovitch [5]-[9] on the suppression of the effect of harmonic disturbances on the regulated output of a system to be controlled. In this series of papers, controllers which are optimal in an LQ sense are derived for the suppression of harmonic disturbances for discrete-time, finite-dimensional, multivariable linear systems. These controllers are shown to enjoy some of the robustness features concomitant with an LQ optimization scheme.

This work can be applied in a variety of settings, for example in active noise control which was studied in [11] for scalar-input scalar-output systems using a minimum variance performance measure. The suppression of harmonic disturbances also arises in the active control of vibrations in helicopters, where a continuous-time model is used in the context of LQ control and Kalman filtering.

Our interest in suppression of harmonic disturbances arose in the development of feedback control strategies for next generation aircraft. In particular, for certain aircraft it is known that the wings exhibit slightly damped flexible behavior which, when excited by a wind gust, produce a disturbance which additively corrupts some important velocity measurements. The frequencies of the slightly damped (exogeneous) signal generator are known, but the wind gust may be modeled as a dirac delta function which forces the exogeneous system into an unknown initial state. The control objective is to track a prescribed trajectory while suppressing the disturbance produced by the exogenous system. This is a slight modification of the standard problem of output regulation, in which the reference trajectory itself is also assumed to be generated by an exosystem. As we shall show, this modification is easily accomodated within the present framework of output regulation. However, since output regulation ensures some asymptotically stability of the error while the tracking and disturbance rejection problem requires a finite time horizon in practice, we have assumed the exosystem is undamped and consists of a finite bank of harmonic oscillators.

As part of an on going research effort, we are developing a solution to the problem for a nonlinear system which incorporates both the rigid body dynamics and certain aerodynamic states. In this paper, we illustrate our use of the internal model principle to solve this problem for continuous-time linear systems. Specifically, in Section 5.2 we introduce the basics of internal model based design for a scalar-input, scalar-output system. In Section 5.3, we illustrate this design for two particular systems where the desired trajectory idealizes a take-off and landing maneuver. In Section 5.4 we give some brief remarks about the internal model principle. One remarkable feature of this principle is its ability to systematically produce classical design approaches in a variety of particular control problems. In the case at hand, it produces a notch filter for the transfer function from the disturbance to the tracking error, with notches, or (blocking) zeroes, at the natural frequencies of the exogenous system. Of course, the real power of this method is its ability to produce nonclassical, multivariable designs. Indeed, in Section 5.4 we consider three-input, three-output, minimum phase systems with vector relative degree $(2, 2, 2)$. Interestingly, the internal model based controller design leads to a Linear Matrix Inequlaity (LMI) constraint on the design parameters, yielding

a convex problem which is easily solved.

5.2 The Case of a SISO System

Consider a linear single-input single-output system having relative degree two and asymptotically stable zero dynamics. Any system of this kind can always be put, after a suitable change of coordinates in the state space, in the form

$$\begin{aligned} \dot{z} &= Az + B_1 x_1 + B_2 x_2 \\ \dot{x}_1 &= x_2 \\ \dot{x}_2 &= Cz + D_1 x_1 + D_2 x_2 + Ku \\ y &= x_1 \end{aligned} \tag{5.1}$$

in which $K \neq 0$ and the matrix A is Hurwitz.

It is well-known that single-input single-output systems having relative degree two and stable zero dynamics can always be asymptotically stabilized by means of a proportional-derivative output feedback, i.e. by means of a control law of the form

$$u = k_1 y + k_2 \dot{y} \, .$$

It is also well-known that a proportional-derivative output feedback can be used to solve the problem of having the output $y(t)$ to track a prescribed trajectory $y_{\text{ref}}(t)$, if the latter and its first and second derivatives are available in real-time.

In this section, we describe how the problem of tracking a prescribed trajectory $y_{\text{ref}}(t)$ can be solved, for a system of the form (5.1), in case the information about the output y (i.e. x_1) is noise-free, but the information about the *rate of change* y (i.e. x_2) is *corrupted by harmonic noise*, i.e. if only x_1 and the quantity

$$x_{2,\text{noisy}} = x_2 + d \, , \tag{5.2}$$

where d is a superposition of sinuosoidal functions of time, are available for feedback.

Define

$$\begin{aligned} e_1 &= y - y_{\text{ref}} &&= x_1 - y_{\text{ref}} \\ e_2 &= \dot{y} - \dot{y}_{\text{ref}} &&= x_2 - \dot{y}_{\text{ref}} \end{aligned}$$

in which e_1 is indeed the *tracking error* and e_2 its rate of change. Then, system (5.1) can be rewritten as

$$\begin{aligned} \dot{z} &= Az + B_1(e_1 + y_{\text{ref}}) + B_2(e_2 + \dot{y}_{\text{ref}}) \\ \dot{e}_1 &= e_2 \\ \dot{e}_2 &= Cz + D_1(e_1 + y_{\text{ref}}) + D_2(e_2 + \dot{y}_{\text{ref}}) + Ku - \ddot{y}_{\text{ref}} \, . \end{aligned} \tag{5.3}$$

Let z_{ref} be any solution of

$$\dot{z}_{\text{ref}} = Az_{\text{ref}} + B_1 y_{\text{ref}} + B_2 \dot{y}_{\text{ref}} \tag{5.4}$$

and define

$$e_z = z - z_{\text{ref}} .$$

Then, system (5.3) can be rewritten as

$$\begin{aligned} \dot{e}_z &= Ae_z + B_1 e_1 + B_2 e_2 \\ \dot{e}_1 &= e_2 \\ \dot{e}_2 &= Ce_z + Cz_{\text{ref}} + D_1 e_1 + y_{\text{ref}} + D_2(e_2 + \dot{y}_{\text{ref}}) + Ku - \ddot{y}_{\text{ref}} . \end{aligned} \tag{5.5}$$

Choose now a control of the form

$$\begin{aligned} u &= u(x_1, x_{2,\text{noisy}}, y_{\text{ref}}, \dot{y}_{\text{ref}}, \ddot{y}_{\text{ref}}, z_{\text{ref}}) \\ &= \tfrac{1}{K}[-Cz_{\text{ref}} - (D_1 + k_1)y_{\text{ref}} - D_2\dot{y}_{\text{ref}} \\ &\quad + \ddot{y}_{\text{ref}} - D_1(x_1 - y_{\text{ref}}) - k_2(x_{2,\text{noisy}} - \dot{y}_{\text{ref}}) + v] \end{aligned}$$

in which v is an additional input to be determined later. With this control u, system (5.5) becomes

$$\begin{aligned} \dot{e}_z &= Ae_z + B_1 e_1 + B_2 e_2 \\ \dot{e}_1 &= e_2 \\ \dot{e}_2 &= Ce_z - k_1 e_1 - \bar{k}_2 e_2 - k_2 d + v , \end{aligned} \tag{5.6}$$

where

$$\bar{k}_2 = k_2 - D_2 .$$

It is known that $d(t)$, which is a superposition of sinusoidal functions, can be put in the form

$$d(t) = \Psi^T w(t)$$

with $w(t)$ solution of the homogeneous equation

$$\dot{w}(t) = Sw(t) \tag{5.7}$$

in which S is a matrix of the form

$$S = \begin{pmatrix} S_1 & 0 & \cdot & 0 \\ 0 & S_2 & \cdot & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdot & S_r \end{pmatrix}$$

and

$$S_i = \begin{pmatrix} 0 & \omega_i \\ -\omega_i & 0 \end{pmatrix} .$$

Under our assumption that the zero dynamics is either vacuous or is asymptotically stable, we know that any transmission zeros of the plant to be controlled lie in the left half plane. Since the eigenvalues of the exosystem (5.7) that generate

the disturbance lie on the imaginary axis, we see that the standard nonresonance conditions for solution of the regulator problem are satisfied.

Matters being so, the output regulation problem for $y_r = 0$ is solvable, when the additional input v is chosen as the output of an *internal model* of the form

$$\begin{aligned} \dot{\xi} &= S\xi - Gk_1 e \\ v &= \Psi^T \xi \end{aligned}$$

where G is a vector of free design parameters. We claim that this also yields a solution for nontrivial y_r. Indeed, introduction of v yields a closed loop system which, changing the variable ξ into

$$\chi = \xi - k_2 w$$

is described by set of equations

$$\begin{aligned} \dot{e}_z &= Ae_z + B_1 e_1 + B_2 e_2 \\ \dot{\chi} &= S\chi - Gk_1 e_1 \\ \dot{e}_1 &= e_2 \\ \dot{e}_2 &= Ce_z + \Psi^T \chi - k_1 e_1 - \bar{k}_2 e_2 \, . \end{aligned} \tag{5.8}$$

At this point, it remains to show that, if k_1, k_2 (or, what is the same, $\bar{k}_2$), and G is appropriately chosen, system (5.8) is asymptotically stable. If this is the case, in fact, we have in particular

$$\lim_{t\to\infty} e_1(t) = 0 \, ,$$

which is the required tracking goal, regardless of the fact that the measurement of $\dot{y}$ is corrupted by noise.

To prove our stability result, set

$$\Psi^T = (\, \Psi_1^T \quad \Psi_2^T \quad \cdots \quad \Psi_r^T \,)$$

and

$$G^T = (\, G_1^T \quad G_2^T \quad \cdots \quad G_r^T \,)$$

where the partitions indicated are consistent with those of S.

Then, the following result holds.

PROPOSITION 5.1
Consider system (5.8). Choose $G_1, \ldots, G_r$ in such a way that

$$\Psi_i^T S_i^{-1} G_i > 0 \, , \qquad i = 1, \ldots, r \tag{5.9}$$

$$\sum_{i=1}^{r} \Psi_i^T S_i^{-1} G_i < 1 \tag{5.10}$$

which is always possible. If $k_1 > 0, \bar{k}_2 > 0$ are sufficiently large, the system is asymptotically stable. □

The proof of this Proposition is a straightforward consequence of the following Lemma, proven for instance in [2].

LEMMA 5.1
Consider the matrix

$$J = \begin{pmatrix} A & 0 & \cdot & 0 & 0 & B \\ 0 & S_1 & \cdot & 0 & 0 & P_1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & S_r & 0 & P_r \\ 0 & 0 & \cdot & 0 & 0 & P_0 \\ C & Q_1 & \cdot & Q_r & Q_0 & -k \end{pmatrix}. \tag{5.11}$$

If

$$P_i Q_i < 0 \qquad \text{for all } i = 0, \ldots, r \tag{5.12}$$

there exists k^* such that, if $k > k^*$, the matrix J is Hurwitz. □

To prove the Proposition, we change coordinates in system (5.8) using

$$\begin{aligned} \tilde{e}_z &= e_z + A^{-1}B_1 e_1 \\ \tilde{\chi} &= \chi - k_1 S^{-1} G e_1 \end{aligned}$$

which yields

$$\begin{aligned} \dot{\tilde{e}}_z &= A\tilde{e}_z + (A^{-1}B_1 + B_2)e_2 \\ \dot{\tilde{\chi}} &= S\tilde{\chi} - k_1 S^{-1} G e_1 \\ \dot{e}_1 &= e_2 \\ \dot{e}_2 &= C\tilde{e}_z + \Psi^T \tilde{\chi} - (CA^{-1}B_1 + k_1(1 - \Psi^T S^{-1} G))e_1 - \bar{k}_2 e_2 \,, \end{aligned} \tag{5.13}$$

i.e. a system of the form

$$\dot{x} = Jx$$

with J a matrix having the same structure as the matrix J in the Lemma if we set

$$\begin{aligned} B &= (A^{-1}B_1 + B_2) \\ Q_i &= \Psi_i^T && i = 1, \ldots, r \\ P_i &= -k_1 S_i^{-1} G_i && i = 1, \ldots, r \\ Q_0 &= -(CA^{-1}B_1 + k_1(1 - \Psi^T S^{-1} G)) \\ P_0 &= 1 \\ k &= \bar{k}_2 \end{aligned}$$

Thus, conditions $Q_i P_i < 0$, for $i = 1, \ldots, r$, of the Lemma become

$$k_1 \Psi_i^T S_i^{-1} G_i > 0$$

while condition $Q_0 P_0 > 0$ of the Lemma becomes

$$CA^{-1}B_1 + k_1(1 - \Psi^T S^{-1} G) = CA^{-1}B_1 + k_1\Big(1 - \sum_{i=1}^{r} \Psi_i^T S_i^{-1} G_i\Big) > 0 \,.$$

Indeed, if G is such that (5.9) and (5.10) hold, the conditions of the Lemma can be met for large $k_1 > 0$ and, consequently, for large $\bar{k}_2 > 0$ the matrix J is Hurwitz.

5.3 A Numerical Example of Take-Off and Landing

In this section we consider some specific numerical examples. In our first example we consider the case in which there is no z, i.e. no zero dynamics, and that $D_1 = D_2 = 0$, $K = 1$ so that the plant reduces to the double integrator. For the second example we consider a one dimensional stable zero dynamics.

For both examples we assume that disturbance d is given as a sum of sinusoids with two different frequencies,

$$d(t) = M_1 \sin(\alpha_1 t + \varphi_1) + M_2 \sin(\alpha_2 t + \varphi_2).$$

In this case we can take

$$S_1 = \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} 0 & 10 \\ -10 & 0 \end{pmatrix}, \quad S = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}$$

and consider

$$\dot{w} = Sw, \quad w(0) = w_0.$$

Then we choose Ψ_1^T, Ψ_2^T and $\Psi^T = \begin{pmatrix} \Psi_1^T & \Psi_2^T \end{pmatrix}$ so that $d(t) = \Psi^T w(t)$.

EXAMPLE 5.1
In our numerical examples we have chosen $\alpha_1 = 2$, $\alpha_2 = 10$,

$$\Psi_1^T = \begin{pmatrix} 0 & 2 \end{pmatrix}, \qquad \Psi_2^T = \begin{pmatrix} 0 & 2 \end{pmatrix}$$

$$w_0 = \begin{pmatrix} 0 \\ 25 \\ 0 \\ 50 \end{pmatrix}$$

so that

$$d(t) = 50\cos(2t) + 100\cos(10t).$$

Next we choose G_1, G_2, k_1 and k_2 so that the conditions of Proposition 5.1 are satisfied.

In the numerical example we have set $k_1 = 10$ and

$$G_1 = \begin{pmatrix} 0.2 \\ -0.2 \end{pmatrix} \qquad G_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

which gives

$$\Psi_1^T S_1^{-1} G_1 = 0.2, \qquad \Psi_2^T S_2^{-1} G_2 = 0.2$$

and

$$\Psi_1^T S_1^{-1} G_1 + \Psi_2^T S_2^{-1} G_2 = .4 < 1$$

so the conditions of Proposition 5.1 are fulfilled.

We proceed to choose $k_2 > 0$ so that the matrix J in (5.11) is Hurwitz. Our choice of k_2 will be based on a root locus design. Since we do not have a z component

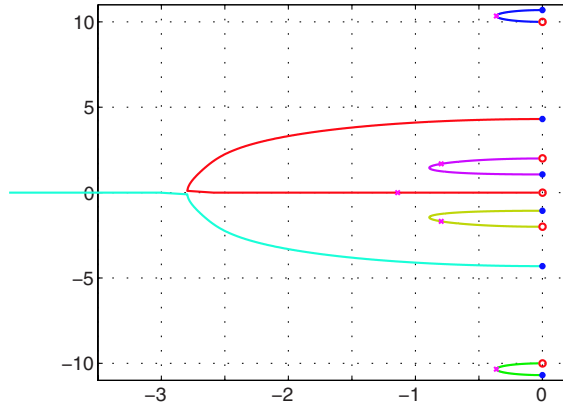


Figure 5.1 plot of root locus for $\alpha_1 = 2$, $\alpha_2 = 10$, $k_1 = 30$, $k_2 = 9.1205$

in this example the matrix J can be reduced by deleteing the first column and first row. Then we set $k_2 = k = 0$ and define the resulting matrix to be J_0, i.e.,

$$J_0 = \begin{pmatrix} S_1 & \cdot & 0 & 0 & P_1 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & S_r & 0 & P_r \\ 0 & \cdot & 0 & 0 & P_0 \\ Q_1 & \cdot & Q_r & Q_0 & 0 \end{pmatrix}.$$

Then we define matrices

$$\mathcal{B}^T = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1), \quad C = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1)$$

and consider the system $\{J_0, \mathcal{B}, C\}$ with feedback law $u = -k_2 y$ and transfer function

$$C(sI - J_0)^{-1}\mathcal{B}.$$

Notice that

$$J = J_0 - k_2 \mathcal{B} C,$$

and, more importantly, the eigenvalues of J are the closed loop poles of this system. If $k_1 > 0$ our Proposition predicts that for large values of k_2 the matrix J is Hurwitz. For $k_1 = 10$ we plot the locus of the roots of the characteristic polynomial of J, viewing k_2 as a gain parameter. For large k_2, five of the 6 roots should approach five (zeros) on the imaginary axis, at $0, \pm 2i, \pm 10i$, while the 6-th one goes to $-\infty$.

In this example we take a simplified model of aircraft take-off and landing with a reference signal given by

$$y_{\text{ref}}(t) = \begin{cases} 2t, & 0 < t < 50 \\ 100, & 50 < t < 100 \\ 300 - 2t, & 100 < t < 150 \end{cases}.$$

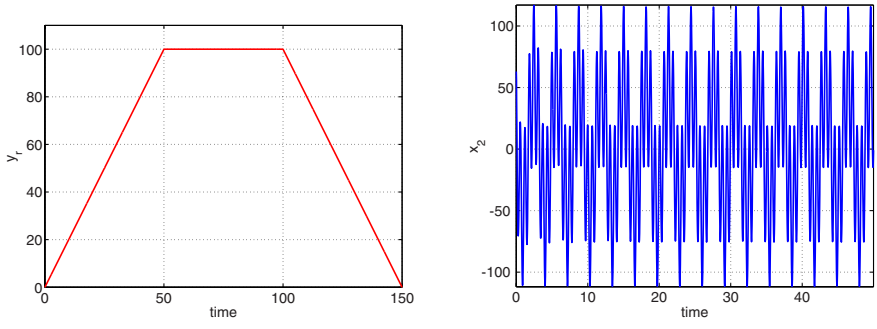


Figure 5.2 (left) plot of y_{ref} , (right) plot of $x_{2,n} = x_2 + d$

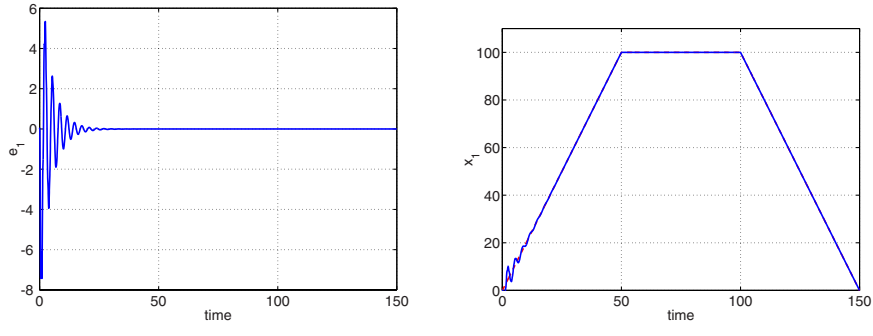


Figure 5.3 (left) plot of $e_1(t) = x_1(t) - y_{\text{ref}}(t)$, (right) plot of x_1 and y_{ref}

Notice that y_{ref} is not differentiable at $t = 50$ and $t = 100$ as it would actually be in practice. Indeed we have not actually used y_{ref} as given above but rather we have used a Hermite interpolation to round the corners at these points. Since it serves no particular good to produce these more complicated formulas we do not present them here.

□

EXAMPLE 5.2

In our second numerical example consider a system with a one dimensional stable zero dynamics. In this case we have set $A = -1$, $C = 1$, $D_1 = 1$ and $D_2 = 1$. The relative values of B_1 and B_2 play a large role in the location of the closed loop poles so we present a "root locus plot" in two different cases: $B_1 = 1/4$, $B_2 = 1/2$ and $B_1 = 1/2$, $B_2 = 1/4$. Also for this example we compute the reference zero dynamics trajectory z_{ref} satisfying (5.4) for a given initial condition $z_{\text{ref}}(0) = 5$. Once again we have set $\alpha_1 = 2$, $\alpha_2 = 10$, and we have taken

$$\Psi_1^T = \begin{pmatrix} 1 & 1 \end{pmatrix}, \qquad \Psi_2^T = \begin{pmatrix} 1 & 1 \end{pmatrix}$$

$$w_0 = \begin{pmatrix} 0 \\ 25 \\ 0 \\ 10 \end{pmatrix}$$

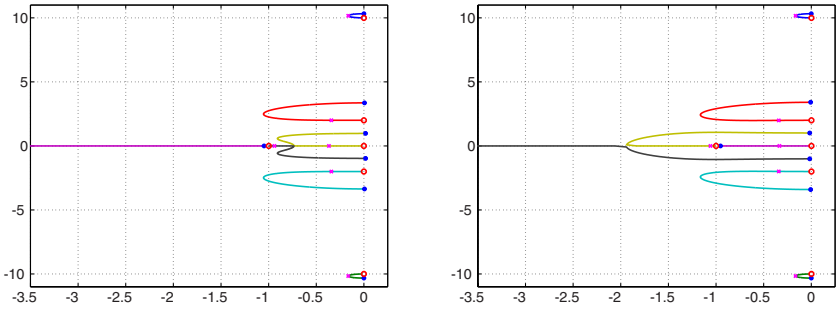


Figure 5.4 (left) $B = (A^{-1}B_1 + B_2) > 0$, $k_1 = 15$, (right) $B = (A^{-1}B_1 + B_2) < 0$, $k_1 = 15$

so that

$$d(t) = 25\sqrt{2}\sin(2t + \pi/4) + 10\sqrt{2}\sin(10t + \pi/4).$$

Next we choose G_1, G_2, k_1 and k_2 so that the conditions of Proposition 5.1 are satisfied.

In the numerical example we have set $k_1 = 20$ and

$$G_1 = \begin{pmatrix} 0.2 \\ -0.2 \end{pmatrix} \qquad G_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

which gives

$$\Psi_1^T S_1^{-1} G_1 = 0.2, \qquad \Psi_2^T S_2^{-1} G_2 = 0.2$$

and

$$\Psi_1^T S_1^{-1} G_1 + \Psi_2^T S_2^{-1} G_2 = .4 < 1$$

so the conditions of Proposition 5.1 are fulfilled.

Just as in the previous example we choose $k_2 > 0$ so that the matrix J in (5.11) is Hurwitz. In the present case more care must be used in making our choice for k_2 as is seen in the root locus plots in Figure 5.4. Just as in Example 5.1, if we first set $k_2 = k = 0$ in J and define the resulting matrix to be J_0, i.e.,

$$J_0 = \begin{pmatrix} A & 0 & \cdot & 0 & 0 & B \\ 0 & S_1 & \cdot & 0 & 0 & P_1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & S_r & 0 & P_r \\ 0 & 0 & \cdot & 0 & 0 & P_0 \\ C & Q_1 & \cdot & Q_r & Q_0 & 0 \end{pmatrix}.$$

Then we define matrices

$$\mathcal{B}^T = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1), \quad C = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1)$$

and consider the system $\{J_0, \mathcal{B}, C\}$ with feedback law $u = -k_2 y$ and transfer function

$$C(sI - J_0)^{-1}\mathcal{B}.$$

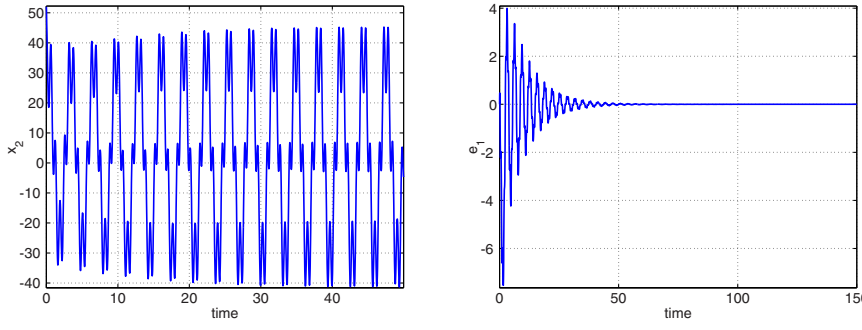


Figure 5.5 (left) $x_{2,n} = x_2 + d$, (right) $e_1(t) = x_1(t) - y_{\text{ref}}(t)$

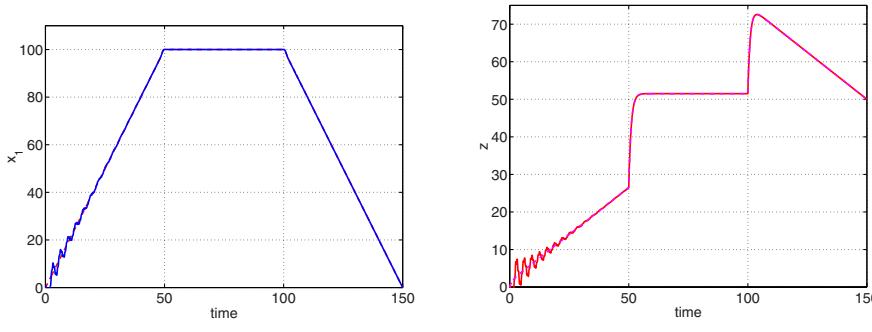


Figure 5.6 (left) x_1 and y_{ref} , (right) z

Notice that

$$J = \mathcal{J}_0 - k_2 \mathcal{B}C,$$

and, more importantly, the eigenvalues of J are the closed loop poles of this system.

Unlike the first example more care must be exercised in choosing k_1 and k_2. As an example, if we choose $k_1 > 0$, set $A = -1$, $C = 1$, then the the sign of the real parts of the eigenvalues of J_0 depend on whether $B = (A^{-1}B_1 + B_2)$ is greater than or less than zero.

If $k_1 > 0$ our Proposition still predicts that for large values of k_2 the matrix J is Hurwitz. For $k_1 = 15$ and for $B_1 = 1/4$, $B_2 = 1/2$, $B = (A^{-1}B_1 + B_2) > 0$ for which the some of the open loop poles are in the right half plane while for $k_1 = 15$ and for $B_1 = 1/2$, $B_2 = 1/4$, $B = (A^{-1}B_1 + B_2) < 0$ all open loop poles are in the left half plane. For large k_2, five of the 7 closed loop poles approach the five (zeros) on the imaginary axis, at $0, \pm 2i, \pm 10i$, while the 6-th one goes to $-\infty$ and the 7-th approaches the zero at -1. In this example we take a simplified model of aircraft take-off and landing with a reference signal given by

$$y_{\text{ref}}(t) = \begin{cases} 2t, & 0 < t < 50 \\ 100, & 50 < t < 100 \\ 300 - 2t, & 100 < t < 150 \end{cases} .$$

□

5.4 Remarks on the Internal Model Principle

One of the remarkable features of the Internal Model Principle is its ability to generate, in a systematic fashion, classical control designs. More importantly, it can also be used in multivariable, nonlinear and infinite-dimensional contexts.

As a first illustration of of this point, consider a relative degree 1, minimum phase scalar-input scalar-output system

$$\begin{aligned}\dot{z} &= Az + By\\ \dot{y} &= Cz + Dy + ku\end{aligned}$$

where $y_{\text{noisy}} = y + d$ with $d(t)$ being an unknown constant. Setting

$$e = y - y_{\text{ref}}$$

we obtain the relative degree one analog of (5.3)-(5.6). In this case, the constant disturbance takes the form

$$d = \psi w$$

where w is the state of the3 exogenous system

$$\dot{w} = 0.$$

This generates the internal model

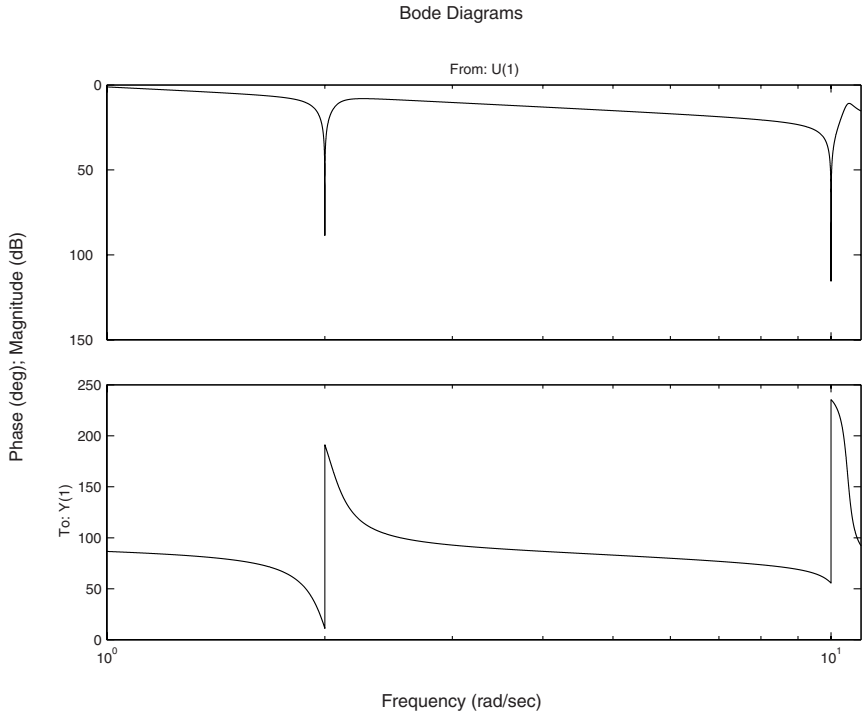
$$\begin{aligned}\dot{\xi} &= -Ge\\ v &= \psi e\end{aligned}$$

which is a proportional integral (PI) controller. Stability of the closed loop system

$$\begin{aligned}\dot{e}_z &= Ae_z + Be\\ \dot{\chi} &= -Ge\\ \dot{e} &= Ce_z + \psi\chi - ke\end{aligned}$$

is guaranteed, whenever $\psi G > 0$ by Lemma 2.1 (see [2]).

As a second illustration, consider the system described in Example 5.1. The disturbance is a sum of two sinusoids having frequencies $\alpha_1 = 2$ and $\alpha_2 = 10$. The charts below depict the Bode magnitude and phase plots of the closed-loop transfer function from the disturbance to the tracking error, obtained by using an internal model based design. The magnitude plot is, of course, that of a typical notch filter having notches (or transmission zeroes) at precisely $2i$ and $10i$, reflecting the fact that the filter absorbs sinusoids at those frequencies.



5.5 The Case of a MIMO System

In the following we briefly sketch how the analysis developed in section 5.3 can be easily extended to deal also with multi inputs-multi outputs systems. We consider the class of minimum-phase linear systems with three inputs and three outputs with vector relative degree $[2, 2, 2]$. As is well-known, this class of systems can always be put, after a suitable change of coordinates, in the normal form

$$\begin{aligned} \dot{z} &= Az + B_1 x_1 + B_2 x_2 \\ \dot{x}_1 &= x_2 \\ \dot{x}_2 &= Cz + D_1 x_1 + D_2 x_2 + Ku \\ y &= x_1 \end{aligned} \tag{5.14}$$

where $z \in \mathbb{R}^r$,

$$x_1 = \begin{pmatrix} x_{11} & x_{12} & x_{13} \end{pmatrix}^T \in \mathbb{R}^3 \qquad x_2 = \begin{pmatrix} x_{21} & x_{22} & x_{23} \end{pmatrix}^T \in \mathbb{R}^3 ,$$

$u \in \mathbb{R}^3$ is the vector of control inputs and $y \in \mathbb{R}^3$ of controlled outputs, with the matrix A which is Hurwitz and the high frequency gain matrix $K \in \mathbb{R}^3 \times \mathbb{R}^3$ which is non singular. Our goal is to force the output vector $\mathbf{y}$ to asymptotically track a vector of known reference signals

$$y_{\text{ref}} = \begin{pmatrix} y_{\text{ref},1} & y_{\text{ref},2} & y_{\text{ref},3} \end{pmatrix}^T$$

with the design of a proportional-derivative output feedback control law. Similarly to section 5.4 the challenging aspect of this apparently simple tracking problem is given by the presence of additive noise on the velocity measures. More precisely we suppose, as above, that the measure of the position vector $\boldsymbol{x}_1$ is noise-free, but the measured velocity vector is corrupted by additive harmonic noise, namely

$$\boldsymbol{x}_{2,\text{noisy}}(t) = \boldsymbol{x}_2(t) + \boldsymbol{d}(t) \qquad \text{with} \qquad \boldsymbol{d}(t) = \begin{pmatrix} d_1(t) & d_2(t) & d_3(t) \end{pmatrix}^T$$

where the signals d_i, $i = 1, 2, 3$, are given as superposition of sinusoidal functions of time. In particular the signals $d_i(t)$, $i = 1, 2, 3$, are thought as generated by three disjoint exosystems of the form

$$d_i(t) = \Psi_i^T w_i(t) \qquad \text{with} \qquad \dot{w}_i(t) = S_i w_i(t) \qquad w_i \in \mathbb{R}^{2r_i}$$

where S_i is a block-diagonal square matrix of dimension $2r_i$ which describes a bank of r_i oscillators at different frequencies.

By mimicking the design methodology of the SISO case we define by z_{ref} any solution of

$$\dot{z}_{\text{ref}} = A z_{\text{ref}} + B_1 \boldsymbol{y}_{\text{ref}} + B_2 \dot{\boldsymbol{y}}_{\text{ref}}$$

and we choose the preliminary control law

$$\boldsymbol{u} = K^{-1} \left(-C z_{\text{ref}} - D_1 \boldsymbol{x}_1 - D_2 \dot{\boldsymbol{y}}_{\text{ref}} + \ddot{\boldsymbol{y}}_{\text{ref}} + \boldsymbol{v} \right) \tag{5.15}$$

where $\boldsymbol{v}$ is a vector of residual control inputs. Defining

$$\begin{array}{rcccl} \boldsymbol{e}_1 & = & \boldsymbol{y} - \boldsymbol{y}_{\text{ref}} & = & \boldsymbol{x}_1 - \boldsymbol{y}_{\text{ref}} \\ \boldsymbol{e}_2 & = & \dot{\boldsymbol{y}} - \dot{\boldsymbol{y}}_{\text{ref}} & = & \boldsymbol{x}_2 - \dot{\boldsymbol{y}}_{\text{ref}} \end{array}$$

and $\tilde{z} = z - z_{\text{ref}}$, simple computations show that system (5.14) under the feedback (5.15), in the new error coordinates, reads as

$$\begin{array}{rcl} \dot{\tilde{z}} & = & A\tilde{z} + B_1 \boldsymbol{e}_1 + B_2 \boldsymbol{e}_2 \\ \dot{\boldsymbol{e}}_1 & = & \boldsymbol{e}_2 \\ \dot{\boldsymbol{e}}_2 & = & C\tilde{z} + D_2 \boldsymbol{e}_2 + \boldsymbol{v} \end{array} \tag{5.16}$$

The design of the vector of inputs $\boldsymbol{v}$ able to asymptotically stabilize the system (5.16), and hence to solve the tracking problem, can be easily carried out by means of a three-steps procedure in which, at each step, just one control variable v_i is designed, following exactly the design methodology illustrated in the SISO case.

To sketch how this procedure can be carried out consider, for the first control input v_1, the following dynamic law

$$\begin{array}{rcl} \dot{\xi}_1 & = & S_1 \xi_1 - G_1 k_{11} e_{11} \\ v_1 & = & -k_{11}(x_{11} - y_{\text{ref},1}) - k_{21}(x_{21,\text{noisy}} - \dot{y}_{\text{ref},1}) + \Psi_1 \xi_1 \\ & = & -k_{11} e_{11} - k_{21} e_{21} + \Psi_1 \xi_1 - k_{21} \Psi_1 w_1 \end{array} \tag{5.17}$$

in which k_{11}, k_{21} and the entries of G_1 are design parameters. Simple computations show that, having defined

$$\chi_1 = \xi_1 - k_{21} w_1 \,,$$

system (5.16) with the partial control law (5.17) can be rewritten as

$$\begin{aligned} \dot{\bar{z}} &= \bar{A}\bar{z} + \bar{B}_1 \bar{e}_1 + \bar{B}_2 \bar{e}_2 \\ \dot{\bar{e}}_1 &= \bar{e}_2 \\ \dot{\bar{e}}_2 &= \bar{C}\bar{z} + \bar{D}_2 \bar{e}_2 + \bar{v} \end{aligned} \tag{5.18}$$

where

$$\bar{z} = \begin{pmatrix} \tilde{z} & \chi_1 & e_{11} & e_{21} \end{pmatrix}^T \qquad \bar{e}_i = \begin{pmatrix} e_{i,2} & e_{i,3} \end{pmatrix}^T \qquad \bar{v} = \begin{pmatrix} v_2 & v_3 \end{pmatrix}^T$$

the matrix $\bar{A}$ is defined as

$$\tilde{A} = \begin{pmatrix} A & 0 & B_1' & B_2' \\ 0 & S_1 & -G_1 k_{11} & 0 \\ 0 & 0 & 0 & 1 \\ C' & \Psi_1^T & -k_{11} & -(k_{21} - D_2') \end{pmatrix} \tag{5.19}$$

and $\bar{B}_1$, $\bar{B}_2$, $\bar{D}_1$, $\bar{D}_2$, B_1', B_2', C', D_2' are suitably defined matrices.

Now it can be easily realized that it is possible to choose k_{11}, k_{21} and G_1 so that the matrix $\bar{A}$ is Hurwitz. To this end just note that the matrix characterizing system (5.8) has the same structure of (5.19). Hence, by Proposition 5.1, we can claim the existence of a vector G_1 such that for suitably large k_{11} and k_{22}, the matrix $\bar{A}$ is Hurwitz.

From this, once G_1, k_{11} and k_{22} have been fixed, the procedure can be iterated in order to design v_2 and v_3. As a matter of fact the system (5.18) exhibits the same structure of the original error system (5.16), with the state variables $\bar{e}_i$ and $\bar{z}$ which replace e_i and $\tilde{z}$, with the vector of control input $\bar{v}$ which replaces v, and with the Hurwitz matrix $\bar{A}$ which replaces A. This allows us to design also v_2, and in the third step v_3, using the same procedure used for the SISO case and come out with a proportional-derivative stabilizer of the form

$$\begin{aligned} \dot{\xi} &= S\xi - GK_1 e_1 \\ v &= -K_1 e_1 - K_2 e_2 + \Psi\xi \end{aligned} \tag{5.20}$$

with

$$S = \begin{pmatrix} S_1 & 0 & 0 \\ 0 & S_2 & 0 \\ 0 & 0 & S_3 \end{pmatrix} \quad K_1 = \begin{pmatrix} k_{11} & 0 & 0 \\ 0 & k_{12} & 0 \\ 0 & 0 & k_{13} \end{pmatrix} \quad K_2 = \begin{pmatrix} k_{21} & 0 & 0 \\ 0 & k_{22} & 0 \\ 0 & 0 & k_{23} \end{pmatrix}$$

$$G = \begin{pmatrix} G_1 & 0 & 0 \\ 0 & G_2 & 0 \\ 0 & 0 & G_3 \end{pmatrix} \qquad \Psi = \begin{pmatrix} \Psi_1 & 0 & 0 \\ 0 & \Psi_2 & 0 \\ 0 & 0 & \Psi_3 \end{pmatrix} .$$

Hence the overall control law is obtained from the composition of (5.15)and (5.20).

EXAMPLE 5.3
In our 3-dimensional numerical examples we have chosen $\alpha_1 = 2$, $\alpha_2 = 3$, $\alpha_3 = 4$, $\Psi_1^T = \Psi_2^T = \Psi_3^T = (2 \quad 2)$,

$$w_0 = \begin{pmatrix} 0 \\ 25 \\ 0 \\ 50 \end{pmatrix}$$

so that

$$d(t) = \begin{bmatrix} 5(\sin(\alpha_1 t) + \cos(\alpha_1 t) \\ 10(\sin(\alpha_2 t) + \cos(\alpha_2 t) \\ 5(\sin(\alpha_3 t) + \cos(\alpha_3 t) \end{bmatrix}.$$

Next we choose G_1, G_2, G_3 k_1 and k_2 so that the conditions of Proposition 5.1 are satisfied. We have also set $D_1 = D_2 = I_{3\times 3}$.

In the numerical example we have set $k_1 = 10$ and

$$G_1 = \begin{pmatrix} 0.2 \\ -0.2 \end{pmatrix}, \qquad G_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad G_3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

which gives

$$\Psi_1^T S_1^{-1} G_1 = 1/5, \qquad \Psi_2^T S_2^{-1} G_2 = 2/3, \qquad \Psi_3^T S_3^{-1} G_3 = 1/10$$

and

$$\Psi_1^T S_1^{-1} G_1 + \Psi_2^T S_2^{-1} G_2 + \Psi_3^T S_3^{-1} G_3 = \frac{87}{100} < 1$$

so the conditions of Proposition 5.1 are fulfilled.

From the theory developed in Section 5.4 we see that the closed loop system can be written as $\dot{\boldsymbol{X}} = \boldsymbol{J}\boldsymbol{X}$ where

$$\boldsymbol{X} = \begin{bmatrix} \xi \\ e_1 \\ e_2 \end{bmatrix} \in \mathbb{R}^{12},$$

and, just an in Example 5.1, since there is no z component, the matrix $\boldsymbol{J}$ can be reduced by deleteing the first column and first row, to obtain

$$J = \begin{bmatrix} S & -GK_1 & 0_{6\times 3} \\ 0_{3\times 6} & 0_{3\times 3} & I_{3\times 3} \\ \Psi & -K_1 & -(K_2 - D_2) \end{bmatrix}.$$

Since this is a MIMO example we cannot directly appeal to root locus methods but if we set

$$K_2 = k_2 I_{3\times 3}, \quad \text{and set } k = (k_2 - 1)$$

then it is still simple to compute a "root locus type plot" of the closed loop poles as functions of the scalar parameter $k > 0$.

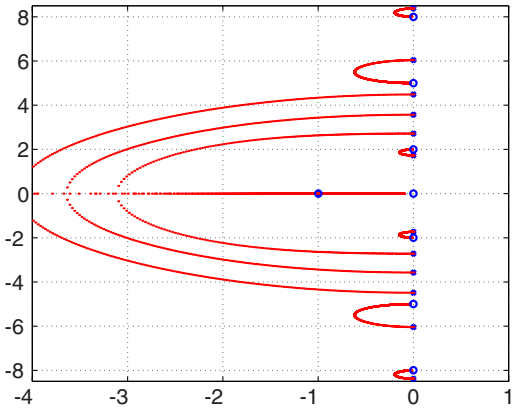


Figure 5.7 plot of root locus

Namely, we define the matrix to be $\boldsymbol{J}_0$ by

$$\boldsymbol{J}_0 = \begin{bmatrix} S & -GK_1 & 0_{6\times 3} \\ 0_{3\times 6} & 0_{3\times 3} & I_{3\times 3} \\ \Psi & -K_1 & 0_{3\times 3} \end{bmatrix}.$$

Then we define matrices

$$\boldsymbol{B}^T = (\,0_{1\times 9} \quad 1 \quad 1 \quad 1\,), \quad \boldsymbol{C} = (\,0_{1\times 9} \quad 1 \quad 1 \quad 1\,)$$

so that

$$\boldsymbol{J} = \boldsymbol{J}_0 - k\boldsymbol{B}\boldsymbol{C}$$

and consider the eigenvalues of $\boldsymbol{J}$ as the closed loop poles of this system.

For $k_1 = 10$ we plot the locus of the roots of the characteristic polynomial of $\boldsymbol{J}$, viewing k_2 (or k) as a gain parameter.

In this example we track a three dimensional reference trajectory

$$y^{\text{ref}}(t) = \left[-10\cos\left(\frac{3\pi t}{100}\right) \quad 10\sin\left(\frac{3\pi t}{100}\right) \quad y_3^{\text{ref}}(t) \right]^T$$

where where

$$y_3^{\text{ref}}(t) = \begin{cases} t, & 0 < t < 40 \\ 40 + 5\sin\left(\dfrac{3\pi(t-40)}{60}\right), & 40 < t < 100 \\ -\dfrac{4}{5}(t-150), & 100 < t < 150 \end{cases}.$$

□

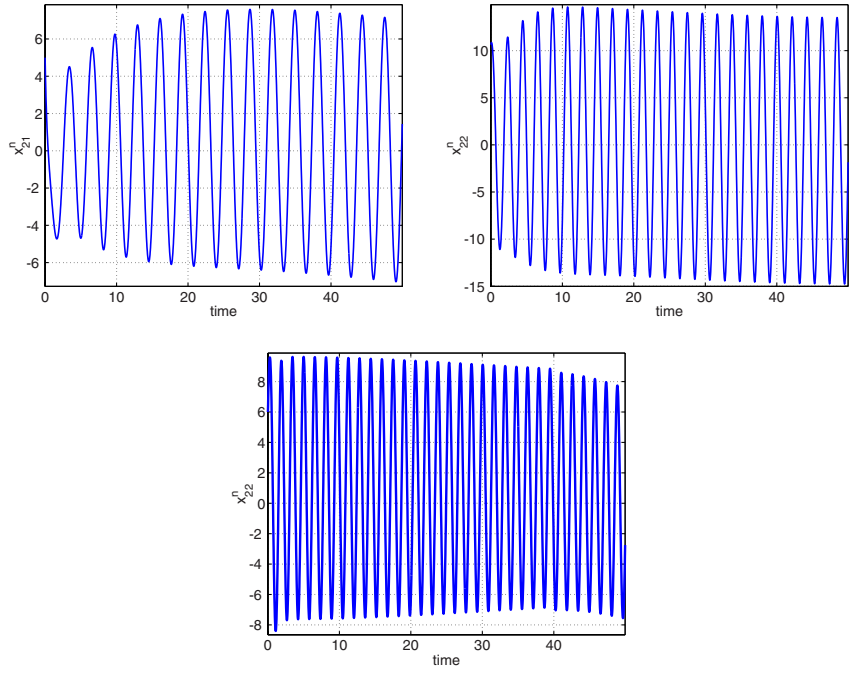


Figure 5.8 $x_{21,n} = x_{21,n} + d_1$, $x_{22,n} = x_{22,n} + d_2$ $x_{23,n} = x_{23,n} + d_3$

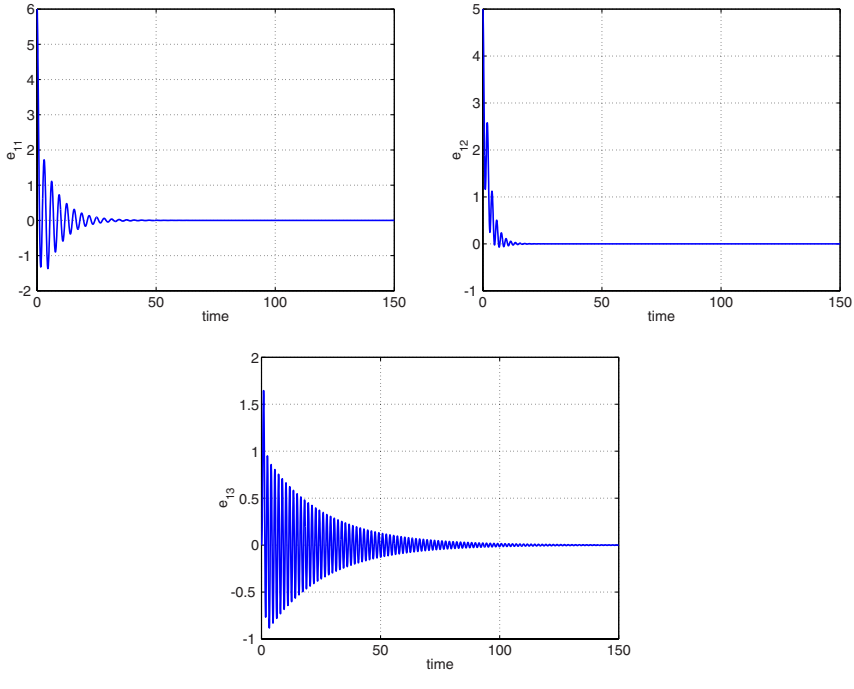


Figure 5.9 $e_{1j}(t) = x_{1j}(t) - y_j^{\text{ref}}(t)$ for $j = 1, 2, 3$

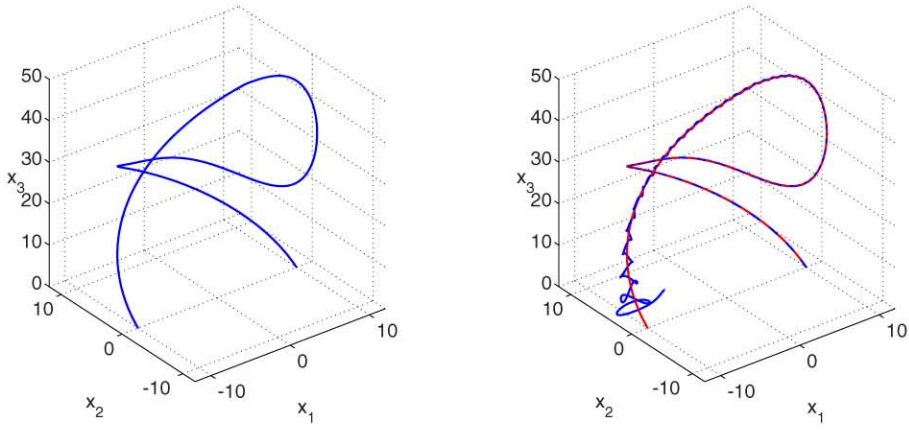


Figure 5.10 y^{ref}, x_1 and y^{ref}

Acknowledgment

Supported in part by the AFOSR and the Boeing Foundation.

5.6 References

[1] S. Bittanti, F. Lorito, S. Strada, "An LQ approach to active control of vibrations in helicopters," *Trans. ASME, J. Dynamical Systems, Measurement and Control*, vol. 118, pp. 482-488,1996.

[2] C.I. Byrnes, A. Isidori, "Bifurcation analysis of the zero dynamics and the practical stabilization of nonlinear minimum-phase systems," *Asian J. of Control*, to appear (2002).

[3] E.J. Davison, A. Goldenberg, "Robust control of a general servomechanism problem: The servo compensator," *Automatica*, vol. 11, pp. 461-471, 1975.

[4] B.A. Francis, W.M. Wonham, "The internal model principle of control theory," *Automatica*, vol. 12, pp. 457-465, 1977.

[5] A. Lindquist, V.A. Yakubovich, "Universal regulators for optimal tracking in discrete-time systems affected by harmonic disturbance," *IEEE Trans. Aut. Contrl.*, vol. 44, No. 9, pp. 1688-1704, 1999.

[6] A. Lindquist, V.A. Yakubovich, "Universal regulators for optimal tracking in linear discrete systems," *Dolk. Akad. Nauk.*, 361: 2, 1998.

[7] A. Lindquist, V.A. Yakubovich, "Universal regulators for optimal damping of forced oscillations in linear discrete systems," *Doklady Mathematics*, 88: 1 1997, pp. 156-159 (in Russian).

[8] A. Lindquist, V.A. Yakubovich, "Optimal damping of forced oscillations by output feedback," *Stochastic Differential and Difference Equations,* I. Csiszár and Gy. Michaletzky, editors, Progress in Systems and Control, vol. 23, Birkhäuser, 1997, pp. 203-231.

[9] A. Lindquist, V.A. Yakubovich, "Optimal damping of forced oscillations in discrete-time systems," *IEEE Trans. Aut. Contrl.*, vol. 42, pp 786-802, 1997.

[10] A.V. Savkin, I.R. Petersen, "Robust control with rejection of harmonic disturbances," IEEE Trans. Automat. Contr., vol. 40, pp. 1968-1971, 1995.

[11] R. Shoureshi, L. Brackney, N. Kubota, G. Batta, "A modern control approach to active noise control," *Trans. ASME J. Dynamical Systems, Measurement and Control*, vol. 115, pp. 673-678, 1993.

[12] V.A. Yakubovich, "A Frequency theorem in control theory," *Sibirskij Mat. Zh.*, vol. 4, pp. 386-419, 1973, (in Russian); English ranslation in *Sibirian Math. J.*

6

Conditional Orthogonality and Conditional Stochastic Realization

Peter E. Caines *R. Deardon* *H. P. Wynn*

Abstract

The concept of conditional orthogonality for the random variables x, y with respect to a third random variable z is extended to the case of a triple x, y, z of processes and is shown to be equivalent to the property that the space spanned by the conditioning process z splits the spaces generated by the conditionally orthogonal processes x, y. The main result is that for jointly wide sense stationary processes x, y, z, conditional orthogonality plus a strong feedback free condition on (z, x) and (z, y), or, equivalently, splitting plus this condition, is equivalent to the existence of a stochastic realization for the joint process (x, y, z) in the special class of so-called conditionally orthogonal stochastic realizations.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 71-84, 2003.
© Springer-Verlag Berlin Heidelberg 2003

6.1 Introduction

Conditional independence for three random variables, x, y, z means that x and y are independent given z and is written $x \perp\!\!\!\perp y \mid z$. This notion is fundamental to the theory of Bayes Nets (see [11] for an excellent summary.) The purpose of this paper is to extend conditional independence to time series models as a prelude to a more detailed development of graphical models for time series. Important recent work is by [7] and [8, 9]. For time series results we refer particularly to [1]. There are also links with early work on causation [10, 6, 3, 4, 2, 14].

We will distinguish different types of independence, particularly conditional independence of the present of the given x and y processes on the past of the z process, which will be referred to as conditional independence at time k, and conditional independence of the past and present of x and y given the present and past of z, which will be referred to as conditional independence up to time k. Since the theory in this paper is solely concerned with second order properties we shall replace conditional independence with conditional orthogonality. Clearly for zero mean Gaussian processes these two conditions are equivalent.

Let H denote an ambient Hilbert space of L^2 random variables with (matricial) inner product $(a, b) \underline{\Delta} Eab^T$ for vector random variables $a, b \in H$. a is orthogonal to b if and only if $(a, b) = 0$ which is written $a \perp b$. For an n component vector a and a closed subspace B of H, $(a \mid B)$ shall denote the vector of orthogonal projections $\{(a_i \mid B); 1 \leq i \leq n\}$ of the components of a on the subspace B. A_p shall denote the linear span of the H-valued components of the vector a_p in the sequence a, and A^p shall denote the linear span of the elements $\{a_k; k \leq p,\ k, p \in N\}$ (also written $\{a_m\}_{m=-\infty}^{k}$) of the sequence a. For closed subspaces A and B of H, $(A \mid B)$ denotes the orthogonal projection of A on B, that is to say, the linear span of $(a|B)$ for all $a \in A$. Henceforth, H shall denote the closure of the space generated by the N component vector process (x, y, z). All subspaces of any Hilbert space under consideration shall be assumed to be closed.

The notation x or $\{x_k\}$ shall be used for a stochastic processes and, according to context, we shall write, for example, $\{x_k\} \perp \{y_k\} \mid \{z_m\}_{m=-\infty}^{m=l}$ if we are referring to processes and (equivalently) $X_k \perp Y_k \mid Z_k$ if referring to subspaces.

The main results of the paper will relate conditional orthogonality directly to two other species of condition: splitting conditions, familiar from systems theory [12, 13] and the existence of stochastic realizations ([12, 13]) with a so-called conditional orthogonal structure (COS) stochastic realizations. We show that conditional orthogonality is equivalent to splitting, and that for jointly wide sense stationary processes either condition taken together with a strongly feedback free condition [1] is equivalent to the existence of a COS stochastic realization. The appropriate definitions will be given in the next section.

The algebra of projections underlies the theory, although we do not exploit it fully in this paper. In another paper [5] this is explained in full detail. However, it is worth giving, without proofs (which are straightforward) a sample of the use of projections.

Consider three zero mean finite dimensional random variables x, y, z of dimension n, m, p, respectively, with a full-rank joint covariance matrix. Then conditional orthogonality, $x \perp y \mid z$ (conditional independence in the Gaussian case), can be expressed in terms of conditions on the appropriate projections.

We shall use the shorthand notation P_{13} for the projection corresponding to the subspace $X+Z$, P_{23} for $Y+Z$ and P_3 for Z, etc. Let $N=n+m+p$, then without loss of generality we can take the $N\times N$ identity, I to be the projector on the N dimensional space $X+Y+Z$. The following conditions are then equivalent to $x \perp y \mid z$:

1. $I = P_{13} + P_{23} - P_3$
2. $(P_{13} - P_3)(P_{23} - P_3) = 0$
3. P_{13} and P_{23} commute
4. $P_{13}P_{23} = P_3$
5. $P_{23}P_1 = P_3P_1$
6. $P_{13}P_2 = P_3P_2$

Condition 2 is the condition for the orthogonality of the innovations $x-(x \mid z)$ and $y-(y \mid z)$. Conditions 5 and 6 will be seen to be precisely the splitting condition for this case, namely $(x \mid Y+Z) = (x \mid Z)$ and its equivalent version $(y \mid X+Z) = (y \mid Z)$ (see Lemma 6.1). The fact that either of these is equivalent to conditional orthogonality, converted to the process case, is the starting point to the theory of the next section.

6.2 Main Results

We relate conditional independence to the concept of splitting subspaces as employed in the theory of stochastic realization as developed by [12, 13] and others. We refer the reader to [1] for an exposition of the theory. The main result below gives a stochastic realization characterization of a strong form of the conditional independence condition.

Unless otherwise stated, all processes appearing in this section are zero mean wide sense stationary processes which possess a rational spectral density matrix which necessarily has no poles on the unit circle; in particular this is the case for the joint process $\{x,y,z\}$. Since, as stated, the theory developed here only deals with the second order properties, we formulate a corresponding second order version of the notion of conditional independence.

DEFINITION 6.1
The processes x and y are *conditionally orthogonal (CO) at $k\in N$ with respect to z up to $l\in N\cup\infty$* if

$$\{x_k\} \perp \{y_k\} \mid \{z_m\}_{m=-\infty}^{m=l} \;\equiv\; \{x_k\} \perp \{y_k\} \mid \{z^l\} \tag{6.1}$$

or, equivalently,

$$x_k - (x_k|Z^l) \perp y_k - (y_k|Z^l) \tag{6.2}$$

These conditions are referred to as *local* when $l<\infty$ and *global* when $l=\infty$.

□

Clearly (1.2) holds if and only if the same property is satisfied by all linear functions of x_k and y_k. Hence x and y are conditionally orthogonal at k with

respect to z up to l if and only if the spaces X_k and Y_k are conditionally orthogonal with respect to z up to l, i.e. with respect to Z^l.

In stochastic realization theory the notion of a *splitting subspace* plays a key role. We extend Definition 3.1, Chapter 4, [1] to the case of an arbitrary ambient space as for so-called *external stochastic realizations*, see e.g. [1, page 238].

DEFINITION 6.2
Let A, B, C be subspaces of a given Hilbert space, then C is said to be a *splitting subspace for* A and B if and only if

$$(A|B+C) \;=\; (A|C). \tag{6.3}$$

□

For brevity, when this condition holds we shall say C *splits* A and B. We then have the following standard lemma (see e.g. [1], page 217) which also shows that the definition above is symmetric in the spaces A and B.

LEMMA 6.1
C is a splitting subspace for A and B if and only if

$$(\alpha,\beta) \;=\; ((\alpha|C),(\beta|C)) \quad \text{for all } \alpha\in A,\ \beta\in B. \tag{6.4}$$

□

Splitting and conditional orthogonality are linked via the next result.

LEMMA 6.2

$$x_k-(x_k|Z^l)\perp y_k-(y_k|Z^l) \tag{6.5}$$

i.e. x and y are *conditionally orthogonal (CO) at k with respect to z up to l*, if and only if

$$(\alpha,\beta) \;=\; ((\alpha|Z^l),(\beta|Z^l)) \quad \text{for all } \alpha\in X_k,\ \beta\in Y_k. \tag{6.6}$$

□

Proof Since $((\alpha|Z^l),\beta) \;=\; ((\alpha,(\beta|Z^l)) \;=\; ((\alpha|Z^l),(\beta|Z^l))$ for all $\alpha,\beta\in H$,

$$(\alpha-(\alpha|Z^l),\beta-(\beta|Z^l)) \;=\; (\alpha,\beta)-((\alpha|Z^l),(\beta|Z^l)) \text{ for all } \alpha\in X_k,\ \beta\in Y_k. \tag{6.7}$$

Equating the right and left hand sides of this equation respectively to 0 yields the necessary and sufficient directions of the lemma. □

Lemmas 6.1 and 6.2 immediately yield

LEMMA 6.3
x and y are conditionally orthogonal (CO) at k with respect to z up to l if and only if Z^l is a splitting subspace for X_k and Y_k.

□

We next introduce

DEFINITION 6.3
The processes x and y are *conditionally orthogonal (CO) up to k with respect to z up to l* if

$$x_p - (x_p|Z^l) \perp y_q - (y_q|Z^l) \quad \text{for all} \quad p,q \leq k, \tag{6.8}$$

or equivalently

$$x_p - (x_p|Z^l) \perp y_q - (y_q|Z^l) \quad \text{for all} \quad p,q \leq k. \tag{6.9}$$

□

In exact analogy with the results above we have

LEMMA 6.4
x and y are conditionally orthogonal (CO) up to k with respect to z up to l if and only if Z^l is a splitting subspace for X^k and Y^k. □

We recall [1] that the jointly second order zero mean stochastic processes u, v are *strongly feedback free (SFF)* if, for all for all $k \in N$ and $s \geq 0$, $(u_k|V^{k+s}) = (u_k|V^{k-1})$.

LEMMA 6.5
Let x, y, z be jointly wide sense stationary processes.

If (i) for some $k \in N$, x and y are conditionally orthogonal up to k with respect to z up to $k+\tau$ for some $\tau \geq 0$, and (ii) (z,x) and (z,y) are strongly feedback free, then x and y are conditionally orthogonal up to k with respect to z up to $k-1$.

□

Proof For all $p,q \leq k-1$, and for the given $\tau \geq 0$

$$x_p - (x_p|Z^{k+\tau}) = x_p - (x_p|Z^{k-1}) \qquad \text{(by } (z,x) \text{ SFF)}, \tag{6.10}$$

and

$$y_q - (y_q|Z^{k+\tau}) = y_q - (y_q|Z^{k-1}) \qquad \text{(by } (z,y) \text{ SFF)}. \tag{6.11}$$

But then $x_p - (x_p|Z^{k-1})$ is orthogonal to $y_q - (y_q|Z^{k-1})$, $p,q \leq k-1$, by the hypothesis that x and y are conditionally orthogonal up to k with respect to z up to $k+\tau$. □

DEFINITION 6.4
The *innovations space I(A;B) of (the space) A relative to (the space) B*, equivalently the *orthogonal complement of B* with respect to A, is the space $(I - P_B)A$, where I denotes the identity projection on H and P_B denotes orthogonal projection into the space B. □

We note immediately that $I(A;B) = I(A+B;B)$, and $(I(A;B),B) = 0$, namely, $I(A;B)$ and B are orthogonal. Furthermore, by (6.3), if C splits A and B, then $I(A;B+C) = I(A;C)$.

Concerning notation, in the following it assumed that the formation of the linear span of a countable set of subspaces of H also involves the formation of the closure of that space; the resulting operation is denoted by $\biguplus$.

LEMMA 6.6
Let a, b, c be zero mean second order processes.

(i) If the processes a, c are such that (c, a) is strongly feedback free, then

$$I(A^k; C^k) = \biguplus_{j \le k} I(a_j; C^j) \equiv \biguplus_{j \le k} I(A_j; C^j) \tag{6.12}$$

(ii) If, further, a, b up to k are conditionally orthogonal with respect to c up to k, for all $k \in N$, then

$$I(A^k; B^k + C^k) = \biguplus_{j \le k} I(a_j; C^j) \tag{6.13}$$

Moreover,

(iii) Under the hypotheses of (ii)

$$I(A^k; B^k + C^{k+1}) = \biguplus_{j \le k} I(a_j; C^{j+1}) = \biguplus_{j \le k} I(a_j; C^j) \tag{6.14}$$

□

Proof For part (i),

$$I(A^k; C^k) = \biguplus_{j \le k} (I - P_{C^k}) a_j \, . \tag{6.15}$$

(since A^k is generated by $\{a_j; j \le k\}$)

$$= \biguplus_{j \le k} (I - P_{C^j}) \, a_j \, , \tag{6.16}$$

as required, where the last equality holds since (c, a) strongly feedback free implies the second equality in

$$P_{C^k} a_j = (a_j | C^k) = (a_j | C^{j-1}) = P_{C^{j-1}} a_j, \tag{6.17}$$

for all $k \ge j - 1$, and hence, by taking first taking the instance $k = j$, we obtain

$$P_{C^k} a_j = (a_j | C^k) = (a_j | C^{j-1}) = P_{C^j} a_j \quad j \le k \, .$$

Part (ii) follows since the CO hypothesis implies that C^k splits A^k, B^k, and so

$$I(A^k; B^k + C^k) = \biguplus_{j \le k} (I - P_{B^k + C^k})\, a_j \tag{6.18}$$

$$= \biguplus_{j \le k} (I - P_{C^k})\, a_j \ , \tag{6.19}$$

and the argument proceeds as before.

(iii) is obtained by first taking the instance $k = j + 1$ in (6.17), which then gives

$$(a_j | C^k) = (a_j | C^{j-1}) = (a_j | C^j) = (a_j | C^{j+1}), \qquad j \le k. \tag{6.20}$$

□

A *conditional stochastic realization* of a (vector processes) a, b with respect to the subprocess b is a stochastic realization where (i) the state space of the realization is the sum $A + B$ of the spaces A and B, (ii) A and B are invariant with respect to the realization system matrix, and (iii) the system input matrix, the state output matrix and the observation noise input matrix have a block triangular structure corresponding to $A+B, B$. The realization is a *conditional orthogonal (CO) stochastic realization* if the subprocesses of the orthogonal input process corresponding to the A, B subspaces are mutually orthogonal and similarly for the orthogonal observation noise process. (In the main theorem below, the space B shall be identified with the span of the z process and A with the span of the innovations of x and y with respect to z, see (6.61), (6.62).)

In order to be self-contained, the statement of the main result below contains a reiteration of our standing assumptions on wide sense stationary processes. The state process construction in the proof of this result follows that of basic stochastic realization results using splitting subspaces (see e.g. [1], Chapter 4, Theorem 4.1). In the analysis here, however, we first construct the innovations processes of x and y relative to their respective pasts plus that of z, then we exploit the conditional orthogonality of x and y with respect to z together with the feedback free property of each pair (z, x), (z, y) in order to obtain the required CO stochastic realization.

THEOREM 6.1
Let (x^T, y^T, z^T) be a $(n + m + p)$ zero mean wide sense stationary processes with a rational spectral density matrix. Then,

(i) x and y are conditionally orthogonal up to k with respect to z up to k,

and

(ii) (z, x) and (z, y) are strongly feedback free,

if and only if

(x^T, y^T, z^T) possesses a stochastic realization of the conditional orthogonal form given in equations (6.61), (6.62), where the system input and observation noise

process is a zero mean wide sense stationary orthogonal process and there exists a stationary covariance for the state process s in (6.61).

□

Proof We first prove the only if part of the theorem. Let the process $\tilde{x}$ of the innovations of x with respect to y, z be defined as

$$\tilde{x} \ \underline{\Delta} \ \{\tilde{x_k} = x_k - (x_k|Y^k + Z^{k+1}); k \in N\}. \tag{6.21}$$

Then the splitting hypothesis that x and y are conditionally orthogonal up to k with respect to z up to k for any $k \in N$ implies that

$$\tilde{x} \ \underline{\Delta} \ \{\tilde{x_k} = x_k - (x_k|Z^{k+1}); \ k \in N\}. \tag{6.22}$$

Using splitting again, the definition on the left below yields the second equality in

$$\tilde{y} \ \underline{\Delta} \ \{\tilde{y_k} \ = \ y_k - (y_k|X^k + Z^{k+1}); \ k \in N\} = \{y_k - (y_k|Z^{k+1}); \ k \in N\}, \tag{6.23}$$

The next step is to construct the following subspaces of H for each $k \in N$:

$$H_k^{-\tilde{x}} \ \underline{\Delta} \ \biguplus \ \{\tilde{x_j}; j \ \leq \ k \ \}, \qquad H_k^{+\tilde{x}} \ \underline{\Delta} \ \biguplus \ \{\tilde{x_j}; j \ \geq \ k \ \}, \tag{6.24}$$

together with the associated sequence of spaces

$$S_k^{\tilde{x}} \ \underline{\Delta} \ (H_k^{+\tilde{x}} \mid H_{k-1}^{-\tilde{x}}), \qquad k \in N. \tag{6.25}$$

It is readily verified that $S_k^{\tilde{x}}$ splits $H_k^{\tilde{x}}$ and $H_{k-1}^{-\tilde{x}}$, for all $k \in N$.

We proceed by making the following orthogonal decomposition of the ambient space H for each $k \in N$, where we note that in the absence of processes other than x, y, z, H may be taken to be $X^\infty + Y^\infty + Z^\infty$. Part (iii) of Lemma 6.6 is invoked to obtain the first term in the third line below and the last line is included for completeness.

$$H = (X^{k-1} + Y^{k-1} + Z^k) \ \oplus \ (X^{k-1} + Y^{k-1} + Z^k)^\perp \tag{6.26}$$

$$= I(X^{k-1}; Y^{k-1} + Z^k) \ \oplus \ (Y^{k-1} + Z^k) \ \oplus \ (X^{k-1} + Y^{k-1} + Z^k)^\perp \ . \tag{6.27}$$

$$= I(X^{k-1}; Z^k) \ \oplus \ I(Y^{k-1}; Z^k) \ \oplus \ Z^k \ \oplus \ (X^{k-1} + Y^{k-1} + Z^k)^\perp \tag{6.28}$$

$$= I(X^{k-1}; Z^k) \ \oplus \ I(Y^{k-1}; Z^k) \ \oplus \ Z^k \ \oplus \ (X^{k-1} + Y^{k-1} + Z^k)^\perp \ . \tag{6.29}$$

Construction of the System Observation Equations

To produce the system state observation equation generating the observed process x we project x_k, for any $k \in N$, as follows:

$$\begin{aligned} x_k &= (x_k|H) \\ &= (x_k|I(X^{k-1}; Y^{k-1}+Z^k)) \oplus (x_k|(Y^{k-1}+Z^k)) \\ &\qquad \oplus (x_k|(X^{k-1}+Y^{k-1}+Z^k)^{\perp}) \qquad (6.30) \\ &= (x_k|I(X^{k-1}; Y^{k-1}+Z^k)) \oplus (x_k|Z^k) \\ &\qquad \oplus (x_k|I(X^k; X^{k-1}+Y^{k-1}+Z^k)), \qquad (6.31) \end{aligned}$$

since, by the splitting hypotheses,

$$(x_k|(Y^{k-1}+Z^k)) = (x_k|Z^k). \qquad (6.32)$$

Next, by an application of part (iii) of Lemma (6.6),

$$I(X^{k-1}; Y^{k-1}+Z^k) = \bigvee_{j\le k-1} I(x_j; Z^{j+1}) = \bigvee_{j\le k-1} I(x_j; Z^j), \qquad (6.33)$$

since, by hypothesis, the processes pair (z, x) is strongly feedback free and x, y up to k are conditionally orthogonal with respect to z up to k.

We now make the crucial observation that since (z, x) is strongly feedback free implies that

$$(x_k \mid Z^{k+1}) = (x_k \mid Z^{k-1}) \in Z^{k-1}, \qquad k \in N. \qquad (6.34)$$

and since $Z^{k-1} \perp I(X^{k-1}; Y^{k-1}+Z^k))$,

$$(x_k|I(X^{k-1}; Y^{k-1}+Z^k)) = (x_k - (x_k|Z^{k+1})|I(X^{k-1}; Y^{k-1}+Z^k)), \quad k \in N. \qquad (6.35)$$

Invoking the representations of $I(X^{k-1}; Y^{k-1}+Z^k$ above we conclude that

$$(x_k \mid I(X^{k-1}; Y^{k-1}+Z^k)) = (\tilde{x_k} \mid \bigvee_{j\le k-1} I(x_j; Z^j)), \qquad k \in N, \qquad (6.36)$$

and observe that by the definition of the $\tilde{x}$ process and of $H_k^{-\tilde{x}}$

$$(\tilde{x_k} \mid \biguplus_{j \le k-1} I(x_j; Z^j)) \;=\; (\tilde{x_k} \mid H_{k-1}^{-\tilde{x}}), \qquad k \in N. \tag{6.37}$$

Now since the process (x^T, y^T, z^T) possesses a rational spectral density matrix the sequence of splitting spaces

$$S_k^{\tilde{x}} \;\underline{\Delta}\; (H_k^{+\tilde{x}} | H_{k-1}^{-\tilde{x}}), \qquad k \in N, \tag{6.38}$$

for $\{(H_k^{+\tilde{x}} | H_{k-1}^{-\tilde{x}}); k \in N\}$, are of constant finite dimension. Choose a basis $s_0^{\tilde{x}}$ for $S_0^{\tilde{x}}$ and hence, by time shift, a basis process $\{s_k^{\tilde{x}}; k \in N\}$ for $\{S_k^{\tilde{x}}; k \in N\}$. Since $\tilde{x_k} \in H_k^{+\tilde{x}}$ we have

$$(\tilde{x_k} \mid H_k^{-\tilde{x}}) \;=\; ((\tilde{x_k} \mid H_k^{+\tilde{x}}) \mid H_k^{-\tilde{x}}) \;=\; (\tilde{x_k} \mid S_k^{\tilde{x}}) \;=\; H_{\tilde{x}}^{x} s_k^{\tilde{x}}, \quad k \in N. \tag{6.39}$$

for some constant matrix $H_{\tilde{x}}^{x}$.

Finally, the observation error process $N_{\tilde{x}}^{x} v^{\tilde{x}}$ shall be defined by

$$\begin{aligned} N_{\tilde{x}}^{x} v_k^{\tilde{x}} \;&=\; (x_k \;\mid\; (X^{k-1} + Y^{k-1} + Z^k)^{\perp}\,) \\ &=\; (x_k \;\mid\; I(X^k; X^{k-1} + Y^{k-1} + Z^k)), \quad k \in N. \end{aligned} \tag{6.40}$$

for some constant matrix $N_{\tilde{x}}^{x}$, where $v_k^{\tilde{x}}$ spans $I(X^k; X^{k-1} + Y^{k-1} + Z^k)$. Taking (6.38), (6.39), and (6.40), we obtain the following state space system output (i.e. observation) equation generating x,

$$x_k \;=\; H_{\tilde{x}}^{x} s_k^{\tilde{x}} \;+\; (x_k | Z^k) \;+\; N_{\tilde{x}}^{x} v_k^{\tilde{x}}, \qquad k \in N. \tag{6.41}$$

Next, a parallel construction projecting y_k on H is performed using the orthogonal decomposition

$$H \;=\; (X^{k-1} + Y^{k-1} + Z^k) \;\oplus\; (X^{k-1} + Y^{k-1} + Z^k)^{\perp} \tag{6.42}$$

$$=\; I(Y^{k-1}; X^{k-1} + Z^k) \;\oplus\; (X^{k-1} + Z^k) \;\oplus\; (X^{k-1} + Y^{k-1} + Z^k)^{\perp} \tag{6.43}$$

Here, in analogy with the previous case, the projection of y_k results in

$$\begin{aligned} y_k \;&=\; (y_k | H) \\ &=\; (y_k | I(Y^{k-1}; X^{k-1} + Z^k)) \\ &\quad\oplus\; (y_k | (X^{k-1} + Z^k)) \\ &\quad\oplus\; (y_k | I(Y^k; X^{k-1} + Y^{k-1} + Z^k)). \end{aligned} \tag{6.44}$$

Again by splitting,

$$(y_k | X^{k-1} + Z^k)) \;=\; (y_k | Z^k), \tag{6.45}$$

and hence

$$y_k = (y_k|I(Y^{k-1}; X^{k-1} + Z^k)) \oplus (y_k|Z^k) \oplus (y_k|I(Y^k; X^{k-1} + Y^{k-1} + Z^k)), \tag{6.46}$$

which, with an analogous definition of the $v^{\tilde{y}}$ process, gives

$$y_k = H_{\tilde{y}}^{y} s_k^{\tilde{y}} + (y_k|Z^k) + N_{\tilde{y}}^{y} v_k^{\tilde{y}}, \qquad k \in N, \tag{6.47}$$

for a constant matrices $H_{\tilde{y}}^{\tilde{y}}, N_{\tilde{y}}^{y}$.

Let $\{S_k^z; k \in N\}$ be the sequence of splitting subspaces given by

$$S_k^z \triangleq (H_k^z|Z^{k-1}), \qquad k \in N, \tag{6.48}$$

where $H_k^{+z} \triangleq \biguplus \{z_j; j \geq k\}$. Choose a basis s_0^z for S_0^z and hence, by time shift, a basis process $\{s_k^z; k \in N\}$ for $\{S_k^z; k \in N\}$. Then the z process evidently satisfies

$$z_k = H_z^z s_k^z + w_k^z, \qquad k \in N, \tag{6.49}$$

for a constant matrix H_z^z and a process w^z, where $w_k^z \in I(Z^k; S_k^z) = I(Z^k; Z^{k-1})$ is orthogonal to Z^{k-1} and hence w^z itself is an orthogonal process.

Construction of the State Space Equations

Employing the orthogonal decomposition of H used for the representation of the x process (6.30) for the decomposition of the $s^{\tilde{x}}$ process yields

$$\begin{aligned} s_{k+1}^{\tilde{x}} &= (s_{k+1}^{\tilde{x}}|H) \\ &= (s_{k+1}^{\tilde{x}}|I(X^{k-1}; Y^{k-1} + Z^k)) \\ &\quad \oplus (s_{k+1}^{\tilde{x}} \mid (Y^{k-1} + Z^k)) \\ &\quad \oplus (s_{k+1}^{\tilde{x}} \mid (X^{k-1} + Y^{k-1} + Z^k)^{\perp}) \qquad (6.50) \\ &= (s_{k+1}^{\tilde{x}}|H_{k-1}^{-\tilde{x}}) \oplus (s_{k+1}^{\tilde{x}} \mid Z^k) \\ &\quad \oplus (s_{k+1}^{\tilde{x}} \mid I(X^k; X^{k-1} + Y^{k-1} + Z^k)). \qquad (6.51) \end{aligned}$$

where the first expression on the right hand side above holds by virtue of

$$I(X^{k-1}; Y^{k-1} + Z^k) = \biguplus_{j \leq k-1} I(x_j; Z^j) = H_{k-1}^{-\tilde{x}},$$

and the second is given by the fact that

$$s_{k+1}^{\tilde{x}} \in S_{k+1}^{\tilde{x}} \subset H_k^{-\tilde{x}} \subset H_k^{-x} + Z^{k+1},$$

where splitting gives $(H_k^{-x}|(Y^{k-1} + Z^k)) = (H_k^{-x}|Z^k)$, while (z, y) SFF implies $(Z^{k+1}|Y^{k-1} + Z^k) = (Z^{k+1}|Z^k)$.

The key step in the use of splitting in the construction of the state space equations is the following set of equations, which uses $H_{k+1}^{+\tilde{x}} \subset H_k^{+\tilde{x}}$ to obtain the second equality.

$$\begin{aligned}(S_{k+1}^{\tilde{x}} \mid H_{-1}^{-\tilde{x}}) &= ((H_{k+1}^{+\tilde{x}} \mid H_k^{-\tilde{x}}) \mid H_{-1}^{-\tilde{x}}) \\ &= ((H_{k+1}^{+\tilde{x}} \mid H_k^{-\tilde{x}}) \mid (H_k^{+\tilde{x}} \mid H_{-1}^{-\tilde{x}})) \\ &= (S_{k+1}^{\tilde{x}} \mid S_k^{\tilde{x}}). \end{aligned} \tag{6.52}$$

This implies that

$$(s_{k+1}^{\tilde{x}} | H_{-1}^{-\tilde{x}}) = (s_{k+1}^{\tilde{x}} \mid S_k^{\tilde{x}}) = F_{\tilde{x}}^{\tilde{x}} s_k^{\tilde{x}}, \tag{6.53}$$

for some constant matrix $F_{\tilde{x}}^{\tilde{x}}$. Since it is readily verified using the strong feedback free property for (z, x) that $H_k^{-\tilde{x}} \perp Z^k$ for all $k \in N$, we obtain

$$(s_{k+1}^{\tilde{x}} \mid Z^k) = 0, \qquad k \in N. \tag{6.54}$$

So finally, the state space recursion for the $\{s^{\tilde{x}}\}$ process is obtained by defining the orthogonal process

$$M_{\tilde{x}}^{x} v_k^{\tilde{x}} = (s_{k+1}^{\tilde{x}} \mid I(X^k; X^{k-1} + Y^{k-1} + Z^k)), \tag{6.55}$$

for some constant matrix $M_{\tilde{x}}^{x}$, to yield

$$s_{k+1}^{\tilde{x}} = F_{\tilde{x}}^{\tilde{x}} s_k^{\tilde{x}} + M_{\tilde{x}}^{x} v_k^{\tilde{x}}, \qquad k \in N, \tag{6.56}$$

An exactly analogous line of argument leads to

$$s_{k+1}^{\tilde{y}} = F_{\tilde{y}}^{\tilde{y}} s_k^{\tilde{y}} + M_{\tilde{y}}^{y} v_k^{\tilde{y}}, \tag{6.57}$$

with the corresponding orthogonal process defined by

$$M_{\tilde{y}}^{y} v_k^{\tilde{y}} = (s_{k+1}^{\tilde{y}} \mid I(Y^k; X^{k-1} + Y^{k-1} + Z^k)), \qquad k \in N. \tag{6.58}$$

Decomposing H as $H = Z^k + (Z^k)^{\perp}$, and using the splitting subspaces $\{S_k^z; k \in N\}$, gives the final state space recursion

$$s_{k+1}^{z} = F_z^z s_k^z + M_z^z w_k^z, \qquad k \in N, \tag{6.59}$$

for some constant matrices F_z^z, M_z^z, where we recall the process w is orthogonal since $w_k \in I(Z^k; Z^{k-1}); k \in N$.

It remains to account for the processes $\{x_k \mid Z^k; k \in N\}$ and $\{y_k \mid Z^k; k \in N\}$. Since $H_k^{-\tilde{x}} \perp Z^k$ for all $k \in N$, the process x has a representation of the form

$$x = W_{\tilde{x}}(\delta) v^{\tilde{x}} \oplus W_z(\delta) w^z, \tag{6.60}$$

where $W_{\tilde{x}}(\delta), W_z(\delta)$ are non-anticipative functions in the shift operator δ.

By hypothesis, the process x^T, z^T has a rational spectral density and hence has a finite dimensional minimal stochastic realization (see e.g. [1]); hence the process $W_z(\delta)w^z$ has a finite dimensional stochastic realization. We shall denote a chosen finite dimensional state process for $W_z(\delta)w^z$ by $s^{x|z}$; evidently the input process and the observation innovations process for the realization may be taken to be the orthogonal process w^z. The rational spectrum hypothesis also implies that the dimension of the isometrically isomorphic state spaces $\{S^{\tilde{x}}_k; k \in N\}$ is finite for each k; the same holds for the spaces $\{S^{\tilde{y}}_k; k \in N\}$ and $\{S^z_k; k \in N\}$ for the the $\tilde{y}$ and z processes respectively.

Let us now combine the processes $s^{\tilde{x}}$ and $s^{x|z}$ into the joint state process s^x and combine the corresponding system matrices into the block diagonal system matrix F^x_x; the system matrix F^z_z is necessarily asymptotically stable and the system matrix F^x_x may be chosen to be asymptotically stable. An analogous procedure is also carried out for the y process. Taking the state recursion (6.56) and the output equation (6.41) together with the stochastic realization of $s^{x|z}$ gives (i) the first row of the state recursion array (6.61) below for a constant matrix M^x_z and (ii) the first row of the output equation array (6.62) for corresponding constant matrices H^x_x, N^x_z.

A parallel sequence of constructions for the y process then yields the second row in (6.61) and (6.62). Finally the process z has a stochastic realization purely in terms of the w^z process as shown in the last row of the two arrays.

$$\begin{bmatrix} s^x_{k+1} \\ s^y_{k+1} \\ s^z_{k+1} \end{bmatrix} = \begin{bmatrix} F^x_x & 0 & 0 \\ 0 & F^y_y & 0 \\ 0 & 0 & F^z_z \end{bmatrix} \begin{bmatrix} s^x_k \\ s^y_k \\ s^z_k \end{bmatrix} + \begin{bmatrix} M^x_{\tilde{x}} & 0 & M^x_z \\ 0 & M^y_{\tilde{y}} & M^y_z \\ 0 & 0 & M^z_z \end{bmatrix} \begin{bmatrix} v^{\tilde{x}}_k \\ v^{\tilde{y}}_k \\ w^z_k \end{bmatrix}, \tag{6.61}$$

$$\begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} = \begin{bmatrix} H^x_x & 0 & 0 \\ 0 & H^y_y & 0 \\ 0 & 0 & H^z_z \end{bmatrix} \begin{bmatrix} s^x_k \\ s^y_k \\ s^z_k \end{bmatrix} + \begin{bmatrix} N^x_{\tilde{x}} & 0 & N^x_z \\ 0 & N^y_{\tilde{y}} & N^y_z \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} v^{\tilde{x}}_k \\ v^{\tilde{y}}_k \\ w^z_k \end{bmatrix}. \tag{6.62}$$

It is evident from the construction of the stochastic realization above that the process $((v^{\tilde{x}})^T, (v^{\tilde{y}})^T, (w^z)^T)$ is an orthogonal process and that further the process $((v^{\tilde{x}})^T, (v^{\tilde{y}})^T)$ is orthogonal to w^z. Hence (6.61), (6.62) constitutes a CO stochastic realization which by the asymptotic stability of the system matrix possesses an invariant state covariance. Clearly also, $v^{\tilde{x}}$ is orthogonal to $v^{\tilde{y}}$.

For the if part of the theorem, we see by inspection that a CO stochastic realization of the form above, with asymptotically stable state matrix and innovations process satisfying the given orthogonality conditions, gives rise to a wide sense stationary process (x^T, y^T, z^T) with a rational spectral density. It may be verified

that the resulting process (x^T, y^T, z^T) satisfies the conditional orthogonality and strong feedback free conditions of the theorem statement and so the converse part of the theorem holds. □

Acknowledgments

The work contained within this paper has been partially funded by the EU as part of the Fifth Framework MAPP project (Multivariate Approach for Statistical Process Control and Cleaner Production. Code: IST-1999-11990.)

6.3 References

[1] P. E. Caines. *Linear Stochastic Systems*. New York; Chichester: Wiley, 1988.

[2] P. E. Caines and C. W. Chan. Estimation, identification and feedback. *Transactions on Automatic Control*, AC-20(4):498–508, 1975.

[3] P. E. Caines and C. W. Chan. Feedback between stationary stochastic processes. *Transactions on Automatic Control*, AC-20(4):498–508, 1975.

[4] P. E. Caines and C. W. Chan. *System Identification: Advances and Cases Studies*, pages 349–405. Academic Press, New York, 1976. Chapter: Estimation, identification and feedback.

[5] P. E. Caines, R. Deardon, and H. P. Wynn. Conditional independence for time series graphical models: algebraic methods. In manuscript, 2002.

[6] C. W. Chan. The identification of closed loop systems with application to econometric problems. Master's thesis, University of Manchester Institute of Science and Technology, Manchester, UK, 1972.

[7] R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51:157–172, 2000.

[8] M. Eichler. *Graphical Models in Time Series Analysis*. PhD thesis, University of Heidelberg, 1999.

[9] M. Eichler. Granger-causality graphs for multivariate time series. Preprint, University of Heidelberg, 2001.

[10] C. W. J. Granger. Investigating causal relations by econometric models and cross spectral methods. *Econometrica*, 37:424–438, 1969.

[11] S. L. Lauritzen. *Graphical Models*. Oxford Univerosty Press, 1996.

[12] A. Lindquist and G. Picci. On the stochastic realization problem. *SIAM J. Control Optim. Theory*, 17(3):365–389, 1979.

[13] A. Lindquist and G. Picci. State space models for Gaussian stochastic processes in *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, pages 169–204. Pub: Reidel, Dordrecht, 1981. Ed: M. Hazewinkel and J. C. Willems.

[14] C. A. Sims. Money, income and causality. *American Economic Review*, 62:540–552, 1972.

7

Geometry of Oblique Splitting Subspaces, Minimality and Hankel Operators

Alessandro Chiuso *Giorgio Picci*

Abstract

Stochastic realization theory provides a natural theoretical background for recent identification methods, called *subspace methods*, which have shown superior performance for multivariable state-space model-building. The basic steps of subspace algorithms are geometric operations on certain vector spaces generated by observed input-output time series which can be interpreted as "sample versions" of the abstract geometric operations of stochastic realization theory. The construction of the state space of a stochastic process is one such basic operation.

In the presence of exogenous inputs the state should be constructed starting from input-output data observed on a finite interval. This and other related questions still seems to be not completely understood, especially in presence of *feedback* from the output process to the input, a situation frequently encountered in applications. This is the basic motivation for undertaking a first-principle analysis of the stochastic realization problem with inputs, as presented in this paper. It turns out that stochastic realization with inputs is by no means a trivial extension of the well-established theory for stationary processes (time-series) and there are fundamentally new concepts involved, e.g. in the construction of the state space under possible presence of feedback from the output process to the input. All these new concepts lead to a richer theory which (although far from being complete) substantially generalizes and puts what was known for the time series setting in a more general perspective.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 85-126, 2003.
© Springer-Verlag Berlin Heidelberg 2003

7.1 Introduction

In this paper we shall study the stochastic realization problem with inputs. Our aim will be to discuss procedures for constructing state space models for a stationary process $\boldsymbol{y}$ "driven" by an exogenous observable input signal $\boldsymbol{u}$, also modelled as a stationary process, of the form

$$\left\{ \begin{array}{rcl} \boldsymbol{x}(t+1) & = & A\boldsymbol{x}(t) + B\boldsymbol{u}(t) + G\boldsymbol{w}(t) \\ \mathbf{y}(t) & = & C\boldsymbol{x}(t) + D\boldsymbol{u}(t) + J\boldsymbol{w}(t) \end{array} \right. \tag{7.1}$$

where $\boldsymbol{w}$ is a normalized white noise process. We will especially be interested in coordinate-free (i.e. "geometric") procedures by which one could abstractly *construct* state space models of the form (7.1), starting from certain Hilbert spaces of random variables generated by the "data" of the problem, namely the processes $\boldsymbol{y}$ and $\boldsymbol{u}$. We shall also characterize some structural properties of state space models of this kind, like minimality absence of feedback etc.

Stochastic realization theory lies at the grounds of a recent identification methodology, called *subspace identification*, which has shown superior performance especially for multivariable state-space model-building, and has been intensively investigated in the past ten years [13, 21, 22, 20, 31]. The basic steps of subspace algorithms are geometric operations on certain vector spaces generated by observed input-output time series. These operations can be interpreted as "sample versions" of certain abstract geometric operations of stochastic realization theory [20, 26, 28]. In fact, it is by now well understood that stochastic realization theory provides a natural theoretical background for subspace identification of time-series (no inputs). The celebrated subspace algorithm of [21] uses the sample-version of a standard geometric construction of the state space (projection of the future onto the past) and computes the G, J parameters of the model by solving the Riccati equation of stochastic realization.

The situation is unfortunately not as clear for identification in the presence of exogenous inputs. Some basic conceptual issues underlying the algorithms remain unclear (see [4]). One such issue is how the state space of a stochastic process in the presence of exogenous inputs should be constructed starting from input-output data observed on a finite interval. This and other related questions are examined in the recent paper [4]. On the basis of the analysis of this paper, one may be led to conclude that all identification procedures with inputs appeared so far in the literature use *ad hoc* approximations of the basic step of state-space construction of the output process $\boldsymbol{y}$, and can only lead to suboptimal performance.

This state of affairs is the basic motivation for undertaking a first-principle analysis of the stochastic realization problem with inputs, as presented in this paper. We warn the reader that stochastic realization with inputs is *not* a trivial extension of the well-established theory for stationary processes (time-series) as there are fundamentally new concepts involved, relating to the construction of the state space, like the possible presence of *feedback* [7, 6] from the output process to the input, the diverse notion of minimality etc.. All these new concepts lead to a richer theory which substantially generalizes and puts what was known for the time series setting in a more general perspective.

In order to construct state-space descriptions of $\boldsymbol{y}$ driven by a non-white process $\boldsymbol{u}$ of the above form it is necessary to generalize the theory of stochastic realization of [18, 19]. The construction presented here is based on an extension of the idea of Markovian splitting subspace which will be called *oblique Markovian splitting subspace*, a concept introduced in [28, 12]. For this reason we shall start the paper by studying oblique projections in a Hilbert space context.

7.2 Oblique Projections

Let $\mathcal{H}$ be a Hilbert space of real zero-mean random variables with inner product

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle := \mathrm{E}\,\{\boldsymbol{x}\boldsymbol{z}\} \tag{7.2}$$

the operator E denoting mathematical expectation. All through this paper we shall denote direct sum of subspaces by the symbol $+$. The symbol $\oplus$ will be reserved for *orthogonal* direct sum. Consider a pair of closed subspaces $\mathcal{A}, \mathcal{B}$ of $\mathcal{H}$ which are in direct sum, i.e. $\mathcal{A} \cap \mathcal{B} = \{0\}$ so that every element $\boldsymbol{v} \in \mathcal{A} + \mathcal{B}$ can be uniquely decomposed in the sum

$$\boldsymbol{v} = \boldsymbol{v}_{\mathcal{A}} + \boldsymbol{v}_{\mathcal{B}}, \quad \boldsymbol{v}_{\mathcal{A}} \in \mathcal{A} \;\; \boldsymbol{v}_{\mathcal{B}} \in \mathcal{B}$$

It follows that the orthogonal projection of a random variable $\boldsymbol{z} \in \mathcal{H}$, on $\mathcal{A} + \mathcal{B}$ admits the unique decomposition

$$\mathrm{E}\,[\,\boldsymbol{z} \mid \mathcal{A} + \mathcal{B}\,] = \boldsymbol{z}_{\mathcal{A}} + \boldsymbol{z}_{\mathcal{B}}$$

the two components $\boldsymbol{z}_{\mathcal{A}}$ and $\boldsymbol{z}_{\mathcal{B}}$ being, by definition, the oblique projection of $\boldsymbol{z}$ onto $\mathcal{A}$ along $\mathcal{B}$ and the oblique projection of $\boldsymbol{z}$ onto $\mathcal{B}$ along $\mathcal{A}$, denoted by the symbols

$$\boldsymbol{z}_{\mathcal{A}} = \mathrm{E}_{\parallel \mathcal{B}}\,[\,\boldsymbol{z} \mid \mathcal{A}\,] \quad , \quad \boldsymbol{z}_{\mathcal{B}} = \mathrm{E}_{\parallel \mathcal{A}}\,[\,\boldsymbol{z} \mid \mathcal{B}\,]$$

If $\mathcal{A}$ and $\mathcal{B}$ are orthogonal, then the oblique projection becomes orthogonal, i.e.

$$\boldsymbol{z}_{\mathcal{A}} = \mathrm{E}_{\parallel \mathcal{B}}\,[\,\boldsymbol{z} \mid \mathcal{A}\,] = \mathrm{E}\,[\,\boldsymbol{z} \mid \mathcal{A}]$$

which, trivially, does not depend on $\mathcal{B}$.

Projections of one subspace onto another subspace will be encountered frequently. We shall denote these objects by

$$\mathrm{E}\,[\,\mathcal{B} \mid \mathcal{A}\,] := \overline{\mathrm{span}}\,\{\,\mathrm{E}\,[\boldsymbol{z} \mid \mathcal{A}] \mid \boldsymbol{z} \in \mathcal{B}\,\}$$

and

$$\mathrm{E}_{\parallel \mathcal{B}}\,[\,C \mid \mathcal{A}\,] := \overline{\mathrm{span}}\,\{\,\mathrm{E}_{\parallel \mathcal{B}}\,[\,\boldsymbol{z} \mid \mathcal{A}\,] \mid \boldsymbol{z} \in C\,\}$$

The following lemma will be extensively used in the following.

LEMMA 7.1
Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ be closed subspaces of $\mathcal{H}$, where $\mathcal{B} \subset \mathcal{C}$. Assume that

$$\mathcal{D} \cap \mathcal{C} = \{0\} \tag{7.3}$$

and

$$\mathrm{E}\,[\mathcal{A} \mid \mathcal{D} + \mathcal{C}] = \mathrm{E}\,[\mathcal{A} \mid \mathcal{D} + \mathcal{B}] \tag{7.4}$$

then

$$\mathrm{E}_{\parallel \mathcal{C}}\,[\mathcal{A} \mid \mathcal{D}] = \mathrm{E}_{\parallel \mathcal{B}}\,[\mathcal{A} \mid \mathcal{D}] \tag{7.5}$$

□

Proof From (7.3) every $a \in \mathcal{A}$ can be uniquely decomposed as $a = (a_{\mathcal{D}} + a_{\mathcal{C}}) \oplus \tilde{a}$ where $a_{\mathcal{D}} \in \mathcal{D}$, $a_{\mathcal{C}} \in \mathcal{C}$, and $\tilde{a} \perp (\mathcal{C} + \mathcal{D})$. It follows from (7.4) that $a_{\mathcal{C}} \in \mathcal{B}$ and therefore $a_{\mathcal{B}} = a_{\mathcal{C}}$, or, more precisely,

$$\mathrm{E}_{\parallel \mathcal{D}}\,[\,a \mid \mathcal{C}\,] \;=\; \mathrm{E}_{\parallel \mathcal{D}}\,[\,a \mid \mathcal{B}\,] \qquad a \in \mathcal{A}$$

By uniqueness of the orthogonal projection,

$$\mathrm{E}\,[a \mid \mathcal{D} + \mathcal{C}] = a_{\mathcal{D}} + a_{\mathcal{B}} = \mathrm{E}\,[a \mid \mathcal{D} + \mathcal{B}]$$

which implies $\mathrm{E}_{\parallel \mathcal{C}}\,[a \mid \mathcal{D}] = \mathrm{E}_{\parallel \mathcal{B}}\,[a \mid \mathcal{D}]$. This equality obviously holds for any finite linear combinations of elements of $\mathcal{A}$. To complete the proof just take closure with respect to the inner product (7.2). □

Note that in general the converse implication (7.5) $\Rightarrow$ (7.4) is not true since $\mathrm{E}_{\parallel \mathcal{B}}\,[a_{\mathcal{C}} \mid \mathcal{D}] = 0$, does not imply that $a_{\mathcal{C}} \in \mathcal{B}$ but only that $\mathrm{E}\,[a_{\mathcal{C}} \mid \mathcal{D} + \mathcal{B}] = \mathrm{E}\,[a_{\mathcal{C}} \mid \mathcal{B}] = a_{\mathcal{B}}$, i.e. $a_{\mathcal{C}} = \mathrm{E}\,[a_{\mathcal{C}} \mid \mathcal{B}] \oplus \tilde{a}_{\mathcal{B}}$ where $\tilde{a}_{\mathcal{B}} \perp \mathcal{D} + \mathcal{B}$ which is for instance always the case if $\mathcal{C} \ominus \mathcal{B} \perp \mathcal{D}$.

Also the following lemma will be of primary importance.

LEMMA 7.2
Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ be closed subspaces of $\mathcal{H}$ where

$$\mathcal{C} \cap \mathcal{D} = \{0\} \tag{7.6}$$

If $\mathcal{B} \subset \mathcal{C}$ then the following conditions are equivalent:

1. $\mathrm{E}_{\parallel \mathcal{D}}\,[\mathcal{A} \mid \mathcal{C}] = \mathrm{E}_{\parallel \mathcal{D}}\,[\mathcal{A} \mid \mathcal{B}]$;
2. $\mathrm{E}\,[\mathcal{A} \mid \mathcal{C} + \mathcal{D}] = \mathrm{E}\,[\mathcal{A} \mid \mathcal{B} + \mathcal{D}]$

□

Proof $(1 \Rightarrow 2)$ By assumption (7.6) every $a \in \mathcal{A}$ can be uniquely decomposed as $a = (a_{\mathcal{C}} + a_{\mathcal{D}}) \oplus \tilde{a}$ where $a_{\mathcal{C}} \in \mathcal{C}$, $a_{\mathcal{D}} \in \mathcal{D}$, and $\tilde{a} \perp (\mathcal{C} + \mathcal{D})$. It follows from (1.) that $a_{\mathcal{C}} \in \mathcal{B}$; in fact (1.) implies that $a_{\mathcal{C}} = \mathrm{E}_{\parallel \mathcal{D}}\,[a_{\mathcal{C}} \mid \mathcal{C}] = \mathrm{E}_{\parallel \mathcal{D}}\,[a_{\mathcal{C}} \mid \mathcal{B}]$. Therefore $a_{\mathcal{B}} = a_{\mathcal{C}}$. This condition insures that $\mathrm{E}\,[a \mid \mathcal{C} + \mathcal{D}] = a_{\mathcal{B}} + a_{\mathcal{D}} = \mathrm{E}\,[a \mid \mathcal{B} + \mathcal{D}]$ by uniqueness of the orthogonal projection, which, taking closure, implies (2.).
$(2 \Rightarrow 1)$ By the same decomposition of a, (2.) implies $\mathrm{E}\,[a \mid \mathcal{C} + \mathcal{D}] = a_{\mathcal{C}} + a_{\mathcal{D}} = \mathrm{E}\,[a \mid \mathcal{B} + \mathcal{D}]$ and hence by uniqueness $a_{\mathcal{C}} \in \mathcal{B}$. Therefore $a_{\mathcal{B}} = a_{\mathcal{C}}$. The above condition implies that $\mathrm{E}_{\parallel \mathcal{D}}\,[a \mid \mathcal{C}] = a_{\mathcal{B}} = \mathrm{E}_{\parallel \mathcal{D}}\,[a \mid \mathcal{B}]$. To complete the proof just take closure with respect to the inner product, and everything goes trough since all subspaces are closed. □

REMARK 7.1
While Lemma 7.1 gives conditions for reducing the subspace *along which* we project, Lemma 7.2 gives conditions for reducing the subspace *onto which* we project. In Lemma 7.2 the conditions are equivalent, while in Lemma 7.1 only one implication holds. The reason for this is explained in the proof of lemma 7.1 and, roughly speaking, it amounts to the fact that condition (7.5) only guarantees that the component a_C lying on $\mathcal{C}$ of any element in $\mathcal{A}$ (which is uniquely defined), splits uniquely in the orthogonal sum $a_C = \mathrm{E}\,[a_C \mid \mathcal{B}] \oplus \tilde{a}_{\mathcal{B}}$ where $\tilde{a}_{\mathcal{B}} \perp \mathcal{D}$ and not $\tilde{a}_{\mathcal{B}} = 0$, which would be necessary to prove the opposite implication. In lemma 7.2 instead the condition on the oblique projection actually guarantees that the component a_C lying on $\mathcal{C}$ of any element in $\mathcal{A}$ is indeed in $\mathcal{B}$. □

How oblique projections can be computed in a finite dimensional setting is addressed in Lemma 1 of [12].

7.3 Notations and Basic Assumptions

The Hilbert space setting for the study of second-order stationary processes is standard. Here we shall work in discrete time $t = \ldots, -1, 0, 1, \ldots$, and make the assumption that all processes involved are jointly (second-order) stationary and with zero mean. The $m+p$-dimensional joint process $[\boldsymbol{y}\,\boldsymbol{u}]'$ will be assumed purely non deterministic and of full rank [30]. Sometimes we shall make the assumption of rational spectral densities in order to work with finite-dimensional realizations, however the geometric theory described in this paper is completely general and works also in the infinite dimensional case.

For $-\infty < t < +\infty$ introduce the linear subspaces of second order random variables

$$\begin{aligned} \mathcal{U}_t^- &:= \overline{\mathrm{span}}\,\{\boldsymbol{u}_k(s);\ k = 1, \ldots, p,\ s < t\,\} \\ \mathcal{Y}_t^- &:= \overline{\mathrm{span}}\,\{\boldsymbol{y}_k(s);\ k = 1, \ldots, m,\ s < t\,\} \end{aligned}$$

where the bar denotes closure with respect to the metric induced by the inner product (7.2). These are the Hilbert spaces of random variables spanned by the infinite past of $\boldsymbol{u}$ and $\boldsymbol{y}$ up to time t. By convention the past spaces do not include the present. We shall call

$$\mathcal{P}_t := \mathcal{U}_t^- \vee \mathcal{Y}_t^-$$

(the $\vee$ denotes closed vector sum) the *joint past space* of the input and output processes at time t.

Subspaces spanned by random variables at just one time instant are simply denoted $\mathcal{U}_t$, $\mathcal{Y}_t$, etc. while the spaces generated by the whole time history of $\boldsymbol{u}$ and $\boldsymbol{y}$ we shall use the symbols $\mathcal{U}, \mathcal{Y}$, respectively.

The *shift operator* σ is a unitary map defined on a dense subset of $\mathcal{U} \vee \mathcal{Y}$ by the assignment

$$\sigma\left(\textstyle\sum_k a_k' \boldsymbol{y}(t_k) + \sum b_j' \boldsymbol{u}(t_j)\right) := \left(\textstyle\sum_k a_k' \boldsymbol{y}(t_k + 1) + \sum b_j' \boldsymbol{u}(t_j + 1)\right)$$

$$a_k \in \mathbb{R}^m,\ b_j \in \mathbb{R}^p,\ t_k,\ t_j \in \mathbb{Z}$$

Because of stationarity σ can be extended to the whole space as a unitary operator, see e.g. [30].

The processes $\boldsymbol{y}$ and $\boldsymbol{u}$ propagate in time by the shift operator (e.g. $\boldsymbol{y}(t) = \sigma^t \boldsymbol{y}(0)$); this in particular implies that all relations involving random variables of $\mathcal{U} \vee \mathcal{Y}$ which are valid at a certain instant of time t, by applying the shift operator on both sides of the relation, are seen to be also automatically valid at any other time. For this reason all definitions and statements this paper involving subspaces or random variables defined at a certain time instant t are to be understood as holding also for arbitrary $t \in \mathbb{Z}$.

Conditional Orthogonality

We say that two subspaces $\mathcal{A}$ and $\mathcal{B}$ of a Hilbert space $\mathcal{H}$ are *conditionally orthogonal* given a third subspace $\mathcal{X}$ if

$$\langle \alpha - \mathrm{E}^{\mathcal{X}} \alpha, \beta - \mathrm{E}^{\mathcal{X}} \beta \rangle = 0 \quad \text{for} \quad \alpha \in \mathcal{A},\, \beta \in \mathcal{B} \tag{7.7}$$

and we shall denote this $\mathcal{A} \perp \mathcal{B} \mid \mathcal{X}$. When $\mathcal{X} = 0$, this reduces to the usual orthogonality $\mathcal{A} \perp \mathcal{B}$. Conditional orthogonality is orthogonality after subtracting the projections on $\mathcal{X}$. Using the definition of the projection operator $\mathrm{E}^{\mathcal{X}}$, it is straightforward to see that (7.7) may also be written

$$\langle \mathrm{E}^{\mathcal{X}} \alpha, \mathrm{E}^{\mathcal{X}} \beta \rangle = \langle \alpha, \beta \rangle \quad \text{for} \quad \alpha \in \mathcal{A},\, \beta \in \mathcal{B}. \tag{7.8}$$

The following lemma is a trivial consequence of the definition.

LEMMA 7.3
If $\mathcal{A} \perp \mathcal{B} \mid \mathcal{X}$, then $\mathcal{A}_0 \perp \mathcal{B}_0 \mid \mathcal{X}$ for all $\mathcal{A}_0 \subset \mathcal{A}$ and $\mathcal{B}_0 \subset \mathcal{B}$. □

Let $\mathcal{A} \oplus \mathcal{B}$ denote the *orthogonal* direct sum of $\mathcal{A}$ and $\mathcal{B}$. If $\mathcal{C} = \mathcal{A} \oplus \mathcal{B}$, then $\mathcal{B} = \mathcal{C} \ominus \mathcal{A}$ is the orthogonal complement of $\mathcal{A}$ in $\mathcal{C}$. The following Proposition from [18, 19] describes some useful alternative characterizations of conditional orthogonality.

LEMMA 7.4
The following statements are equivalent.

(i) $\mathcal{A} \perp \mathcal{B} \mid \mathcal{X}$

(ii) $\mathcal{B} \perp \mathcal{A} \mid \mathcal{X}$

(iii) $(\mathcal{A} \vee \mathcal{X}) \perp \mathcal{B} \mid \mathcal{X}$

(iv) $\mathrm{E}^{\mathcal{A} \vee \mathcal{X}} \beta = \mathrm{E}^{\mathcal{X}} \beta \quad \text{for } \beta \in \mathcal{B}$

(v) $(\mathcal{A} \vee \mathcal{X}) \ominus \mathcal{X} \perp \mathcal{B}$

(vi) $\mathrm{E}^{\mathcal{A}} \beta = \mathrm{E}^{\mathcal{A}} \mathrm{E}^{\mathcal{X}} \beta \quad \text{for } \beta \in \mathcal{B}$

□

Feedback

Let $\boldsymbol{y}$ and $\boldsymbol{u}$ be two jointly stationary vector stochastic processes. In general one may express both $\boldsymbol{y}(t)$ and $\boldsymbol{u}(t)$ as a sum of the best linear estimate based on the past of the other variable, plus "noise"

$$\boldsymbol{y}(t) = \mathrm{E}\left[\boldsymbol{y}(t) \mid \mathcal{U}_{t+1}^{-}\right] + \boldsymbol{d}(t) \tag{7.9a}$$
$$\boldsymbol{u}(t) = \mathrm{E}\left[\boldsymbol{u}(t) \mid \mathcal{Y}_{t+1}^{-}\right] + \boldsymbol{r}(t) \tag{7.9b}$$

so that each variable $\boldsymbol{y}(t)$ and $\boldsymbol{u}(t)$ can be expressed as a sum of a causal linear transformation of the past of the other, plus "noise". Here the noise terms are uncorrelated with the past of $\boldsymbol{u}$ and $\boldsymbol{y}$ respectively, but may in general be mutually correlated.

Since both linear estimators above can be expressed as the output of linear filters, represented by causal transfer functions $F(z)$ and $H(z)$, the joint model (7.9) corresponds to a *feedback interconnection* of the type

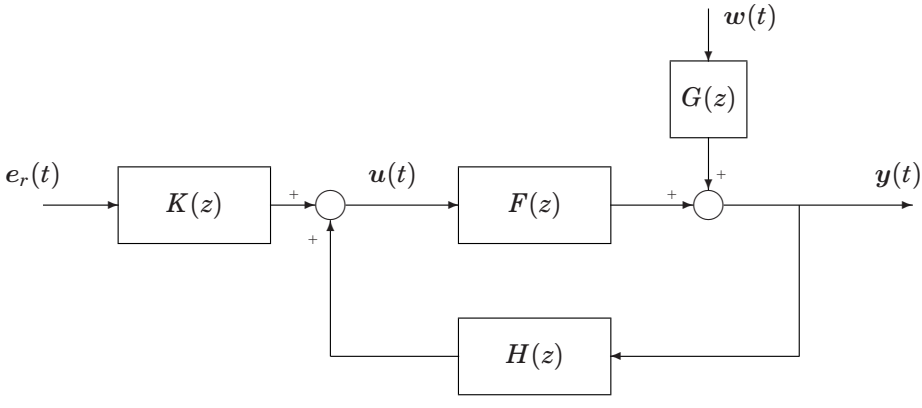


Figure 7.1 Feedback model of the processes $\boldsymbol{y}$ and $\boldsymbol{u}$.

where (in symbolic "Z-transform" notation)

$$\boldsymbol{d}(t) = G(z)\boldsymbol{w}(t) \qquad \boldsymbol{r}(t) = K(z)\boldsymbol{e}_r(t)$$

in which $G(z)$, $K(z)$ may be assumed, without loss of generality, minimum phase spectral factors of the spectra $\Phi_d(z)$ and $\Phi_r(z)$ of the two stationary "error" (or "disturbance") signals d and r.

In this scheme the "errors" $\boldsymbol{r}$ and $\boldsymbol{d}$ are in general correlated. More useful are feedback schemes which involve uncorrelated error processes, since in physical models of feedback systems this will more usually be the situation. It can be shown that any pair of jointly stationary processes can also be represented by schemes of this last type. In this case however, although the overall interconnection must be internally stable, the individual transfer functions $F(z)$ and $H(z)$ may well be unstable; see [6] for a detailed discussion.

Following Granger [8], and subsequent work by Caines, Chan, Anderson, Gevers etc. [3, 7, 6] we say that *there is no feedback from* $\boldsymbol{y}$ *to* $\boldsymbol{u}$ if the future of $\boldsymbol{u}$

is conditionally uncorrelated with the past of $\boldsymbol{y}$, given the past of $\boldsymbol{u}$ itself. In our Hilbert space framework this is written as

$$\mathcal{U}_t^+ \perp \mathcal{Y}_t^- \mid \mathcal{U}_t^- \tag{7.10}$$

This condition expresses the fact that the future time evolution of the process $\boldsymbol{u}$ is not "influenced" by the past of $\boldsymbol{y}$ once the past of $\boldsymbol{u}$ is known. This captures in a coordinate free way the absence of feedback (from $\boldsymbol{y}$ to $\boldsymbol{u}$). Taking $\mathcal{A} = \mathcal{U}_t^+$ in condition (iii) of Lemma 7.4 above, the feedback-free condition is seen to be equivalent to $\mathcal{Y}_t^- \perp \mathcal{U} \mid \mathcal{U}_t^-$ and hence, from (iv), to $\mathrm{E}\{\mathcal{Y}_t^- \mid \mathcal{U}\} = \mathrm{E}\{\mathcal{Y}_t^- \mid \mathcal{U}_t^-\}$, so that

$$\mathrm{E}\{\boldsymbol{y}(t) \mid \mathcal{U}\} = \mathrm{E}\{\boldsymbol{y}(t) \mid \mathcal{U}_{t+1}^-\}, \quad \text{for all } t \in \mathbb{Z}, \tag{7.11}$$

meaning that $\mathrm{E}\{\boldsymbol{y}(t) \mid \mathcal{U}\}$ depends only on the past and present values of the process $\boldsymbol{u}$ but not on its future history. We take this as the definition of *causality*. In this case (and only in this case), it is appropriate to call $\boldsymbol{u}$ an *input* variable. One says that there is *causality from* $\boldsymbol{u}$ *to* $\boldsymbol{y}$ (or $\boldsymbol{u}$ "causes" $\boldsymbol{y}$).

Let us consider the feedback interconnection of fig. 7.1 and let $F = N_F D_F^{-1}$, $H = D_H^{-1} N_H$ be coprime matrix fraction descriptions of the transfer functions of the direct and feedback channel. An important technical condition often used in this this paper is that the input process is "sufficiently rich", in the sense that $\mathcal{U}$ admits the direct sum decomposition $\mathcal{U} = \mathcal{U}_t^- + \mathcal{U}_t^+$ for all t. Obviously, sufficient richness is equivalent to

$$\mathcal{U}_t^- \cap \mathcal{U}_t^+ = \{0\} \tag{7.12}$$

Various conditions ensuring sufficient richness are known. For example, it is well-known that for a full-rank p.n.d. process $\boldsymbol{u}$ to be sufficiently rich it is necessary and sufficient that the determinant of the spectral density matrix Φ_u should have no zeros on the unit circle [9]. Using this criterion the following result follows readily.

LEMMA 7.5
The input process is sufficiently rich, i.e. Φ_u does not have unitary zeros, if and only if both of the following conditions are satisfied:

1. D_F has no zeros on the unit circle
2. $N_H G$ and $D_H K$ do not have common zeros on the unit circle with coincident corresponding left zero directions (this we simply say: do not vanish simultaneously on the unit circle).

□

7.4 Oblique Markovian Splitting Subspaces

The idea of (stochastic) state space is the fundamental concept of stochastic realization theory. In the classical setting, i.e. stochastic time series modelling, a state space is characterized by the property of being a *Markovian splitting subspace* [18, 19]. This idea captures in a coordinate-free way the structure of stochastic

state-space models and lies at the grounds of their many useful properties. Realization of stochastic processes (without inputs) has been investigated by a number of authors, including [25, 2, 1, 15, 16, 18, 19].

The intuitive idea of a stochastic state space model with inputs is different from that of a state-space model of a single process since in the former type of models one wants to describe the effect of the exogenous input $\boldsymbol{u}$ on the output process *without modeling the dynamics of* $\boldsymbol{u}$. This is also in agreement with the aim of most identification experiments, where one is interested in describing the dynamics of the "open loop" system only and does not want to worry about finding a dynamic description of $\boldsymbol{u}$ at all. Hence the concept of state-space has to be generalized to the new setting. This will lead to the introduction of the concepts of oblique (conditional) Markov and oblique (conditional) splitting. The idea behind these definitions is to factor out the dynamics of the input process, which should not be modeled explicitly. When applying classical realization theory to the joint input-output process $(\boldsymbol{y}\ \boldsymbol{u})$ the dynamics of $\boldsymbol{u}$ will also be modeled.

At the end of the paper we shall see some connections between classical stochastic realizations of the joint input-output process and realization of $\boldsymbol{y}$ in terms of $\boldsymbol{u}$ as we are studying in this chapter.

Note that since the input is an observed variable, one is generally interested in realization which are causal with respect to $\boldsymbol{u}$. For this reason in this more general setting one should not expect the same mathematical symmetry between past and future as in the time-series case.

A source of difficulty in dealing with realization with inputs is the possible presence of feedback from $\boldsymbol{y}$ to $\boldsymbol{u}$. In this paper we shall strive to keep a certain level of generality without making too restrictive assumptions regarding the presence of feedback between $\boldsymbol{y}$ and $\boldsymbol{u}$. Dealing with feedback is a necessity in the design of identification algorithms and the complications we incur are not searched for just for academic sake of generality.

Later on we shall specialize to the case when there is no feedback from $\boldsymbol{y}$ to $\boldsymbol{u}$ and a much simpler and elegant theory will emerge.

All Hilbert spaces that we consider will be subspaces of an ambient Hilbert space $\mathcal{H}$ containing $\mathcal{U}$ and $\mathcal{Y}$ and equipped with a shift operator under which $\boldsymbol{y}$ and $\boldsymbol{u}$ are jointly stationary. We shall also assume that $\mathcal{H}$ has a finite number of generators i.e. there are $N < \infty$ random variables $\{h_1, \ldots, h_N\}$ such that

$$\overline{\text{span}}\{\sigma^t h_k \mid k = 1, \ldots, N,\ t \in \mathbb{Z}\} = \mathcal{H}$$

This is sometimes referred to by saying that $\mathcal{H}$ (or the shift σ acting on $\mathcal{H}$) has *finite multiplicity*. The multiplicity is certainly finite in the interesting special case where

$$\mathcal{H} = \mathcal{Y} \vee \mathcal{U} \tag{7.13}$$

Let $\mathcal{X}$ be a subspace of $\mathcal{H}$ and define the *stationary family of translates*, $\{\mathcal{X}_t\}$, by: $\mathcal{X}_t := \sigma^t \mathcal{X}, \quad t \in \mathbb{Z}$. The *past* and *future* of $\{\mathcal{X}_t\}$ at time t are

$$\mathcal{X}_t^- := \vee_{s<t} \mathcal{X}_s, \qquad \mathcal{X}_t^+ := \vee_{s \geq t} \mathcal{X}_s.$$

Generalizing a construction of stochastic realization theory (see e.g. [18]), we define a pair of subspaces $(\mathcal{S}, \bar{\mathcal{S}})$ attached to a given $\mathcal{X}$, as follows:

$$\mathcal{S} = \mathcal{P}^- \vee \mathcal{X}^- \tag{7.14}$$

where $\mathcal{P}^-$ is a shorthand for the *joint past* space $\mathcal{P}_t^- := \mathcal{U}_t^- \vee \mathcal{Y}_t^-$ at time $t = 0$; $\mathcal{S}$ will be called the *incoming* subspace associated to $\mathcal{X}$, while

$$\bar{\mathcal{S}} = \mathcal{Y}^+ \vee \mathcal{X}^+ \tag{7.15}$$

will be called the corresponding *outgoing* subspace.

Recall that in classical stochastic realization the subspaces $\mathcal{S}$ and $\bar{\mathcal{S}}$ are incoming and outgoing subspaces for the shift σ, in the sense of Lax-Phillips Scattering Theory [24]. In the present setting $\bar{\mathcal{S}}$ is not outgoing as it does not satisfy $\cup_t \bar{\mathcal{S}}_t = \mathcal{H}$. However, it will turn out to be convenient to keep the same terminology to make connections with classical stochastic realization theory, especially when studying minimality. Later we shall also introduce "extended" versions of $\mathcal{S}$ and $\bar{\mathcal{S}}$ which will in fact be incoming and outgoing subspaces in the sense of Lax-Phillips.

DEFINITION 7.2
The family $\{\mathcal{S}_t\}$ is (forward) *purely-non-deterministic* (p.n.d.) if

$$\bigcap_{t<0} \mathcal{S}_t = \{0\} \tag{7.16}$$

The subspace $\mathcal{X}$ is then called (forward) *purely-non-deterministic* (p.n.d.) whenever the associated incoming subspace has the p.n.d. property. □

Let us define the sequence of *wandering* subspaces $\mathcal{W}_t = \sigma^t \mathcal{W}$ associated to $\{\mathcal{S}_t\}$ as:

$$\mathcal{W}_t := \mathcal{S}_{t+1} \ominus (\mathcal{S}_t + \mathcal{U}_t)\,. \tag{7.17}$$

LEMMA 7.6
The wandering subspaces are pairwise orthogonal, i.e. $\mathcal{W}_t \perp \mathcal{W}_s$, $\forall\, t \neq s$. □

Proof Let us assume $t > s$; by construction we have $\mathcal{W}_s \subseteq \mathcal{S}_{s+1}$ while clearly $\mathcal{W}_t \perp \mathcal{S}_t$. Since $\mathcal{S}_t$ is non-decreasing (backward-shift invariant), i.e. $\mathcal{S}_s \subseteq \mathcal{S}_t$, it follows that $\mathcal{W}_t \perp \mathcal{S}_s$ and hence $\mathcal{W}_t \perp \mathcal{W}_s$. □

It follows from (7.17) that the incoming subspace admits the decomposition

$$\mathcal{S}_{t+1} = (\mathcal{S}_t + \mathcal{U}_t) \oplus \mathcal{W}_t \tag{7.18}$$

For future reference we note the following fact:

FACT 7.1
The future $\mathcal{W}_t^+$ is orthogonal to $\mathcal{S}_t + \mathcal{U}_t$. □

Since $\mathcal{H}$ has finite multiplicity, the wandering subspace $\mathcal{W}$ is finite-dimensional and admits an orthonormal basis $w(0)$. It follows that $\mathcal{W}_t^- = \mathbf{H}_t^-(w)$ where $w(t) = \sigma^t w(0)$ is a normalized white noise process which is called the (forward) *generating process* of $\mathcal{X}$.

The following are basic definitions which capture the notion of state space in presence of exogenous inputs.

DEFINITION 7.3
The subspace $\mathcal{X}$ is (forward) *oblique Markovian*, if $\mathcal{U}_t \cap \left(\mathcal{X}_{t+1}^- \vee \mathcal{U}_t^-\right) = \{0\}$ and the following equality holds:

$$E_{||\mathcal{U}_t}[\mathcal{X}_{t+1} \mid \mathcal{X}_t^- \vee \mathcal{U}_t^-] = E_{||\mathcal{U}_t}[\mathcal{X}_{t+1} \mid \mathcal{X}_t]. \tag{7.19}$$

We shall say that $\mathcal{X}$ is *causal* oblique Markovian if $\mathcal{X}_t \subseteq \mathcal{P}_t^-$. □

Note that the condition $\mathcal{U}_t \cap \mathcal{U}_t^- = \{0\}$, necessary for the oblique projection to be well defined, is implied by the richness Assumption 7.12.

The oblique Markovian condition can be written in terms of conditional orthogonality as follows:

PROPOSITION 7.1
The oblique Markovian property (7.19) is equivalent to

$$E\left[\mathcal{X}_{t+1} \mid \mathcal{X}_{t+1}^- \vee \mathcal{U}_{t+1}^-\right] = E[\mathcal{X}_{t+1} \mid \mathcal{X}_t + \mathcal{U}_t] \tag{7.20}$$

and hence to the following conditional orthogonality property

$$\mathcal{X}_{t+1} \perp (\mathcal{X}_t^- \vee \mathcal{U}_t^-) \mid (\mathcal{X}_t + \mathcal{U}_t) \tag{7.21}$$

which can be interpreted by saying that $\mathcal{X}_t$ is *conditionally Markov* given $\mathcal{U}_t$. □

Proof Setting $\mathcal{A} = \mathcal{X}_{t+1}$, $\mathcal{B} = \mathcal{X}_{t+1}$, $\mathcal{C} = \mathcal{X}_t^- \vee \mathcal{U}_t^-$ and $\mathcal{D} = \mathcal{U}_t$ in Lemma 7.2 the oblique Markovian property (7.19) is seen to be equivalent to (7.20). The conditional orthogonality follows from Lemma 7.4. □

To be honest, the oblique Markovian property of the definition should be named "one-step-ahead" oblique Markovian property. For it is in general not guaranteed that the sufficient statistic property of $\mathcal{X}_t$ holds also when one wants to predict random variables in the distant future $\mathcal{X}_{t+k}$, $k > 1$. However we shall see later that the extension of the "one-step-ahead" oblique Markovian property to an arbitrary number of steps ahead, holds when there is no feedback from $\boldsymbol{x}$ to $\boldsymbol{u}$, in which case condition (7.19) is equivalent to

$$E_{||\mathcal{U}_t^+}\left[\mathcal{X}_t^+ \mid \mathcal{X}_{t+1}^- \vee \mathcal{U}_t^-\right] = E_{||\mathcal{U}_t^+}\left[\mathcal{X}_t^+ \mid \mathcal{X}_t\right]$$

Unfortunately in general (7.19) is not equivalent to the above. In fact, we shall see that the property of sufficient statistics for the whole future will hold only "conditionally", given the subspace generated by *all* future inputs to the realization with state space $\mathcal{X}_t$, where "inputs" now means input signal which may be observable and not. To make this precise, we shall have to introduce the "extended" or *joint future input space*

$$\mathcal{F}_t^+ := \left(\mathcal{U}_t^+ \vee \mathcal{W}_t^+\right)$$

of the p.n.d. subspace $\mathcal{X}$.

The joint future plays a role in generalizing the fundamental concept of splitting to the *oblique splitting* property defined below.

DEFINITION 7.4
A subspace $\mathcal{X}$ is (forward) *oblique splitting* for $(\mathcal{Y}, \mathcal{U})$, if

$$\mathcal{Y}_t^+ \perp \mathcal{P}_t^- \mid [\mathcal{X}_t \vee \mathcal{F}_t^+] \tag{7.22}$$

We will say that $\mathcal{X}$ is a *causal* oblique splitting subspace if $\mathcal{X}_t \subseteq \mathcal{P}_t^-$. □

Condition (7.22) says that once the (real plus unobservable white) future inputs are given, the information in the present state space $\mathcal{X}_t$, is equivalent to the knowledge of all the (joint) past history of state input and output, for the purpose of predicting the future of $\boldsymbol{y}$. Indeed, provided that

$$\mathcal{X}_t \cap \mathcal{F}_t^+ = \{0\}, \tag{7.23}$$

by using Lemma 7.2, it follows that (7.22) can be expressed using oblique projections

$$E_{\|\mathcal{F}_t^+}[\mathcal{Y}_t^+ \mid \mathcal{X}_t \vee \mathcal{P}_t^-] = E_{\|\mathcal{F}_t^+}[\mathcal{Y}_t^+ \mid \mathcal{X}_t]$$

Unfortunately, there may be situations in which condition (7.23) does not hold, for virtually all oblique Markovian splitting subspaces $\mathcal{X}_t$. Intuitively, this will be the case when the transfer function $F(z)$ in the forward loop in Fig. 7.1 is not stable (which can happen, even if the feedback interconnection is internally stable).

Another difficulty related to the presence of feedback is that it may happen that $(\mathcal{U}^+ \vee \mathcal{W}^+) \cap (\mathcal{U}^- \vee \mathcal{W}^-) \neq \{0\}$. This fact would make some oblique projection formulas meaningless.

A sufficient condition for zero intersection is given in following Proposition, whose proof we shall leave to the reader.

PROPOSITION 7.2
The joint spectral density matrix $\Phi\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{w}\end{bmatrix}$, has no zeros on the unit circle, or, equivalently, $(\mathcal{U}^+ \vee \mathcal{W}^+) \cap (\mathcal{U}^- \vee \mathcal{W}^-) = \{0\}$, if and only if, with the same notation of remark 7.5, $D_F D_H K$ does not vanish on the unit circle. □

Note that if there are no inputs, $\mathcal{X}_t \vee \mathcal{F}_t^+ = \mathcal{X}_t \oplus \mathcal{W}_t^+$ and condition (7.22) reduces to

$$\mathcal{Y}_t^+ \perp \mathcal{Y}_t^- \mid (\mathcal{X}_t \oplus \mathcal{W}_t^+)$$

and since $\mathcal{W}_t^+ \perp \mathcal{Y}_t^-$ the latter is in turn equivalent to

$$\mathcal{Y}_t^+ \perp \mathcal{Y}_t^- \mid \mathcal{X}_t$$

which is the usual splitting property.

The oblique Markov and the splitting conditions separately are not enough, in general, to fully characterize the state space in the presence of inputs. A condition which implies both (7.22) and (7.19) is the *oblique Markovian splitting* property, defined below.

DEFINITION 7.5
A subspace $\mathcal{X}$ is an *oblique Markovian splitting subspace* for the pair $(\mathcal{Y}, \mathcal{U})$ if

$$\mathcal{U}_t \cap (\mathcal{X}_{t+1}^- \vee \mathcal{P}_t^-) = \{0\} \tag{7.24}$$

and

$$E_{||\mathcal{U}_t}[\mathcal{Y}_t \vee \mathcal{X}_{t+1} \,|\, \mathcal{X}_{t+1}^- \vee \mathcal{P}_t^-] = E_{||\mathcal{U}_t}[\mathcal{Y}_t \vee \mathcal{X}_{t+1} \,|\, \mathcal{X}_t]. \tag{7.25}$$

We shall say that $\mathcal{X}$ is a *causal* oblique Markovian splitting subspace if $\mathcal{X} \subseteq \mathcal{P}^-$ □

This condition is precisely what is needed for the space $\mathcal{X}$ to qualify as a state space for a stochastic model described by equations of the form (7.1)

Note that the "extended richness condition"

$$\mathcal{U}_t \cap \mathcal{P}_t^- = \{0\} \tag{7.26}$$

is a necessary condition for the oblique projection to be well-defined. This property will be necessary in order to be able to derive unique state-space equations.

PROPOSITION 7.3
If the joint spectrum of $\boldsymbol{y}$ and $\boldsymbol{u}$ is coercive then the zero intersection property (7.26) holds. □

Proof If the joint spectrum is coercive, then one has that

$$(\mathcal{U}_t^+ \vee \mathcal{Y}_t^+) \cap (\mathcal{U}_t^- \vee \mathcal{Y}_t^-) = \{0\}$$

and then in particular (7.26). □

REMARK 7.6
Again, the property that we would like to hold is

$$E_{||\mathcal{U}_t^+}[\mathcal{Y}_t^+ \vee \mathcal{X}_{t+1}^+ \,|\, \mathcal{X}_{t+1}^- \vee \mathcal{P}_t^-] = E_{||\mathcal{U}_t^+}[\mathcal{Y}_t^+ \vee \mathcal{X}_{t+1}^+ \,|\, \mathcal{X}_t] \tag{7.27}$$

however this condition is in general not equivalent to (7.25). In fact, when feedback is present (the past output space $\mathcal{Y}_t^-$ is not conditionally uncorrelated with the future inputs $\mathcal{U}_t^+$ given the past past inputs $\mathcal{U}_t^-$), one has

$$E[\mathcal{Y}_t \,|\, \mathcal{X}_t^- \vee \mathcal{P}_t^- \vee \mathcal{U}_t^+] \neq E[\mathcal{Y}_t \,|\, , \mathcal{X}_t^- \vee \mathcal{P}_t^- \vee \mathcal{U}_t]$$

so that requiring the stronger condition (7.27) would make the state space "unnecessarily large". We shall see later (see Lemma 7.11) that when there is no feedback from $\begin{bmatrix} \boldsymbol{x}^\top & \boldsymbol{y}^\top \end{bmatrix}$ to $\boldsymbol{u}$, condition (7.25) is equivalent to (7.27). □

The result which follows is in the same spirit of Proposition 7.1.

PROPOSITION 7.4
The oblique Markovian splitting property (7.25) is equivalent to

$$E\left[\mathcal{X}_{t+1} \vee \mathcal{Y}_t \,|\, \mathcal{X}_{t+1}^- \vee \mathcal{U}_t \vee \mathcal{P}_t^-\right] = E[\mathcal{X}_{t+1} \vee \mathcal{Y}_t \,|\, \mathcal{X}_t + \mathcal{U}_t]. \tag{7.28}$$

and hence to the conditional orthogonality property:

$$(\mathcal{X}_{t+1} \vee \mathcal{Y}_t) \perp (\mathcal{X}_t^- \vee \mathcal{P}_t^-) \mid (\mathcal{X}_t + \mathcal{U}_t)$$

□

Proof Letting $\mathcal{A} = (\mathcal{X}_{t+1} \vee \mathcal{Y}_t)$, $\mathcal{B} = \mathcal{X}_t$, $\mathcal{C} = \mathcal{X}_{t+1}^- \vee \mathcal{P}_t^-$ and $\mathcal{D} = \mathcal{U}_t$ in Lemma 7.2, the oblique Markovian splitting property (7.25) is seen to be equivalent to (7.28). □

PROPOSITION 7.5
Under condition (7.23) the oblique Markovian splitting property (7.25) is equivalent to

$$E\left[\mathcal{X}_{t+1}^+ \vee \mathcal{Y}_t^+ \,|\, \mathcal{X}_{t+1}^- \vee \mathcal{F}_t^+ \vee \mathcal{P}_t^-\right] = E[\mathcal{X}_{t+1}^+ \vee \mathcal{Y}_t^+ \,|\, \mathcal{X}_t + \mathcal{F}_t^+]. \tag{7.29}$$

which can also be written in form of conditional orthogonality as

$$\left(\mathcal{X}_{t+1}^+ \vee \mathcal{Y}_t^+\right) \perp (\mathcal{X}_t^- \vee \mathcal{P}_t^-) \mid \left(\mathcal{X}_t + \mathcal{F}_t^+\right)$$

□

Proof Assume (7.29) holds. It follows that

$$E[\mathcal{X}_{t+1} \vee \mathcal{Y}_t \,|\, \mathcal{X}_{t+1}^- \vee \mathcal{P}_t^- \vee \mathcal{F}_t^+] \subseteq \mathcal{X}_t + \mathcal{F}_t^+$$

However, from $\mathcal{X}_{t+1} \vee \mathcal{Y}_t \subseteq \mathcal{S}_{t+1}$ we obtain

$$E[\mathcal{X}_{t+1} \vee \mathcal{Y}_t \,|\, \mathcal{X}_{t+1}^- \vee \mathcal{P}_t^- \vee \mathcal{F}_t^+] \subseteq \mathcal{S}_t + \mathcal{U}_t \oplus \mathcal{W}_t.$$

Using the fact that (7.23) implies $\mathcal{S}_t \cap \mathcal{F}_t^+ = \{0\}$ it must be that

$$E[\mathcal{X}_{t+1} \vee \mathcal{Y}_t \,|\, \mathcal{X}_{t+1}^- \vee \mathcal{P}_t^- \vee \mathcal{U}_t \vee \mathcal{W}_t] = E[\mathcal{X}_{t+1} \vee \mathcal{Y}_t \,|\, \mathcal{X}_{t+1}^- \vee \mathcal{P}_t^- \vee \mathcal{F}_t^+] \subseteq \mathcal{X}_t + \mathcal{U}_t \oplus \mathcal{W}_t$$

which clearly implies (7.25). The proof of the converse will be given after Theorem 7.7. □

COROLLARY 7.1
Under condition (7.23) the oblique Markovian splitting property (7.25) is equivalent to

$$E_{\| \mathcal{F}_t^+}\left[\mathcal{X}_{t+1}^+ \vee \mathcal{Y}_t^+ \,|\, \mathcal{X}_{t+1}^- \vee \mathcal{P}_t^-\right] = E_{\| \mathcal{F}_t^+}\left[\mathcal{X}_{t+1}^+ \vee \mathcal{Y}_t^+ \,|\, \mathcal{X}_t\right]. \tag{7.30}$$

□

Proof Setting $\mathcal{A} = \left(\mathcal{X}_{t+1}^{+} \vee \mathcal{Y}_t^{+}\right)$, $\mathcal{B} = \mathcal{X}_t$, $\mathcal{C} = \mathcal{X}_{t+1}^{-} \vee \mathcal{P}_t^{-}$ and $\mathcal{D} = \mathcal{F}_t^{+}$ in Lemma 7.2, we have that (7.30) is equivalent to (7.29) and therefore, from Proposition 7.5 to (7.25). □

The following result is a coordinate-free version of the equivalence between oblique Markovian splitting property and representability by a state space model of the form (7.1). The Theorem holds without finite dimensionality assumptions.

THEOREM 7.7
Let $\mathcal{X}_t$ be a p.n.d. oblique Markovian splitting subspace for $(\mathcal{Y}, \mathcal{U})$; then the following inclusions hold

$$\mathcal{X}_{t+1} \subseteq (\mathcal{X}_t + \mathcal{U}_t) \oplus \mathcal{W}_t \tag{7.31}$$

$$\mathcal{Y}_t \subseteq (\mathcal{X}_t + \mathcal{U}_t) \oplus \mathcal{W}_t \tag{7.32}$$

□

Proof Since $\mathcal{X}_{t+1} \subseteq \mathcal{S}_{t+1}$ using the decomposition $\mathcal{S}_{t+1} = (\mathcal{S}_t + \mathcal{U}_t) \oplus \mathcal{W}_t$ we obtain :

$$\begin{aligned}
\mathcal{X}_{t+1} &= E\left[\mathcal{X}_{t+1} \mid \mathcal{S}_{t+1}\right] \\
&= E\left[\mathcal{X}_{t+1} \mid (\mathcal{S}_t + \mathcal{U}_t) \oplus \mathcal{W}_t\right] \\
&\subseteq E\left[\mathcal{X}_{t+1} \mid (\mathcal{S}_t + \mathcal{U}_t)\right] \oplus \mathcal{W}_t \\
&\subseteq \left(E_{\| \mathcal{U}_t}\left[\mathcal{X}_{t+1} \mid \mathcal{S}_t\right] + \mathcal{U}_t\right) \oplus \mathcal{W}_t \\
&\subseteq (\mathcal{X}_t + \mathcal{U}_t) \oplus \mathcal{W}_t
\end{aligned}$$

where the last equality follows from (7.25). A completely analogous derivation holds for the second inclusion. □

We can now complete the proof of Proposition 7.5. To this purpose it will be handy to introduce a notation for vector sum of subspaces of the type

$$\mathcal{U}_{[t,t+k)} := \mathcal{U}_t + \mathcal{U}_{t+1} + \ldots + \mathcal{U}_{t+k-1}$$

Similar notations will be used without further comments in the following.

Proof of Proposition 7.5 Conversely, assume (7.25) holds. It follows from Theorem 7.7 that, for every $k \geq 0$

$$\mathcal{X}_{t+k+1} \vee \mathcal{Y}_{t+k} \subseteq \mathcal{X}_t + \mathcal{U}_{[t,t+k)} + \mathcal{W}_{[t,t+k)}$$

which implies that

$$\mathcal{X}_{t+1}^{+} \vee \mathcal{Y}_t^{+} \subseteq \mathcal{X}_t + \mathcal{U}_t^{+} + \mathcal{W}_t^{+}$$

and therefore (7.29). □

When $\mathcal{X}$ is finite-dimensional, we can obtain state-space representations of the form (7.1) just by choosing a basis in the subspaces $\mathcal{X}$ and $\mathcal{W}$. Conversely, given a finite-dimensional state-space model of the form (7.1), it is easy to check that the subspace generated by the components of the state vector

$$\mathcal{X}_t := \operatorname{span}\{x_1(t), \ldots, x_n(t)\}$$

is an oblique Markovian splitting subspace. We shall leave this check to the reader.

By using the representation Theorem 7.7 we can show that the oblique Markovian splitting property implies both the oblique Markovian and the oblique splitting property.

PROPOSITION 7.6
The oblique Markovian splitting property implies oblique Markovian and oblique splitting, i.e. (7.25) implies both (7.19) and (7.22). □

Proof Projecting both members of (7.31) along $\mathcal{U}_t$ we obtain

$$E_{\|\mathcal{U}_t}[\mathcal{X}_{t+1} \mid \mathcal{X}_t \vee \mathcal{U}_t^-] \subset \mathcal{X}_t$$

which is the oblique Markovian property (7.19). Combining (7.32) and (7.31) we obtain

$$\mathcal{Y}_t^+ \subseteq \mathcal{X}_t \vee \mathcal{U}_t^+ \vee \mathcal{W}_t^+ = \mathcal{X}_t \vee \mathcal{F}_t^+$$

which implies (7.22). If (7.23) holds, the projection of any element $y^+ \in \mathcal{Y}_t^+$ onto $\mathcal{X}_t \vee \mathcal{P}_t^-$ along $\mathcal{F}_t^+$ is equal to the projection of its (unique in this case) component in $\mathcal{X}_t$. In other words

$$E_{\|\mathcal{F}_t^+}\left[\, y^+ \,|\, \mathcal{X}_t \vee \mathcal{P}_t^- \right] = E_{\|\mathcal{F}_t^+}\left[\, y^+ \,|\, \mathcal{X}_t \right], \qquad y^+ \in \mathcal{Y}_t^+.$$

□

7.5 Acausality of Realizations with Feedback

Stationary models for the pair $(\boldsymbol{y}, \boldsymbol{u})$ of the form (7.1), or in symbolic notation (z-transform),

$$\boldsymbol{y}(t) = F(z)\boldsymbol{u}(t) + G(z)\boldsymbol{w}(t)$$

tend to give for granted that $\boldsymbol{y}$ depends *causally* on the input signals $\boldsymbol{u}$, $\boldsymbol{w}$. This is in general false if we are in the presence of feedback.

Certainly causality holds as long as $F(z)$ and $G(z)$ are stable, or, equivalently, $|\lambda(A)| < 1$ in the model (7.1). However, in the presence of feedback the pair $(\boldsymbol{y}, \boldsymbol{u})$ may well be stationary even if $F(z)$ is not stable. In this situation, the eigenvalues of A may lie anywhere, in particular some may be (strictly) outside of the unit circle. Then, the customary interpretation of the state space model (7.1) as a *forward difference equation*, does not make sense. Unstable modes must be integrated *backwards* see [29]. In general when there is feedback, past outputs may be influenced by future inputs, which, according to what has been seen above, means that the model is *not causal*. We shall briefly analyze how this acausality shows up and what are the consequences. For simplicity of exposition we analyze only the finite dimensional case.

Consider a basis for (7.1) so that the matrix A is of the block-diagonal form

$$A = \begin{pmatrix} A_- & 0 & 0 \\ 0 & A_0 & 0 \\ 0 & 0 & A_+ \end{pmatrix} \tag{7.33}$$

where $|\lambda(A_-)| < 1$, $|\lambda(A_0)| = 1$, $|\lambda(A_+)| > 1$. Correspondingly we shall denote with $\mathcal{X}_-$ the "stable

manifold", with $\mathcal{X}_+$ the "unstable manifold" and with $\mathcal{X}_0$ the "central manifold" of the state space. The symbols $\boldsymbol{x}_-$, $\boldsymbol{x}_+$ and $\boldsymbol{x}_0$ will denote the bases in the

corresponding spaces. A similar meaning will be attributed to the symbols B_-, B_+, B_0, G_-, G_+ and G_0. Hence the state space equation can be rewritten in decoupled from as follows:

$$\left\{\begin{array}{rcl} x_-(t+1) &=& A_-x_-(t) + B_-u(t) + G_-w(t) \\ x_0(t+1) &=& A_0x_0(t) + B_0u(t) + G_0w(t) \\ x_-(t+1) &=& A_+x_+(t) + B_+u(t) + G_+w(t) \end{array}\right. \tag{7.34}$$

The three difference equation can be thought of as running forward, forward or backward, and backward respectively. The interpretation of the equation on the "central manifold" is somewhat delicate and we shall not insist on this point here, however see [5, p. 105]. May it suffice to say that this component belongs to both past $\mathcal{U}_t^- \vee \mathcal{W}_t^-$ and future $\mathcal{U}_t^+ \vee \mathcal{W}_t^+$. This is not a contradiction as pointed out in remark 7.2. Concerning the first and the third block, it is trivial to recognize that these may be thought as a stable difference equation running forward and an unstable difference equation running backward in time. This implies that $x_-(t) \in \mathcal{U}_t^- \vee \mathcal{W}_t^-$ and $x_+(t) \in \mathcal{U}_t^+ \vee \mathcal{W}_t^+ = \mathcal{F}_t^+$. From this we see that in general one cannot assume that $\mathcal{X}_t \cap \mathcal{F}_t^+ = \{0\}$. This fact is the source of a number of complications.

To avoid these complications we shall henceforth restrict to the case $|\lambda(A)| < 1$ (i.e. we have a feedback interconnection with a stable forward loop transfer function $F(z)$) and postpone the discussion of the general case to future publications.

We formalize this assumption below.

ASSUMPTION 7.8
The joint spectrum of u and w is coercive (see remark 7.2) and the poles of $F(z)$ lie strictly inside the unit circle. □

Observability, Constructibility and Minimality

Minimality is a fundamental property of state space models. The concept can be described purely in geometrical terms as in the following definition

DEFINITION 7.9
An oblique Markovian splitting subspace $\mathcal{X}$ is minimal if it does not contain properly other oblique Markovian splitting subspaces. □

Structural properties which are instrumental in the study of minimality are *observability and constructibility* of an oblique Markovian splitting subspace $\mathcal{X}_t$. One measures the "observability" of $\mathcal{X}_t$ on the basis of its "ability" of predicting future outputs "given" (in an appropriate sense), the future inputs.

Let $\mathcal{X}_t$ be an oblique Markovian splitting subspace, and introduce the (adjoint) *observability operator*

$$\mathbb{O}^* : \mathcal{Y}^+ \to \mathcal{X} \quad , \quad \mathbb{O}^*\lambda := E_{\|\mathcal{F}^+}[\lambda \mid \mathcal{X}], \quad \lambda \in \mathcal{Y}^+ \tag{7.35}$$

The subspace

$$\mathcal{X}_t^o := \overline{\text{Range } \mathbb{O}^*} = E_{\|\mathcal{F}_t^+}[\mathcal{Y}_t^+ \mid \mathcal{X}_t] \tag{7.36}$$

will be called the *observable subspace of $\mathcal{X}_t$ given $\mathcal{F}_t^+$*.

DEFINITION 7.10
We shall say that $\mathcal{X}_t$ is *observable* given $\mathcal{F}_t^+$ if $\mathcal{X}_t^o = \mathcal{X}_t$ or, equivalently, if the operator $\mathbb{O}^*$ has dense range. □

Similarly we may consider the "constructibility" property of $\mathcal{X}_t$. Let

$$\mathbb{K} : \mathcal{X} \to \mathcal{P}^- \quad , \quad \mathbb{K}\xi := \boldsymbol{E}_{\|\mathcal{F}^+}[\xi \mid \mathcal{P}^-], \quad \xi \in \mathcal{X}. \tag{7.37}$$

be the "constructibility" operator. It measures the degree of predictability of an element of $\mathcal{X}$ based on the joint past $\mathcal{P}^-$ given the future inputs $\mathcal{F}^+$.

DEFINITION 7.11
Let $\mathcal{X}_t$ be an oblique Markovian splitting subspace. The closure of the range of the adjoint constructibility operator is called the "constructible part" of $\mathcal{X}_t$ and denoted as $\mathcal{X}_t^c$.

We shall say that $\mathcal{X}_t$ is *constructible* if $\mathcal{X}_t^c = \mathcal{X}_t$. □

PROPOSITION 7.7
The constructible part $\mathcal{X}_t^c$ of $\mathcal{X}_t$ is given by:

$$\mathcal{X}_t^c = \mathcal{X}_t \ominus \text{Ker } \mathbb{K}$$

□

Proof This is immediate from the fact that

$$\mathcal{H} = \overline{\text{Range } \mathfrak{C}^*} \oplus (\text{Ker } \mathbb{K})^\perp$$

for any bounded linear operator. □

A central goal of this paper will be to prove the following criterion for minimality. The proof will be given in the following.

THEOREM 7.12
An oblique Markovian splitting subspace $\mathcal{X}$ is minimal if and only if it is both observable and constructible. □

Causal Oblique Markovian Splitting Subspaces

In this section we will restrict our attention to *causal* oblique Markovian splitting subspaces, namely we will require that

$$\mathcal{X} \subseteq \mathcal{P}^-. \tag{7.38}$$

In this case clearly

$$\bigvee_{t=-\infty}^{\infty} \mathcal{X}_t \subset \mathcal{Y} \vee \mathcal{U}$$

and hence the ambient space can be taken to be

$$\mathcal{H} = \mathcal{Y} \vee \mathcal{U}. \tag{7.39}$$

The corresponding realizations are called "internal". The motivation for this restriction is that in system identification we want to construct the state space from the available data, which (ideally) generate the subspace $\mathcal{Y} \vee \mathcal{U}$.

Once we restrict to the causal situation, since (7.38) implies that $\mathcal{X}^- \subseteq \mathcal{P}^-$, the incoming subspace is given by

$$\mathcal{S} = \mathcal{X}^- \vee \mathcal{P}^- = \mathcal{P}^-.$$

Therefore the incoming subspace coincides with $\mathcal{P}^-$ in the causal situation. We shall denote by $\mathcal{E}_t$ the wandering subspace which generates $\mathcal{P}^-$

$$\mathcal{P}^-_{t+1} = (\mathcal{P}^-_t + \mathcal{U}_t) \oplus \mathcal{E}_t \tag{7.40}$$

Note that $\mathcal{P}^-$ is p.n.d. by assumption.

Assumption 7.8 allows to "iterate" (7.40) so that

$$\mathcal{P}^-_{t+1} = (\mathcal{E}^-_t + \mathcal{U}^-_{t+1}) \oplus \mathcal{E}_t$$

It is common use to take as a basis for $\mathcal{E}_t$ the *innovation* $e(t)$ defined by

$$e(t) := \boldsymbol{y}(t) - E\left[\boldsymbol{y}(t) \mid \mathcal{P}^-_t \vee \mathcal{U}_t\right] \tag{7.41}$$

which is a (non normalized) white noise process whose variance is positive definite by the full-rank assumption.

We are eventually in a position to give a procedure to *construct* an oblique Markovian splitting subspace. The construction is motivated by the definition of observability (7.36) [1]

Define the *oblique predictor space* $\mathcal{X}^{+/-}_t$ at time t as follows; let $\mathcal{G}_t := \mathcal{U}_t \oplus \mathcal{E}_t$ and let

$$\mathcal{X}^{+/-}_t := E_{\|\mathcal{G}^+_t}\left[\mathcal{Y}^+_t \mid \mathcal{P}^-_t\right] \tag{7.42}$$

Obviously $\mathcal{X}^{+/-}_t$ is contained in $\mathcal{P}^-_t$ and is oblique splitting. Let us prove that it is oblique Markovian splitting.

PROPOSITION 7.8
The predictor space $\mathcal{X}^{+/-}_t$ is a causal oblique Markovian splitting subspace □

Proof It suffices to prove that

$$E_{\|\mathcal{G}^+_t}\left[\mathcal{X}^{+/-}_{t+1} \mid \mathcal{P}^-_t\right] \subseteq \mathcal{X}^{+/-}_t$$

but this is trivial since

$$\begin{aligned} E_{\|\mathcal{G}^+_t}\left[\mathcal{X}^{+/-}_{t+1} \mid \mathcal{P}^-_t\right] &= E_{\|\mathcal{G}^+_t}\left[E_{\|\mathcal{G}^+_{t+1}}\left[\mathcal{Y}^+_{t+1} \mid \mathcal{P}^-_{t+1}\right] \mid \mathcal{P}^-_t\right] \\ &= E_{\|\mathcal{G}^+_t}\left[E\left[\mathcal{Y}^+_{t+1} \mid \mathcal{P}^-_{t+1} + \mathcal{G}^+_{t+1}\right] \mid \mathcal{P}^-_t\right] \\ &= E_{\|\mathcal{G}^+_t}\left[\mathcal{Y}^+_{t+1} \mid \mathcal{P}^-_t\right] \\ &\subseteq \mathcal{X}^{+/-}_t \end{aligned}$$

[1] Note that any causal oblique Markovian splitting subspace will be constructible by construction.

where the last equality follows from the fact that $\mathcal{H} = \mathcal{P}_t^- + \mathcal{G}_t^+$. □

As we have anticipated in Theorem 7.12 minimality is equivalent to both constructibility and observability. Clearly in the casual case constructibility is granted for free and therefore one just need to check observability. However, for the oblique predictor space a proof of minimality can be given directly.

PROPOSITION 7.9
The oblique predictor space is the *minimal* causal oblique Markovian splitting subspace, in the sense that

$$\mathcal{X}_t^{+/-} \subseteq \mathcal{X}_t$$

for every causal oblique Markovian splitting subspace $\mathcal{X}$. □

Proof From the fact that $\mathcal{S}_t = \mathcal{X}_t^- \vee \mathcal{P}_t^- = \mathcal{P}_t^-$ we obtain

$$\mathcal{X}_t^{+/-} = E_{\|\mathcal{G}_t^+}\left[\mathcal{Y}_t^+ \mid \mathcal{P}_t^-\right] \subseteq \mathcal{X}_t$$

from which the statement follows. □

Note that in order to construct the oblique predictor space we have used the "innovation" space $\mathcal{E}_t$. Theoretically one could construct the innovation space $\mathcal{E}$ starting form the "data" $\mathcal{Y}$ and $\mathcal{U}$, using (7.40), and after that construct $\mathcal{X}_t^{+/-}$.

There is, however, a direct construction which, although somewhat complicated, permits to skip the first step of this procedure.

PROPOSITION 7.10
Define the k step ahead (oblique) predictor space

$$\begin{aligned} \mathcal{X}_t^k &:= E_{\|\mathcal{G}_t^+}\left[\mathcal{Y}_{t+k} \mid \mathcal{P}_t^-\right] \\ &= E_{\|\mathcal{U}_t}\left[E_{\|\mathcal{U}_{t+1}}\left[\cdots E_{\|\mathcal{U}_{t+k}}\left[\mathcal{Y}_{t+k} \mid \mathcal{P}_{t+k}^-\right]\cdots \mid \mathcal{P}_{t+1}^-\right] \mid \mathcal{P}_t^-\right] \end{aligned} \tag{7.43}$$

Then the oblique predictor space can be computed as the (closed) infinite vector sum

$$\mathcal{X}_t^{+/-} = \bigvee_{k\geq 0} \mathcal{X}_t^k \tag{7.44}$$

□

Proof We just need to show that (7.43) holds true. From Theorem 7.7 we have

$$\mathcal{Y}_{t+k} \subseteq \left(\mathcal{X}_{t+k}^{+/-} + \mathcal{U}_{t+k}\right) \oplus \mathcal{E}_{t+k}$$

and

$$\mathcal{X}_{t+h}^{+/-} \subseteq \left(\mathcal{X}_{t+h-1}^{+/-} + \mathcal{U}_{t+h-1}\right) \oplus \mathcal{E}_{t+h-1}.$$

These two conditions imply that

$$E_{\|\mathcal{U}_{t+k}}\left[\mathcal{Y}_{t+k} \mid \mathcal{P}_{t+k}^-\right] \subseteq \mathcal{X}_{t+k}^{+/-}$$

and

$$E_{\|\mathcal{U}_{t+h}}\left[\mathcal{X}_{t+h+1} \mid \mathcal{P}_{t+h}^-\right] \subseteq \mathcal{X}_{t+h}^{+/-}$$

which, together with

$$\mathcal{Y}_{t+k} \subseteq \mathcal{X}_t^{+/-} + \mathcal{U}_{[t,t+k)} + \mathcal{E}_{[t,t+k)}$$

imply (7.43). □

REMARK 7.13
Note that in the finite dimensional case the sum (7.44) can be limited to n terms, where n is the dimension of a minimal causal realization. □

7.6 Scattering Representations of Oblique Markovian Splitting Subspaces

In this section we shall establish some general properties of oblique Markovian splitting subspaces in order to facilitate the study of minimality.

Let $\mathcal{X}$ be an oblique Markovian splitting subspace and let $\mathcal{S}$ defined by (7.14) and $\bar{\mathcal{S}}$, defined by (7.15) be the associated incoming-outgoing pair. The oblique Markovian splitting property (7.25) can be written as

$$E_{||\mathcal{F}^+}[\bar{\mathcal{S}} \,|\, \mathcal{S}] = E_{||\mathcal{F}^+}[\bar{\mathcal{S}} \,|\, \mathcal{X}]. \tag{7.45}$$

The following Lemma gives a formal characterization of any oblique Markovian splitting subspace as the oblique predictor space of the outgoing subspace.

LEMMA 7.7
Let $(\mathcal{S}, \bar{\mathcal{S}})$ and $\mathcal{F}^+$ be as defined above. Then

$$\mathcal{X} = E_{||\mathcal{F}^+}[\bar{\mathcal{S}} \,|\, \mathcal{S}]$$

hence every $\mathcal{X}$ is the oblique predictor space of $\bar{\mathcal{S}}$, given $\mathcal{S}$, along $\mathcal{F}^+$. □

Proof Every element $\bar{s}$ of $\bar{\mathcal{S}}$ has the form $\bar{s} = y + x,\ y \in \mathcal{Y}^+,\ x \in \mathcal{X}^+$ so that $E_{||\mathcal{U}^+}[y \,|\, \mathcal{S}] = E_{||\mathcal{F}^+}[y \,|\, \mathcal{X}] \in \mathcal{X} \subseteq \mathcal{S}$. On the other hand, by definition of oblique splitting we have

$$E_{||\mathcal{F}^+}[x \,|\, \mathcal{S}] = E_{||\mathcal{F}^+}[x \,|\, \mathcal{X}] \qquad x \in \mathcal{X}^+,$$

therefore

$$\overline{\text{span}}\{E_{||\mathcal{F}^+}[\bar{s} \,|\, \mathcal{S}] \,|\, \bar{s} \in \bar{\mathcal{S}}\} = \mathcal{X}.$$

This implies that $\mathcal{X}$ is the oblique predictor space of $\bar{\mathcal{S}}$ given $\mathcal{S}$ along $\mathcal{F}^+$. □

LEMMA 7.8
Let $(\mathcal{S}, \bar{\mathcal{S}})$ be as above. Then

$$\mathcal{X} = \bar{\mathcal{S}} \cap \mathcal{S} \tag{7.46}$$

□

Proof The fact that $\mathcal{X} \subseteq \bar{\mathcal{S}} \cap \mathcal{S}$ is trivial. Let us show the other inclusion. Let $\bar{s} \in \bar{\mathcal{S}} \cap \mathcal{S}$; then $\bar{s} \in \bar{\mathcal{S}}$ and $\bar{s} \in \mathcal{S}$. Therefore

$$\bar{s} = E_{||\mathcal{F}^+}[\bar{s} \,|\, \mathcal{S}] = E_{||\mathcal{F}^+}[\bar{s} \,|\, \mathcal{X}] \in \mathcal{X}.$$

□

The following proposition is a generalization of the perpendicular intersection property known for "orthogonal" splitting subspaces.

PROPOSITION 7.11
Let $\mathcal{X}$ be an oblique Markovian splitting subspace and let $\mathcal{S}$, $\bar{\mathcal{S}}$ be the relative incoming-outgoing pair of subspaces. Then the following *oblique intersection property* holds:

$$\bar{\mathcal{S}} \perp \mathcal{S} \mid (\mathcal{X} + \mathcal{F}^+) \tag{7.47}$$

□

Proof Condition (7.55), is equivalent to (see Lemma 7.2)

$$E[\bar{\mathcal{S}} \mid \mathcal{S} + \mathcal{F}^+] = E[\bar{\mathcal{S}} \mid \mathcal{X} + \mathcal{F}^+]$$

which by (7.46) is precisely the oblique intersection property (7.47). □

The following theorem gives an "almost" one-to-one correspondence between oblique Markovian splitting subspaces and "scattering pairs".

THEOREM 7.14
Let $\mathcal{H}$ be a Hilbert space of random variables with shift operator σ and let $\mathcal{X}$ be a subspace of $\mathcal{H}$ such that

$$\mathcal{H} = \mathcal{Y} \vee \mathcal{U} \vee \left(\bigvee_t \mathcal{X}_t \right).$$

Then $\mathcal{X}$ is an oblique Markovian splitting subspace, if and only if

$$\mathcal{X} = \bar{\mathcal{S}} \cap \mathcal{S}$$

for some pair of subspaces $\mathcal{S}$, $\bar{\mathcal{S}}$ such that the following properties hold

1. Extended past and future property

$$\begin{cases} \mathcal{Y}^+ \subseteq \bar{\mathcal{S}} \\ \mathcal{P}^- \subseteq \mathcal{S} \end{cases} \quad , \quad \mathcal{S} \cap \mathcal{F}^+ = \{0\}$$

2. Shift-invariance

$$\begin{cases} \sigma \bar{\mathcal{S}} \subseteq \bar{\mathcal{S}} \\ \sigma^* \mathcal{S} \subseteq \mathcal{S} \end{cases}$$

3. Oblique intersection at $\mathcal{X}$

$$\bar{\mathcal{S}} \perp \mathcal{S} \mid ((\bar{\mathcal{S}} \cap \mathcal{S}) + \mathcal{F}^+)$$

Conversely, given an oblique Markovian splitting subspace $\mathcal{X}$, a pair of subspaces satisfying conditions 1), 2), 3), can be constructed as follows

$$\mathcal{S} = \mathcal{P}^- \vee \mathcal{X}^- \quad , \quad \mathcal{Y}^+ \vee \mathcal{X}^+ \subseteq \bar{\mathcal{S}} \subseteq \mathcal{Y}^+ \vee \mathcal{X}^+ \vee \mathcal{U}^+. \tag{7.48}$$

The minimal subspace $\bar{\mathcal{S}}$ satisfying 1), 2), and 3) (i.e. contained in any other $\bar{\mathcal{S}}$ satisfying 1), 2), and 3)), is given by

$$\bar{\mathcal{S}} = \mathcal{Y}^+ \vee \mathcal{X}^+$$

□

Proof Let $\mathcal{X}$ be an oblique Markovian splitting subspace, then $\mathcal{S} = \mathcal{P}^- \vee \mathcal{X}^-$ and $\bar{\mathcal{S}} = \mathcal{Y}^+ \vee \mathcal{X}^+$ satisfy the assumptions above and $\mathcal{X} = \bar{\mathcal{S}} \cap \mathcal{S}$. Conversely, let $\mathcal{S}$, $\bar{\mathcal{S}}$ be subspaces of $\mathcal{H}$ which satisfies the conditions above. Define the subspace $\mathcal{X} = \bar{\mathcal{S}} \cap \mathcal{S}$; then by assumptions 1) and 2) we have that $\mathcal{P}^- \vee \mathcal{X}^- \subseteq \mathcal{S}$ and $\mathcal{Y}^+ \vee \mathcal{X}^+ \subseteq \bar{\mathcal{S}}$, which by the oblique intersection property 3) implies that

$$\left(\mathcal{Y}^+ \vee \mathcal{X}^+\right) \perp \left(\mathcal{P}^- \vee \mathcal{X}^-\right) \mid \left(\mathcal{X} + \mathcal{F}^+\right). \tag{7.49}$$

By lemma 7.2 condition (7.49) is equivalent to the oblique Markovian splitting property (7.25).

Let us prove that $\mathcal{S}$ and $\bar{\mathcal{S}}$ are given by (7.48). We have already pointed out that $\mathcal{S}^m := \mathcal{P}^- \vee \mathcal{X}^- \subseteq \mathcal{S}$. Assume the inclusion is strict; then since $\mathcal{S} \subseteq \mathcal{H} = \mathcal{S}^m \vee \bar{\mathcal{S}} \vee \mathcal{U}^+$, $\boldsymbol{s} \in \mathcal{S}$ can be written as: $\boldsymbol{s} = \boldsymbol{s}^m + \bar{\boldsymbol{s}} + \boldsymbol{u}$ where $\boldsymbol{s}^m \in \mathcal{S}^m$, $\bar{\boldsymbol{s}} \in \bar{\mathcal{S}}$, $\boldsymbol{u} \in \mathcal{U}^+$. Therefore we have:

$$\boldsymbol{s} = E_{\| \mathcal{F}^+}[\boldsymbol{s}|\mathcal{S}] = \boldsymbol{s}^m + E_{\| \mathcal{F}^+}[\bar{\boldsymbol{s}}|\mathcal{S}] = \boldsymbol{s}^m + \boldsymbol{x}$$

where $\boldsymbol{x} \in \mathcal{X} \subseteq \mathcal{S}^m$ which contradict the hypothesis that $\boldsymbol{s} \notin \mathcal{S}^m$.
Similarly, we have seen that $\bar{\mathcal{S}}^m := \mathcal{Y}^+ \vee \mathcal{X}^+ \subseteq \bar{\mathcal{S}}$. Assume $(\mathcal{Y}^+ \vee \mathcal{X}^+ \vee \mathcal{U}^+) \subset \bar{\mathcal{S}}$ strictly. Then, since $\mathcal{Y}^+ \vee \mathcal{X}^+ \vee \mathcal{U}^+ \vee \mathcal{S}^m = \mathcal{H}$, there is $\bar{\boldsymbol{s}} \in \bar{\mathcal{S}}$, $\boldsymbol{s} \notin (\mathcal{Y}^+ \vee \mathcal{X}^+ \vee \mathcal{U}^+)$ which lies in $\mathcal{S}^m$. Therefore $\bar{\boldsymbol{s}} \in \mathcal{X}$, and hence $\bar{\boldsymbol{s}} \in \bar{\mathcal{S}}^m$ which contradicts the hypothesis. Requiring $\bar{\mathcal{S}}$ to be minimal implies that $\bar{\mathcal{S}}^m = \bar{\mathcal{S}}$ since $\bar{\mathcal{S}}^m \subseteq \bar{\mathcal{S}}$. □

The question of minimality of oblique Markovian splitting subspaces can be rephrased as a question of minimality for the subspaces $\mathcal{S}$ and $\bar{\mathcal{S}}$. In fact, given a Markovian splitting subspace $\mathcal{X}$ and the corresponding pair $(\mathcal{S}, \bar{\mathcal{S}})$, reducing $(\mathcal{S}, \bar{\mathcal{S}})$ without violating the properties 1), 2) and 3) of theorem 7.14 amounts to constructing an oblique Markovian splitting subspace which is contained in $\mathcal{X}$.

Scattering Pairs and Minimality

Our aim in this section is to adapt to oblique splitting subspaces a construction inspired by a similar procedure in stochastic realization theory, [18], which allows to construct a minimal Markovian splitting subspace starting from an arbitrarily "large" scattering pair $\left(\mathcal{S}, \bar{\mathcal{S}}\right)$ of perpendicularly intersecting subspaces. As it will be clear in a little while, we will only be able to draw a completely parallel construction in case of absence of feedback.

The construction of a minimal Markovian splitting subspace can be done in principle by reducing (in the sense of subspace inclusion) the subspaces $\left(\mathcal{S}, \bar{\mathcal{S}}\right)$ without violating properties 1), 2) and 3) of Theorem 7.14.

Before doing so we shall clarify the geometric meaning of constructibility and observability.

PROPOSITION 7.12
Let $\mathcal{X}$ be an oblique Markovian splitting subspace and let $\left(\mathcal{S}, \bar{\mathcal{S}}\right)$ be the scattering pair associated to it. Let us introduce the *extended scattering pair*, $\mathcal{S}_e := \mathcal{S} + \mathcal{F}^+$ and $\bar{\mathcal{S}}_e := \bar{\mathcal{S}} \vee \mathcal{F}^+$. Then $\mathcal{X}$ is observable if and only if

$$\bar{\mathcal{S}}_e = \mathcal{S}_e^{\perp} \vee \mathcal{Y}^+ \vee \mathcal{F}^+ \tag{7.50}$$

and constructible if and only if

$$\mathcal{S}_e = \bar{\mathcal{S}}_e^{\perp} \vee \mathcal{P}^- \vee \mathcal{F}^+. \tag{7.51}$$

□

Proof Assume (7.50) holds. Since by definition $\mathcal{S}_e^{\perp} \perp (\mathcal{S} + \mathcal{F}^+)$ then

$$\begin{aligned} \mathcal{X} &= E_{\parallel \mathcal{F}^+}\left[\bar{\mathcal{S}} \mid \mathcal{S}\right] \\ &= E_{\parallel \mathcal{F}^+}\left[\bar{\mathcal{S}}_e \mid \mathcal{X}\right] \\ &= E_{\parallel \mathcal{F}^+}\left[\mathcal{Y}^+ \mid \mathcal{X}\right] \end{aligned}$$

which is observability. Conversely, if $\mathcal{X}$ is observable

$$\mathcal{X} + \mathcal{F}^+ = E\left[\mathcal{Y}^+ \vee \mathcal{F}^+ \mid \mathcal{X} + \mathcal{F}^+\right]$$

which is in turn equivalent to

$$\left(\mathcal{Y}^+ \vee \mathcal{F}^+\right)^{\perp} \cap \left(\mathcal{X} + \mathcal{F}^+\right) = \{0\}$$

Taking orthogonal complements we can rewrite

$$\mathcal{Y}^+ \vee \mathcal{F}^+ \vee \left(\mathcal{X} + \mathcal{F}^+\right)^{\perp} = \mathcal{H};$$

since

$$\left(\mathcal{X} + \mathcal{F}^+\right) = \mathcal{S}_e \cap \bar{\mathcal{S}}_e$$

and $\mathcal{S}_e^{\perp} \subseteq \bar{\mathcal{S}}_e$ the following orthogonal decomposition holds

$$\mathcal{H} = \bar{\mathcal{S}}_e^{\perp} \oplus \left(\mathcal{S}_e^{\perp} \vee \mathcal{Y}^+ \vee \mathcal{F}^+\right)$$

from which the conclusion follows.
As far as constructibility is concerned, assume $\mathcal{X}$ is not constructible. There exist $x \in \mathcal{X}$ such that $E_{\parallel \mathcal{F}^+}[x \mid \mathcal{P}^-] = 0$, i.e. $x \in \mathcal{F}^+ \oplus (\mathcal{P}^- + \mathcal{F}^+)^{\perp}$ and therefore can be uniquely decomposed as $x = x_f \oplus \tilde{x}_f$ where $x_f \in \mathcal{F}^+$ and $\tilde{x}_f \in (\mathcal{P}^- + \mathcal{F}^+)^{\perp}$. Note that $\tilde{x}_f \neq 0$ since $\mathcal{X} \cap \mathcal{F}^+ = \{0\}$. It follows that $\tilde{x}_f \in \mathcal{S}_t^e \cap \bar{\mathcal{S}}_t^e$. This condition insures that $\tilde{x}_f \perp \left(\bar{\mathcal{S}}^e\right)^{\perp} \vee \mathcal{P}^- \vee \mathcal{F}^+$ and therefore $\tilde{x}_f \notin \left(\bar{\mathcal{S}}^e\right)^{\perp} \vee \mathcal{P}^- \vee \mathcal{F}^+$ which implies $\mathcal{S}^e \subset \left(\bar{\mathcal{S}}^e\right)^{\perp} \vee \mathcal{P}^- \vee \mathcal{F}^+$ strictly.
Conversely, assume $\mathcal{S}^e \subset \left(\bar{\mathcal{S}}^e\right)^{\perp} \vee \mathcal{P}^- \vee \mathcal{F}^+$ strictly. Then there exists $s \in \mathcal{S}^e$ and $s \in \left[\left(\bar{\mathcal{S}}^e\right)^{\perp} \vee \mathcal{P}^- \vee \mathcal{F}^+\right]^{\perp}$ or, alternatively, $s \in \bar{\mathcal{S}}^e \cap (\mathcal{P}^- + \mathcal{F}^+)^{\perp}$. Therefore $s \in \mathcal{S}_t^e \cap \mathcal{S}^e \cap (\mathcal{P}^- + \mathcal{F}^+)^{\perp} = (\mathcal{X} + \mathcal{F}^+) \cap (\mathcal{P}^- + \mathcal{F}^+)^{\perp}$. The last condition insures that $s = s_x + s_f$, $s_x \in \mathcal{X}$, $s_f \in \mathcal{F}^+$, and, for obvious reasons $s_x \neq 0$. Writing $s_x = s - s_f$ we have that $s_x \in \mathcal{X} \cap \mathcal{F}^+ \oplus (\mathcal{P}^- + \mathcal{F}^+)^{\perp}$ contradicting constructibility, which concludes the proof. □

We shall now introduce an orthogonal intersection property which is implied by the oblique intersection.

LEMMA 7.9
Let $(\mathcal{S}, \bar{\mathcal{S}})$ satisfy the oblique intersection property

$$\mathcal{S} \perp \bar{\mathcal{S}} \mid (\mathcal{X} + \mathcal{F}^+) .$$

Then the extended subspaces $\mathcal{S}_e = \mathcal{S} \vee \mathcal{F}^+$ and $\bar{\mathcal{S}}_e = \bar{\mathcal{S}} \vee \mathcal{F}^+$ intersect perpendicularly, i.e.

$$\mathcal{S}_e \perp \bar{\mathcal{S}}_e \mid (\mathcal{S}_e \cap \bar{\mathcal{S}}_e) .$$

□

Proof The proof follows readily from Theorem 2.1 in [18] □

Making use of this lemma we obtain immediately the following *orthogonal decomposition of the ambient space* $\mathcal{H}$ which is analogous to the one valid in stochastic realization for time series, and plays an important role in many structural questions in stochastic systems theory.

THEOREM 7.15
Let $\mathcal{X}$ be an oblique Markovian splitting subspace and $(\mathcal{S}_e, \bar{\mathcal{S}}_e)$ the associated extended scattering pair. Then the following orthogonal decomposition holds

$$\mathcal{H} = \mathcal{S}_e^{\perp} \oplus (\mathcal{X} + \mathcal{F}^+) \oplus \bar{\mathcal{S}}_e^{\perp} \tag{7.52}$$

□

The characterization of oblique Markovian splitting subspaces in terms of their scattering pair is a fundamental tool to study minimality. As we have seen an oblique Markovian splitting subspace can always be represented as the intersection of $\mathcal{S}$ and $\bar{\mathcal{S}}$. These subspaces, or more precisely their extended versions are related to observability and constructibility. Apparently, failing either of them, these subspace are not "minimal", in the sense of proposition 7.12. At this point, following classical stochastic realization theory, we would need a procedure to reduce, if possible, the subspaces $\mathcal{S}$ and $\bar{\mathcal{S}}$. Unfortunately such a procedure is not yet available and at this point the analogy with the classical theory seems to halt.

Nevertheless a proof of theorem 7.12 can still be given.

Proof of Theorem 7.12 According to theorem 7.14, if (7.50) and (7.51) hold, it is not possible to reduce these subspaces and therefore $\mathcal{X}$ must be minimal. Conversely, if $\mathcal{X}$ is minimal, it cannot be possible to reduce $\mathcal{S}$ and $\bar{\mathcal{S}}$ any further. This implies that (7.50) and (7.51) hold and therefore $\mathcal{X}$ is both observable and constructible. □

7.7 Stochastic Realization in the Absence of Feedback

When there is no feedback from from $\boldsymbol{y}$ to $\boldsymbol{u}$, some of the results presented above simplify considerably. For instance the construction of the oblique predictor space, somewhat complicated in the general setting, can be simplified when there is no feedback. Moreover, specializing some definitions, we shall also be able in this case to give a procedure to reduce the incoming and outgoing subspaces in order to achieve minimality. The following lemma will be useful in this respect.

LEMMA 7.10
Assume there is no feedback from $\boldsymbol{y}$ to $\boldsymbol{u}$. Let $\mathcal{E}$ be the innovation space of $\boldsymbol{y}$ defined by (7.41), then:

$$E_{||\mathcal{U}_t^+}\left[\mathcal{E}_t^+ \,|\, \mathcal{P}_t^-\right] = \{0\}$$

□

Proof Decomposition (7.17) can be rewritten in this causal case as:

$$\mathcal{P}_{t+1}^- = (\mathcal{P}_t^- + \mathcal{U}_t) \oplus \mathcal{E}_t$$

Since $\mathcal{E}_t = \text{span}\,\{e(t)\}$ it suffices to show that

$$E_{||\mathcal{U}_t^+}\left[e(t+k) \,|\, \mathcal{P}_t^-\right] = 0$$

for all $k \geq 0$. As we have already pointed out $e(t) \perp \mathcal{U}_{t+1}^+$ and obviously $e(t) \perp \mathcal{P}_t^- \vee \mathcal{U}_t$, therefore the oblique projection is zero.

This fact can be verified directly since by absence of feedback

$$\begin{aligned} E\left[\boldsymbol{y}(t) \mid \mathcal{P}_t^- \vee \mathcal{U}_t^+\right] &= E\left[\boldsymbol{y}(t) \mid (\mathcal{Y}_s)_t^- \oplus \left(\mathcal{U}_t^- \vee \mathcal{U}_t^+\right)\right] \\ &= E\left[\boldsymbol{y}(t) \mid (\mathcal{Y}_s)_t^-\right] \oplus E\left[\boldsymbol{y}(t) \mid \mathcal{U}\right] \\ &= \hat{\boldsymbol{y}}_s(t) \oplus E\left[\boldsymbol{y}(t) \mid \mathcal{U}_{t+1}^-\right] \\ &= E\left[\boldsymbol{y}(t) \mid \mathcal{P}_t^- \vee \mathcal{U}_t\right] \end{aligned}$$

which proves that $e(t)$ is orthogonal to $\mathcal{U}_{t+1}^+$ and hence $e(t) \perp \mathcal{U}$. □

We have seen that the wandering subspace generated by the innovation is orthogonal to the whole input history in the absence of feedback. Recall that the feedback-free property was defined from an input-output point of view, apparently putting no restrictions on the state $\boldsymbol{x}$. However it is straightforward to see that:

PROPOSITION 7.13
In the causal case, absence of feedback from $\boldsymbol{y}$ to $\boldsymbol{u}$ implies absence of feedback from $\boldsymbol{x}$ to $\boldsymbol{u}$. □

Proof Since $\mathcal{X}_t^- \subseteq \mathcal{P}_t^-$ and $\mathcal{P}_t^- \perp \mathcal{U}_t^+ \mid \mathcal{U}_t^-$, it is also true that

$$\mathcal{X}_t^- \perp \mathcal{U}_t^+ \mid \mathcal{U}_t^- \tag{7.53}$$

holds. □

Even in a non-causal situation, it is useful to restrict our attention to realizations whose state space satisfies the condition (7.53). Let us consider the subspace $\mathcal{Z}_t^- = \mathcal{Y}_t^- \vee \mathcal{X}_t^-$. Henceforth we shall only consider state spaces such that

$$\mathcal{Z}_t^- \perp \mathcal{U}_t^+ \mid \mathcal{U}_t^- \tag{7.54}$$

and we say that the corresponding realizations are *feedback free*. This extended notion of absence of feedback guarantees not only that the innovation $\boldsymbol{e}$ is orthogonal to future inputs, but that so will be any wandering subspace $\mathcal{W}_t$. The following proposition states this formally.

PROPOSITION 7.14
Let $\mathcal{X}$ be an oblique Markovian splitting subspace. The wandering subspace $\mathcal{W}$ which generates $\mathcal{X}$ is orthogonal to the whole input history if and only if the feedback-free condition (7.54) is satisfied. □

Proof (if) Recall that

$$\mathcal{S}_t = \mathcal{X}_t^- \vee \mathcal{Y}_t^- \vee \mathcal{U}_t^-,$$

and, by (7.54),

$$\mathcal{S}_t \perp \mathcal{U}_t^+ \mid \mathcal{U}_t^-.$$

From

$$\mathcal{S}_{t+1} = (\mathcal{S}_t + \mathcal{U}_t) \oplus \mathcal{W}_t$$

we have that $\mathcal{W}_t$ is contained in $\mathcal{S}_{t+1}$ and therefore $\mathcal{W}_t \perp \mathcal{U}_{t+1}^+ \mid \mathcal{U}_{t+1}^-$; since $\mathcal{W}_t \perp \mathcal{U}_{t+1}^-$ by construction, we obtain $\mathcal{W}_t \perp \mathcal{U}_{t+1}^+$ and hence

$$\mathcal{W}_t \perp \mathcal{U}$$

which is the thesis.

(only if) Assume $\mathcal{W}_t \perp \mathcal{U}$, since $\mathcal{S}_t = \mathcal{U}_t^- + \mathcal{W}_t^-$ it follows that

$$\mathcal{S}_t \perp \mathcal{U}_t^+ \mid \mathcal{U}_t^-.$$

Therefore, since $\mathcal{Z}_t^- = (\mathcal{X}_t^- \vee \mathcal{Y}_t^-) \subseteq \mathcal{S}_t$, the thesis follows

$$\mathcal{Z}_t^- \perp \mathcal{U}_t^+ \mid \mathcal{U}_t^-.$$

□

REMARK 7.16
Note that, under hypothesis (7.54), also the incoming subspace $\mathcal{S}_t$ satisfies $\mathcal{S}_t \perp \mathcal{U}_t^+ \mid \mathcal{U}_t^-$. It follows that the richness condition

$$\mathcal{U}_t^+ \cap \mathcal{S}_t = \{0\}$$

is automatically satisfied as long as the input is coercive (Assumption 7.12). □

The following result gives a somehow simpler geometric characterization of the oblique Markovian splitting property in the absence of feedback.

THEOREM 7.17
Let the symbols have the same meaning as above. Assume there is no feedback from $\boldsymbol{y}$ to $\boldsymbol{u}$ and that (7.54) holds. The subspace $\mathcal{X}$ is oblique Markovian splitting if and only if

$$E_{\|\mathcal{U}^+}[\bar{\mathcal{S}} \mid \mathcal{S}] = E_{\|\mathcal{U}^+}[\bar{\mathcal{S}} \mid \mathcal{X}]. \tag{7.55}$$

□

Proof The condition is obviously sufficient since $\mathcal{Y}_t \vee \mathcal{X}_{t+1} \subseteq \bar{\mathcal{S}}_t$ and, by absence of feedback (7.54):

$$E[\mathcal{Y}_t \vee \mathcal{X}_{t+1} \,|\, \mathcal{S}_t + \mathcal{U}_t^+] = E[\mathcal{Y}_t \vee \mathcal{X}_{t+1} \,|\, \mathcal{S}_t + \mathcal{U}_t]$$

which by lemma (7.1) implies that

$$E_{||\mathcal{U}_t^+}[\mathcal{Y}_t \vee \mathcal{X}_{t+1} \,|\, \mathcal{S}_t] = E_{||\mathcal{U}_t}[\mathcal{Y}_t \vee \mathcal{X}_{t+1} \,|\, \mathcal{S}_t]$$

and therefore

$$E_{||\mathcal{U}_t}[\mathcal{Y}_t \vee \mathcal{X}_{t+1} \,|\, \mathcal{S}_t] \subseteq \mathcal{X}_t.$$

To prove the other implication just note that by Theorem 7.7

$$\mathcal{Y}_{t+k} \subseteq (\mathcal{X}_t + \mathcal{U}_{[t,t+k)}) \oplus \mathcal{W}_{[t,t+k)}$$

and

$$\mathcal{X}_{t+k+1} \subseteq (\mathcal{X}_t + \mathcal{U}_{[t,t+k)}) \oplus \mathcal{W}_{[t,t+k)}$$

where the last sum is orthogonal from Proposition 7.14. It follows that for every $k \geq 0$

$$E_{\|\mathcal{U}_t^+}[\mathcal{Y}_{t+k} \vee \mathcal{X}_{t+k+1} \,|\, \mathcal{S}_t] \subseteq \mathcal{X}_t$$

which is equivalent to (7.55). □

The following lemma is the equivalent of Lemma 7.7

LEMMA 7.11
Let $(\mathcal{S}, \bar{\mathcal{S}})$ and $\mathcal{U}^+$ be as defined above. Then

$$\mathcal{X} = E_{||\mathcal{U}^+}[\bar{\mathcal{S}} \,|\, \mathcal{S}]$$

Hence every oblique MArkovian splitting subspace $\mathcal{X}$ is the oblique predictor space for $\bar{\mathcal{S}}$, given $\mathcal{S}$, along $\mathcal{U}^+$. □

Proof In the feedback free case, from Proposition 7.14we get $\mathcal{W}^+ \perp (\mathcal{S} + \mathcal{U}^+)$. Therefore, since $\mathcal{F}^+ = \mathcal{U}^+ \vee \mathcal{W}^+$,

$$E_{||\mathcal{U}^+}[\mathcal{X} \,|\, \mathcal{S}] = E_{||\mathcal{F}^+}[\mathcal{X} \,|\, \mathcal{S}] = \mathcal{X}$$

where the last equality follows from Lemma 7.7. □

The following proposition specializes the concept of oblique intersection to the case when there is no feedback.

PROPOSITION 7.15
Assume there is no feedback from $\boldsymbol{y}$ to $\boldsymbol{u}$. Let $\mathcal{X}$ be an oblique Markovian splitting subspace and let $\mathcal{S}$, $\bar{\mathcal{S}}$ be the incoming and outgoing subspaces attached to it. Then the following holds:

$$\bar{\mathcal{S}} \perp \mathcal{S} \mid ((\bar{\mathcal{S}} \cap \mathcal{S}) + \mathcal{U}^+) \tag{7.56}$$

□

Proof Condition (7.55), is equivalent to (see Lemma 7.2)

$$E[\bar{\mathcal{S}} \mid \mathcal{S} + \mathcal{U}^+] = E[\bar{\mathcal{S}} \mid \mathcal{X} + \mathcal{U}^+]$$

which by (7.46) is precisely the oblique intersection property (7.56). □

REMARK 7.18
It is worth to stress, at this point, that condition (3) (oblique intersection) of Theorem 7.14 can be replaced by condition (7.56). □

The following theorem gives a characterization of the oblique predictor space in absence of feedback.

THEOREM 7.19
In absence of feedback the oblique predictor space can be computed by the formula

$$\mathcal{X}^{+/-} := E_{\|\mathcal{U}^+}[\mathcal{Y}^+ \mid \mathcal{P}^-]. \tag{7.57}$$

□

Proof To show this let us just note that by (7.54) and by lemma 7.1

$$E_{\|\mathcal{U}_{t+h}}[\mathcal{Y}_{t+h} \mid \mathcal{P}^-_{t+h}] = E_{\|\mathcal{U}^+_{t+h}}[\mathcal{Y}_{t+h} \mid \mathcal{P}^-_{t+h}] \subseteq \mathcal{X}_{t+h}$$

for any causal oblique markovian splitting subspace $\mathcal{X}_{t+h}$ and

$$E_{\|\mathcal{U}_{t+h}}[\mathcal{X}_{t+h+1} \mid \mathcal{P}^-_{t+h}] = E_{\|\mathcal{U}^+_{t+h}}[\mathcal{X}_{t+h+1} \mid \mathcal{P}^-_{t+h}]$$

in the absence of feedback; this implies the following:

$$\begin{aligned}
\left(\mathcal{X}^{+/-}_t\right)^k &= E_{\|\mathcal{U}_t}\left[E_{\|\mathcal{U}_{t+1}}\left[\cdots E_{\|\mathcal{U}_{t+k}}\left[\mathcal{Y}_{t+k} \mid \mathcal{P}^-_{t+k}\right] \cdots \mid \mathcal{P}^-_{t+1}\right] \mid \mathcal{P}^-_t\right] \\
&= E_{\|\mathcal{U}^+_t}\left[E_{\|\mathcal{U}^+_{t+1}}\left[\cdots E_{\|\mathcal{U}^+_{t+k}}\left[\mathcal{Y}_{t+k} \mid \mathcal{P}^-_{t+k}\right] \cdots \mid \mathcal{P}^-_{t+1}\right] \mid \mathcal{P}^-_t\right] \\
&= E_{\|\mathcal{U}^+_t}[\mathcal{Y}_{t+k} \mid \mathcal{P}^-_t]
\end{aligned}$$

and hence

$$\begin{aligned}
\mathcal{X}^{+/-}_t &= \bigvee_{k\geq 0} E_{\|\mathcal{U}^+_t}[\mathcal{Y}_{t+k} \mid \mathcal{P}^-_t] \\
&= E_{\|\mathcal{U}^+_t}\left[\mathcal{Y}^+_t \mid \mathcal{P}^-_t\right]
\end{aligned} \tag{7.58}$$

□

Scattering Pairs and Minimality (without Feedback)

In this section we shall give a procedure to reduce the state space when it is not minimal, by reducing the incoming and outgoing subspaces using a two-steps procedure similar to that described in [18].

The construction of a minimal Markovian splitting subspace can be done by reducing (in the sense of subspace inclusion) the subspaces $(\mathcal{S}, \bar{\mathcal{S}})$ without violating properties (1), (2) and (3) (which in this case is equivalent to (7.56)) of Theorem 7.14. We shall first state some technical results which will be needed throughout the section.

First of all, we shall introduce an orthogonal intersection property which is implied by the oblique intersection.

LEMMA 7.12
Let $(\mathcal{S}, \bar{\mathcal{S}})$ satisfy the oblique intersection property

$$\mathcal{S} \perp \bar{\mathcal{S}} \mid ((\mathcal{S} \cap \bar{\mathcal{S}}) + \mathcal{U}^+) .$$

Then the extended subspaces $\mathcal{S}_{eu} := \mathcal{S} \vee \mathcal{U}^+$ and $\bar{\mathcal{S}}_{eu} := \bar{\mathcal{S}} \vee \mathcal{U}^+$ intersect perpendicularly, i.e.

$$\mathcal{S}_{eu} \perp \bar{\mathcal{S}}_{eu} \mid (\mathcal{S}_{eu} \cap \bar{\mathcal{S}}_{eu}) .$$

□

Proof The proof follows readily from Theorem 2.1 in [18] □

Using this lemma we obtain the orthogonal decomposition of the ambient space $\mathcal{H}$ valid in the feedback-free case

$$\mathcal{H} = \mathcal{S}_{eu}^{\perp} \oplus (\mathcal{X} + \mathcal{U}^+) \oplus \bar{\mathcal{S}}_{eu}^{\perp} \tag{7.59}$$

Note that $\mathcal{Y}^+ \vee \mathcal{U}^+ \subset \bar{\mathcal{S}}_{eu}$ and $\mathcal{P}^- \vee \mathcal{U}^+ \subset \mathcal{S}_{eu}$ must hold for every pair of subspaces $(\mathcal{S}, \bar{\mathcal{S}})$ attached to an oblique Markovian splitting subspace.

We shall construct a pair of subspaces $(\mathcal{S}^1, \bar{\mathcal{S}}^1)$, which are contained in $(\mathcal{S}, \bar{\mathcal{S}})$ and which satisfies all the conditions of Theorem 7.14 by defining their "extended" version and then by properly reducing them.

Define

$$\bar{\mathcal{S}}_{eu}^1 := \mathcal{S}_{eu}^{\perp} \vee \mathcal{Y}^+ \vee \mathcal{U}^+ \tag{7.60a}$$

$$\mathcal{S}_{eu}^1 := (\bar{\mathcal{S}}_{eu}^1)^{\perp} \vee \mathcal{P}^- \vee \mathcal{U}^+ \tag{7.60b}$$

and also the related state space

$$\mathcal{X}_{eu}^1 := \bar{\mathcal{S}}_{eu}^1 \cap \mathcal{S}_{eu}^1 \tag{7.61}$$

LEMMA 7.13
The pair of subspaces defined in (7.60) intersect perpendicularly, i.e.

$$\bar{\mathcal{S}}_{eu}^1 \perp \mathcal{S}_{eu}^1 \mid \bar{\mathcal{S}}_{eu}^1 \cap \mathcal{S}_{eu}^1. \tag{7.62}$$

□

Proof Clearly $\mathcal{H} = \bar{\mathcal{S}}_{eu}^1 \vee \mathcal{S}_{eu}^1$ and $(\bar{\mathcal{S}}_{eu}^1)^{\perp} \subset \mathcal{S}_{eu}^1$, which together implies that $\bar{\mathcal{S}}_{eu}^1$ and $\mathcal{S}_{eu}^1$ intersect perpendicularly. □

Of course from (7.60) we can see that $\bar{\mathcal{S}}_{eu}^1 \subseteq \bar{\mathcal{S}}_{eu}$ and from the fact that $(\bar{\mathcal{S}}_{eu}^1)^{\perp} = \mathcal{S}_{eu} \cap (\mathcal{Y}^+)^{\perp} \cap (\mathcal{U}^+)^{\perp}$ also $\mathcal{S}_{eu}^1 \subseteq \mathcal{S}_{eu}$, which implies that

$$\mathcal{U}^+ \subseteq \bar{\mathcal{S}}_{eu}^1 \cap \mathcal{S}_{eu}^1 \subseteq \mathcal{X} + \mathcal{U}^+.$$

Let us define the subspace $\mathcal{X}^1$ as follows:

$$\mathcal{X}^1 := \mathcal{X} \cap (\bar{\mathcal{S}}_{eu}^1 \cap \mathcal{S}_{eu}^1) = \mathcal{X} \cap \mathcal{X}_{eu}^1. \tag{7.63}$$

We shall show that $\mathcal{X}^1$ is a minimal Markovian splitting subspace (clearly contained in $\mathcal{X}$). Before doing so we state the following technical lemma which will be used in the following:

LEMMA 7.14
There holds

$$\mathcal{X}^1_{eu} = \mathcal{X}^1 + \mathcal{U}^+ \tag{7.64}$$

□

Proof Since $\mathcal{X}^1 \subset \mathcal{X}^1_{eu}$ and $\mathcal{U}^+ \subset \mathcal{X}^1_{eu}$, clearly $\mathcal{X}^1 + \mathcal{U}^+ \subset \mathcal{X}^1_{eu}$. Conversely, since $\mathcal{X}^1_{eu} \subset \mathcal{X} + \mathcal{U}^+$, any $\xi \in \mathcal{X}^1_{eu}$ can be written as $\xi = \boldsymbol{x} + \boldsymbol{u}^+$ with $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{u}^+ \in \mathcal{U}^+$. However since $\mathcal{X}^1_{eu} \supset \mathcal{U}^+$, then $\boldsymbol{u}^+ \in \mathcal{X}^1_{eu}$ as well, and hence $\boldsymbol{x}$ belongs to both $\mathcal{X}^1_{eu}$ and $\mathcal{X}$, so that $\boldsymbol{x} \in \mathcal{X}^1$. □

THEOREM 7.20
The pair of subspaces $(\mathcal{S}^1, \bar{\mathcal{S}}^1)$ defined as

$$\begin{aligned} \mathcal{S}_1 &:= \mathcal{S} \cap \mathcal{S}^1_{eu} \\ \bar{\mathcal{S}}_1 &:= \bar{\mathcal{S}} \cap \bar{\mathcal{S}}^1_{eu} \end{aligned} \tag{7.65}$$

satisfy the conditions of theorem 7.14 and $\mathcal{X}^1 = \mathcal{S}^1 \cap \bar{\mathcal{S}}^1$. Therefore $\mathcal{X}^1$ is oblique Markovian splitting subspace and is contained in $\mathcal{X}$. □

Proof First of all let us just note that since $\mathcal{X}^1 \subseteq \mathcal{X}$ then $\mathcal{X}^1 \subseteq \mathcal{S}$ and $\mathcal{X}^1 \subseteq \bar{\mathcal{S}}$, which implies that $\mathcal{X}^1 \subseteq \mathcal{S}^1$ and $\mathcal{X}^1 \subseteq \bar{\mathcal{S}}^1$. Moreover $\mathcal{P}^- \subseteq \mathcal{S}^1$, $\mathcal{Y}^+ \subseteq \bar{\mathcal{S}}^1$ and $\mathcal{U}^+ \cap \mathcal{S}^1 = \{0\}$. Clearly $\mathcal{X}^1 \subseteq \mathcal{S}^1 \cap \bar{\mathcal{S}}^1$, let us show the converse. Since, from Lemma 7.14,

$$\mathcal{S}^1_{eu} \cap \bar{\mathcal{S}}^1_{eu} = \mathcal{X}^1 + \mathcal{U}^+ \supseteq (\mathcal{S}^1 \cap \bar{\mathcal{S}}^1) + \mathcal{U}^+,$$

by the direct sum property we have that

$$\mathcal{S}^1 \cap \bar{\mathcal{S}}^1 \subseteq \mathcal{X}^1$$

and therefore

$$\mathcal{X}^1 = \mathcal{S}^1 \cap \bar{\mathcal{S}}^1 \tag{7.66}$$

Proving the shift invariance properties

$$\begin{cases} \sigma \bar{\mathcal{S}} \subseteq \bar{\mathcal{S}} \\ \sigma^* \mathcal{S} \subseteq \mathcal{S} \end{cases}$$

is just an easy check. The oblique intersection property is immediate from (7.62) and (7.66). □

We have seen a construction which allows us to reduce an oblique Markovian splitting subspace. We shall now prove that indeed this procedure yields a minimal one.

THEOREM 7.21
The subspace $\mathcal{X}^1$ defined above is a minimal oblique Markovian splitting subspace. □

Proof The proof follows the same lines as in the stochastic case. Namely we assume that there exists an oblique Markovian splitting subspace, say $\mathcal{X}^o$, properly contained in $\mathcal{X}^1$. If such a subspace exists, then we could attach to it a pair of subspaces $(\mathcal{S}^o, \bar{\mathcal{S}}^o)$, which obviously satisfy $\mathcal{S}^o \subseteq \mathcal{S}^1$ and $\bar{\mathcal{S}}^o \subseteq \bar{\mathcal{S}}^1$. Then by the same argument we have already used we get that $\left(\bar{\mathcal{S}}^o_{eu}\right)^\perp \subseteq \mathcal{S}^o_{eu}$ and therefore $\mathcal{S}^o_{eu} \supseteq \left(\bar{\mathcal{S}}^o_{eu}\right)^\perp \vee \bar{\mathcal{P}}^- \vee \mathcal{U}^+$. But from the very definition of $\mathcal{S}^1_{eu}$ we have

$$\mathcal{S}^1_{eu} = \left(\bar{\mathcal{S}}^1_{eu}\right)^\perp \vee \bar{\mathcal{P}}^- \vee \mathcal{U}^+ \subseteq \mathcal{S}^o_{eu}$$

which implies that $\mathcal{S}^1_{eu} = \mathcal{S}^o_{eu}$. Moreover since $\mathcal{S}^1_{eu} \subseteq \mathcal{S}_{eu}$, $\bar{\mathcal{S}}^1_{eu} = \mathcal{S}^\perp_{eu} \vee \mathcal{Y}^+ \vee \mathcal{U}^+ \subseteq \bar{\mathcal{S}}^o_{eu}$ and therefore $\bar{\mathcal{S}}^1_{eu} = \bar{\mathcal{S}}^o_{eu}$ which guarantees that $\mathcal{S}^1 = \mathcal{S}^o$ and $\bar{\mathcal{S}}^1 = \bar{\mathcal{S}}^o$ and therefore $\mathcal{X}^o = \mathcal{X}^1$. □

Splitting property and Hankel Operators in the Absence of Feedback

In this section we shall study observability and constructibility of an oblique Markovian splitting subspace in the absence of feedback. The state space property of $\mathcal{X}$ will be interpreted in terms of factorization of a certain Hankel operator defined on the data space.

We first state, without proof, a lemma which provides a representation of the observability and constructibility operators in the absence of feedback.

LEMMA 7.15
Assume there is non feedback from $\boldsymbol{y}$ to $\boldsymbol{u}$. Then

$$\mathbb{O}^*\lambda = E_{\|\mathcal{U}^+}[\lambda \mid \mathcal{X}], \quad \lambda \in \mathcal{Y}^+ \tag{7.67}$$

and

$$\mathbb{K}\xi = E[\xi \mid \mathcal{P}^-], \quad \xi \in \mathcal{X}. \tag{7.68}$$

□

Let us consider also the *Hankel operator* $\mathbb{H} : \mathcal{Y}^+ \to \mathcal{P}^-$ defined as

$$\mathbb{H}\lambda := E_{\|\mathcal{U}^+}[\lambda \mid \mathcal{P}^-] \quad , \quad \lambda \in \mathcal{Y}^+.$$

PROPOSITION 7.16
The splitting property of $\mathcal{X}$ is equivalent to the factorization $\mathbb{H} = \mathbb{K}\mathbb{O}^*$. □

Proof For every $\lambda \in \mathcal{Y}^+$,

$$E\left[\lambda \mid \mathcal{P}^- + \mathcal{U}^+\right] = E\left[E\left[\lambda \mid (\mathcal{P}^- + \mathcal{X}) + \mathcal{U}^+\right] \mid \mathcal{P}^- + \mathcal{U}^+\right]$$

and by the splitting property,

$$E\left[\lambda \mid \mathcal{P}^- + \mathcal{U}^+\right] = E\left[E\left[\lambda \mid \mathcal{X} + \mathcal{U}^+\right] \mid \mathcal{P}^- + \mathcal{U}^+\right]$$

which implies that

$$E_{\|\mathcal{U}^+}[\lambda \mid \mathcal{P}^-] = E_{\|\mathcal{U}^+}\left[E\left[\lambda \mid \mathcal{X} + \mathcal{U}^+\right] \mid \mathcal{P}^-\right]$$

and since $\mathcal{X} \cap \mathcal{U}^+ = 0$ in the feedback-free situation, we have

$$E\left[\lambda \mid \mathcal{X} + \mathcal{U}^+\right] = E_{\| \mathcal{U}^+}\left[\lambda \mid \mathcal{X}\right] + E_{\| \mathcal{X}}\left[\lambda \mid \mathcal{U}^+\right]$$

which implies that

$$E_{\| \mathcal{U}^+}\left[\lambda \mid \mathcal{P}^-\right] = E\left[E_{\| \mathcal{U}^+}\left[\lambda \mid \mathcal{X}\right] \mid \mathcal{P}^-\right]$$

This is the factorization $\mathbb{H} = \mathbb{K}\mathbb{O}^*$. □

As usual we say that this factorization is *canonical* if $\mathbb{O}^*$ has dense range and if $\mathbb{K}$ is injective. Using the well known relations, valid for every bounded linear operator [2]

$$\mathcal{X} = \overline{\text{Range } \mathbb{O}^*} \oplus \text{Ker } \mathbb{O} \qquad \mathcal{X} = \overline{\text{Range } \mathbb{K}^*} \oplus \text{Ker } \mathbb{K}$$

we see that

$$\overline{\text{Range } \mathbb{O}^*} = E_{\| \mathcal{U}^+}\left[\mathcal{Y}^+ \mid \mathcal{X}\right] \quad , \quad \text{Ker } \mathbb{O} = \mathcal{X} \cap \left(E_{\| \mathcal{U}^+}\left[\mathcal{Y}^+ \mid \mathcal{X}\right]\right)^{\perp}$$

and

$$\overline{\text{Range } \mathbb{K}^*} = E\left[\mathcal{P}^- \mid \mathcal{X}\right] \quad , \quad \text{Ker } \mathbb{K} = \mathcal{X} \cap (\mathcal{P}^-)^{\perp}$$

Therefore we shall call $\overline{\text{Range } \mathbb{O}^*}$ the *observable component of* $\mathcal{X}$ and its orthogonal complement Ker $\mathbb{O}$ *the unobservable subspace*. Similarly, $\overline{\text{Range } \mathbb{K}^*}$ is the *constructible component* of the state and Ker $\mathbb{K} = \mathcal{X} \cap (\mathcal{P}^-)^{\perp}$ is the *unconstructible* subspace.

As we can see, in stochastic realization with inputs one is led to consider a "mixture" of the concepts of constructibility and reachability[3]. Let us look at the expression for the unconstructible component. Since by the feedback free property $\mathcal{P}^- = \mathcal{Y}_s^- \oplus \mathcal{U}^-$ we obtain:

$$\text{Ker } \mathbb{K} = \left[\mathcal{X} \cap (\mathcal{Y}_s^-)^{\perp}\right] \cap \left[\mathcal{X} \cap (\mathcal{U}^-)^{\perp}\right]. \tag{7.69}$$

We anticipate here that this concept of constructibility, which, as we shall see later, is the one which is linked to minimality (in the sense of subspace inclusion) does not in general imply constructibility of the stochastic component and hence the condition is not strong enough to characterize stochastic minimality.

In order to get a deeper understanding of the situation, let us define the restricted operators

$$\mathbb{K}_r^* := \mathbb{K}^*_{|\mathcal{Y}_s^-} \tag{7.70}$$

and

$$\mathbb{R}^* := \mathbb{K}^*_{|\mathcal{U}^-} \tag{7.71}$$

[2]Note that the norm of the oblique projection satisfies $\|\mathbb{O}^*\|^2 = (1 - \sigma 2_{max})^{-1}$ where σ_{max} is the maximum canonical correlation coefficient between $\mathcal{U}^+$ and $\mathcal{X}$, which is strictly less than 1 by the zero intersection property $\mathcal{X} \cap \mathcal{U}^+ = \{0\}$.

[3]This is the reason why in the beginning of the section we have used the word "constructibility" between quotes.

the former being related to the "stochastic" component and the latter to the "deterministic" component.

Let us define

$$\mathcal{X}_d := E[\mathcal{X} \mid \mathcal{U}^-] = \mathbb{R}\mathcal{X}, \quad , \quad \mathcal{X}_s := E\left[\mathcal{X} \mid \mathcal{W}^-\right] = \mathbb{R}_w\mathcal{X}, \tag{7.72}$$

where $\mathbb{R}_w$ is the "stochastic" reachability operator

$$\begin{array}{rccl} \mathbb{R}_w & : & \mathcal{X} & \to & \mathcal{W}^- \\ & & \boldsymbol{x} & \to & E\left[\boldsymbol{x} \mid \mathcal{W}^-\right] \end{array}$$

Recall that $\mathcal{S} = \mathcal{U}^- \oplus \mathcal{W}^-$, which implies that $\mathcal{X}_d \perp \mathcal{X}_s$. Let us also note that $\mathcal{X} \subseteq \mathcal{X}_s \oplus \mathcal{X}_d$. From

$$E[\boldsymbol{x} \mid \mathcal{Y}_s^-] = E\left[E\left[\boldsymbol{x} \mid \mathcal{W}^-\right] \mid \mathcal{Y}_s^-\right] \quad , \quad \forall \boldsymbol{x} \in \mathcal{X}$$

the restricted constructibility operator $\mathbb{K}_r$ can be factorized as follows:

$$\mathbb{K}_r = \mathbb{K}_s\mathbb{R}_w \tag{7.73}$$

where $\mathbb{K}_s$ is the usual "stochastic" constructibility operator.

In general neither $\mathbb{K}_r^*$ nor $\mathbb{R}^*$ will have dense range, or equivalently, neither $\mathbb{K}_r$ nor $\mathbb{R}$ will have a trivial kernel. The meaning of the constructibility condition for the state space, is that the intersection of the two kernels must be the zero random variable.

On the other hand, since all processes involved are assumed to be p.n.d., the joint system is reachable, (recall that $\mathcal{X} \subset \mathcal{S}$),[4] and therefore Ker $\mathbb{R}_w \cap$ Ker $\mathbb{R} = \{0\}$.

The following fact clarifies the link between constructibility of the joint model and constructibility of the "stochastic" component.

PROPOSITION 7.17
If the "stochastic component" is constructible, i.e. Ker $\mathbb{K}_s = \{0\}$, then the joint model is so. Geometrically this condition reads as

$$\mathcal{X}_s \cap (\mathcal{Y}_s^-)^\perp = \{0\}.$$

□

Proof This is immediate since Ker $\mathbb{K}$ = Ker $\mathbb{K}_r \cap$ Ker $\mathbb{R}$. Therefore, if $\boldsymbol{x} \in$ Ker $\mathbb{K}$ then, $\boldsymbol{x} \in$ Ker $\mathbb{R}$, $\boldsymbol{x} \notin$ Ker $\mathbb{R}_w$ and hence $\mathbb{R}_w\boldsymbol{x} \in$ Ker $\mathbb{K}_s$ form (7.73). □

[4]Here we assume that there are no p.n.d. components, see [17] for a discussion on this topic.

REMARK 7.22
Note that in general Ker $\mathbb{K} = \{0\}$ does not imply Ker $\mathbb{K}_s = \{0\}$. In fact, assume $\boldsymbol{x} = \boldsymbol{x}_s + \boldsymbol{x}_d$. It might well happen that $\mathbb{K}_r \boldsymbol{x} = \mathbb{K}_s \mathbb{R}_w \boldsymbol{x} = \mathbb{K}_s \mathbb{R}_w \boldsymbol{x}_s = 0$ while $\mathbb{R}\boldsymbol{x} = \mathbb{R}\boldsymbol{x}_d \neq 0$. Geometrically

$$\mathcal{X} \cap (\mathcal{P}^-)^\perp = \{0\}$$

does not imply

$$\mathcal{X}_s \cap (\mathcal{Y}_s^-)^\perp = \{0\}.$$

In fact we might have $\boldsymbol{x}_s \in (\mathcal{Y}_s^-)^\perp$, without $\boldsymbol{x} \in (\mathcal{P}^-)^\perp$. □

In some sense this shows that the definition we have given of "minimality" is not quite complete. Since however the concept of minimality is historically linked to "dimension", or more generally, to inclusion in the infinite dimensional case, we shall introduce a further definition.

DEFINITION 7.23
An oblique Markovian splitting subspace $\mathcal{X}$ is *strongly minimal* if it is minimal and $\mathcal{X}_s$ defined in (7.72) is constructible, i.e. Ker $\mathbb{K}_s = \{0\}$, or, equivalently,

$$\mathcal{X}_s \cap (\mathcal{Y}_s^-)^\perp = \{0\}$$

□

It is apparent that minimality and strong minimality are equivalent in the causal case, as the following proposition states.

PROPOSITION 7.18
Let $\mathcal{X}$ be a minimal causal oblique Markovian splitting subspace, then it is strongly minimal. □

Proof In the causal case $\mathcal{W}^-$ is the space spanned by the past innovation $\mathcal{E}^-$, which is nothing but $\mathcal{Y}_s^-$; therefore $\mathcal{X}_s \cap (\mathcal{Y}_s^-)^\perp = \{0\}$. □

Using decomposition (7.72) one can define the observability operators

$$\mathbb{O}_s^* := E_{\| \mathcal{U}^+} \left[\mathcal{Y}^+ \mid \mathcal{X}_s \right] = E \left[\mathcal{Y}^+ \mid \mathcal{X}_s \right] = E \left[\mathcal{Y}_s^+ \mid \mathcal{X}_s \right] \tag{7.74}$$

and

$$\mathbb{O}_d^* := E_{\| \mathcal{U}^+} \left[\mathcal{Y}^+ \mid \mathcal{X}_d \right] = E_{\| \mathcal{U}^+} \left[\mathcal{Y}_d^+ \mid \mathcal{X}_d \right]. \tag{7.75}$$

It is easy to see that the following factorizations hold

$$\mathbb{O}_s^* = \mathbb{R}_w \mathbb{O}^*$$

and

$$\mathbb{O}_d^* = \mathbb{R}\mathbb{O}^*.$$

Note that the observability conditions for the "deterministic" and "stochastic" component, i.e. Ker $\mathbb{O}_d = \{0\}$ and Ker $\mathbb{O}_s = \{0\}$ do not in general imply that Ker $\mathbb{O} = \{0\}$, while the converse is always true since $(\mathbb{R}_w)_{|\mathcal{X}_s}$ and $\mathbb{R}_{|\mathcal{X}_d}$ have trivial kernel by construction.

The geometric characterizations of minimality given above does not address the question of strong minimality. This has clearly to do only with the "stochastic" component and therefore it can be expressed geometrically in the usual way as

$$\mathcal{Y}_s^- \vee \mathcal{X}_s^- = (\mathcal{Y}_s^+ \vee \mathcal{X}_s^+)^\perp \vee \mathcal{Y}_s^-.$$

where orthogonal complement is taken in $H(\boldsymbol{w})$.

THEOREM 7.24
Let $\mathcal{X}$ be an oblique Markovian splitting subspace and let $\mathcal{S}_{eu} = \mathcal{S} + \mathcal{U}^+$, $\bar{\mathcal{S}}_{eu} = \bar{\mathcal{S}} \vee \mathcal{U}^+$. The following conditions are equivalent:

i) $\mathcal{X}$ is strongly minimal

ii) $\mathcal{X}$ is minimal and $\mathbb{K}_s$ is injective

iii) $\bar{\mathcal{S}}_{eu} = \mathcal{S}_{eu}^\perp \vee \mathcal{Y}^+ \vee \mathcal{U}^+$ and $\mathcal{Y}_s^- \vee \mathcal{X}_s^- = [H(\boldsymbol{w}) \ominus (\mathcal{Y}_s^+ \vee \mathcal{X}_s^+)] \vee \mathcal{Y}_s^-$

iv) $\bar{\mathcal{S}}_{eu} = \mathcal{S}_{eu}^\perp \vee \mathcal{Y}^+ \vee \mathcal{U}^+$ and $\mathcal{S}_{eu} = [H(\boldsymbol{w}) \ominus (\mathcal{Y}_s^+ \vee \mathcal{X}_s^+)] \vee \mathcal{P}^- \vee \mathcal{U}^+$

□

Proof *i*) and *ii*) are equivalent by definition. The fact that $\mathcal{X}$ is minimal implies that it is observable, i.e. $\bar{\mathcal{S}}_{eu} = \mathcal{S}_{eu}^\perp \vee \mathcal{Y}^+ \vee \mathcal{U}^+$, and $\mathbb{K}_s$ injective is equivalent to $\mathcal{Y}_s^- \vee \mathcal{X}_s^- = [H(\boldsymbol{w}) \ominus (\mathcal{Y}_s^+ \vee \mathcal{X}_s^+)] \vee \mathcal{Y}_s^-$, from which condition *iii*). To show that *iii*) implies *iv*) just note that since $\mathcal{Y}_s^- \vee \mathcal{X}_s^- \vee \mathcal{U} = \mathcal{S}_{eu}$, $\mathcal{Y}_s^- \vee \mathcal{X}_s^- = [H(\boldsymbol{w}) \ominus (\mathcal{Y}_s^+ \vee \mathcal{X}_s^+)] \vee \mathcal{Y}_s^-$ implies that $\mathcal{S}_{eu} = [H(\boldsymbol{w}) \ominus (\mathcal{Y}_s^+ \vee \mathcal{X}_s^+)] \vee \mathcal{Y}_s^- \vee \mathcal{U} = [H(\boldsymbol{w}) \ominus (\mathcal{Y}_s^+ \vee \mathcal{X}_s^+)] \vee \mathcal{P}^- \vee \mathcal{U}^+$. Conversely, if *iv*) holds, $\mathcal{S}_{eu} \ominus \mathcal{U} = [H(\boldsymbol{w}) \ominus (\mathcal{Y}_s^+ \vee \mathcal{X}_s^+)] \vee \mathcal{P}^- \vee \mathcal{U}^+ \ominus \mathcal{U}$, i.e. $\mathcal{Y}_s^- \vee \mathcal{X}_s^- = [H(\boldsymbol{w}) \ominus (\mathcal{Y}_s^+ \vee \mathcal{X}_s^+)] \vee \mathcal{Y}_s^-$, which concludes the proof. □

7.8 Reconciliation with Stochastic Realization Theory

So far we have studied state space construction in the presence of exogenous inputs, based on the concept of *Oblique Markovian splitting subspace*. In this section we shall examine the relation between oblique splitting and the classical construction of stochastic realization theory based on (orthogonal) splitting. We shall show that, in the absence of feedback, stochastic realizations with inputs can be constructed directly from stochastic realizations of the joint input-output process.

Let $(\mathcal{S}_J, \bar{\mathcal{S}}_J)$ be an orthogonal scattering pair for the joint process $\begin{bmatrix}\boldsymbol{y}^\top & \boldsymbol{u}^\top\end{bmatrix}^\top$, where the subscript J stands for joint, and let us assume that there is no feedback form $\boldsymbol{y}$ to $\boldsymbol{u}$. The following technical lemmas will be useful.

LEMMA 7.16
Let $\begin{bmatrix}\boldsymbol{y}^\top & \boldsymbol{u}^\top\end{bmatrix}^\top$ be a stationary process, and assume that there is no feedback form $\boldsymbol{y}$ to $\boldsymbol{u}$. Then there exist joint Markovian splitting subspaces $\mathcal{X}_J \equiv (\mathcal{S}_J, \bar{\mathcal{S}}_J)$ such that

$$\mathcal{S}_J \perp \mathcal{U}^+ \mid \mathcal{U}^- \tag{7.76}$$

□

Proof We just need to show that there exists at least one. Let us consider any causal realization, and let $\mathcal{X}_J$ be its state space. It is clear that by causality $\mathcal{X}_J \subset \mathcal{Y}^- \vee \mathcal{U}^-$, which implies that $\mathcal{S}_J = \mathcal{Y}^- \vee \mathcal{U}^-$ and therefore (7.76) follows. □

DEFINITION 7.25
In the sequel we shall say that joint Markovian splitting subspaces (realizations) which satisfy (7.76) are *feedback free*. □

LEMMA 7.17
Let $\mathcal{X}_J \equiv (\mathcal{S}_J, \bar{\mathcal{S}}_J)$ be feedback free, then

$$\mathcal{S}_J \cap \mathcal{U}^+ = \{0\}.$$

□

Proof From (7.76) $\mathrm{E}\left[\mathcal{S}_J \mid (\mathcal{U}^-)^\perp\right] \perp \mathcal{U}$, so that any element $s \in \mathcal{S}_J$ can be uniquely decomposed as $s = \hat{s} + \tilde{s}$ where $\hat{s} := E[s \mid \mathcal{U}^-] \in \mathcal{U}^-$ and $\tilde{s} \perp \mathcal{U}$. Let us assume that $s \in \mathcal{U}^+$; it follows that $\hat{s} = E[E[\hat{s} \mid \mathcal{U}^+] \mid \mathcal{U}^-]$ and therefore $\hat{s} \in \mathcal{U}^+ \cap \mathcal{U}^-$ which, recalling the sufficiently rich assumption (7.12), implies $\hat{s} = 0$ and therefore $\tilde{s} \in \mathcal{U}^+$; hence, $\tilde{s} = 0$ and therefore $s = 0$, which concludes the proof. □

We are now ready to state the following result.

THEOREM 7.26
Let $\mathcal{X}_J := (\mathcal{S}_J, \bar{\mathcal{S}}_J)$ be a feedback free realization of the stationary process $[\boldsymbol{y}^\top \boldsymbol{u}^\top]^\top$ Then $\mathcal{X}_J$ is an oblique Markovian splitting subspace. □

Proof We just need to verify the conditions of Theorem 7.14; $\mathcal{Y}^+ \subseteq \bar{\mathcal{S}}_J$ and $\mathcal{P}^- \subseteq \mathcal{S}_J$ by construction; the fact that $\mathcal{S}_J \cap \mathcal{U}^+ = 0$ follows from Lemma 7.17; forward and backward shift invariance follow from the fact that $(\mathcal{S}_J, \bar{\mathcal{S}}_J)$ is a scattering pair and the oblique intersection property holds since, in particular, also

$$\mathcal{S}_J \perp \bar{\mathcal{S}}_J \mid \mathcal{S}_J \cap \bar{\mathcal{S}}_J$$

holds. □

However, as one may expect, $\mathcal{X}_J$ is not, in general, a minimal oblique Markovian splitting subspace.

In order to construct a minimal oblique Markovian splitting subspace we can follow the procedure described in the previous section.

Let us assume that $\mathcal{X}_J := \mathcal{S}_J \cap \bar{\mathcal{S}}_J$ is a minimal Markovian splitting subspace for the joint process. The reason for $\mathcal{X}_J$ being not minimal oblique splitting, is that it includes the dynamics of the input process $\boldsymbol{u}$. We want to "factor out" this dynamics.

The basic idea in the reduction process is to consider first an "extended" state space

$$\mathcal{X}_{Je} := \mathcal{X}_J + \mathcal{U}^+$$

together with the associated "extended" pair $(\mathcal{S}_{Je}, \bar{\mathcal{S}}_{Je})$, $\mathcal{S}_{Je} := \mathcal{S}_J + \mathcal{U}^+$, $\bar{\mathcal{S}}_{Je} := \bar{\mathcal{S}}_J \vee \mathcal{U}^+$. We proceed to reduce this subspace, subject to the constraint that it must always contain $\mathcal{U}^+$.

Denote, as in the previous section, by $(\mathcal{S}^1_{Je}, \bar{\mathcal{S}}^1_{Je})$ the reduced pair, let $\mathcal{X}^1_{Je} := \mathcal{S}^1_{Je} \cap \bar{\mathcal{S}}^1_{Je}$ and let

$$\mathcal{X}^1 := \mathcal{X}_J \cap \left(\mathcal{S}^1_{Je} \cap \bar{\mathcal{S}}^1_{Je}\right) = \mathcal{X}_J \cap \mathcal{X}^1_{Je}$$

Introduce the generating process, $\mathcal{W}_t$, for $\mathcal{S}^1_{Je}$, as

$$\left(\mathcal{S}^1_{Je}\right)_{t+1} = \left(\mathcal{S}^1_{Je}\right)_t \oplus \mathcal{W}_t.$$

Of course the subspaces $\{\mathcal{W}_t\}$ are pairwise orthogonal, i.e. $\mathcal{W}_t \perp \mathcal{W}_s \quad t \neq s$.

Since $\mathcal{X}^1_{Je}$ is (orthogonally) Markovian splitting, the usual state update equation in geometric form holds

$$\left(\mathcal{X}^1_{Je}\right)_{t+1} \subseteq \left(\mathcal{X}^1_{Je}\right)_t \oplus \mathcal{W}_t.$$

Using Lemma 7.14, the last equation can be written as

$$\mathcal{X}^1_{t+1} + \mathcal{U}^+_{t+1} \subseteq \left(\mathcal{X}^1_t + \mathcal{U}^+_t\right) \oplus \mathcal{W}_t. \tag{7.77}$$

THEOREM 7.27
The subspace $\mathcal{X}^1$ is a minimal oblique splitting subspace. In fact we have

$$\begin{cases} \mathcal{X}^1_{t+1} & \subseteq & (\mathcal{X}^1_t + \mathcal{U}_t) \oplus \mathcal{W}_t \\ \mathcal{Y}_t & \subseteq & (\mathcal{X}^1_t + \mathcal{U}_t) \oplus \mathcal{W}_t. \end{cases}$$

□

Proof We show that (7.77) is equivalent to

$$\begin{cases} \mathcal{X}^1_{t+1} & \subseteq & (\mathcal{X}^1_t + \mathcal{U}_t) \oplus \mathcal{W}_t \\ \mathcal{U}^+_{t+1} & \subseteq & \mathcal{U}^+_t. \end{cases}$$

In fact, if this was not the case, there would be elements $\boldsymbol{x}^1_{t+1} \in \mathcal{X}^1_{t+1} \subseteq (\mathcal{S}_J)_{t+1}$ which could be written as $\boldsymbol{x}^1_{t+1} = (\boldsymbol{x}^1_t + \boldsymbol{u}_t + \boldsymbol{w}_t) + \boldsymbol{u}^+_{t+1}$ where $(\boldsymbol{x}^1_t + \boldsymbol{u}_t + \boldsymbol{w}_t) \in (\mathcal{S}_J)_{t+1}$, $\boldsymbol{u}^+_{t+1} \in \mathcal{U}^+_{t+1}$, which would imply that $\boldsymbol{u}^+_{t+1} = \boldsymbol{x}^1_{t+1} - (\boldsymbol{x}^1_t + \boldsymbol{u}_t + \boldsymbol{w}_t) \in (\mathcal{S}_J)_{t+1}$ contradicting the fact that $(\mathcal{S}_J)_{t+1} \cap \mathcal{U}^+_{t+1} = \{0\}$. Similarly, since $\mathcal{Y}_t \subseteq (\mathcal{S}_J)_{t+1}$ it follows that

$$\begin{cases} \mathcal{Y}_t & \subseteq & (\mathcal{X}^1_t + \mathcal{U}_t) \oplus \mathcal{W}_t \\ \mathcal{U}_t & \subseteq & \mathcal{U}^+_t. \end{cases}$$

Since $\boldsymbol{u}$ is known, we can "drop" the second part of the state equations, i.e. the component lying on $\mathcal{U}^+$ and end up with the equation state din the Theorem. This equation obcviously implies the oblique splitting property of $\mathcal{X}^1$.

Note that, by construction, $\mathcal{X}^1 + \mathcal{U}^+$ is the minimal splitting subspace containing the future of the input process, which, as it turns out, corresponds to the fact that $\mathcal{X}^1$ is the minimal subspace of $\mathcal{X}_J$ which is oblique splitting. □

It is natural to ask when a minimal (feedback free) Markovian splitting subspace for the joint process is a minimal oblique Markovian splitting subspace. It turns out that this is true if and only if the input predictor space is contained in the oblique Markovian splitting state.

THEOREM 7.28
Let $\mathcal{X}_J$ be a (feedback free) minimal Markovian splitting subspace for the joint process $[\boldsymbol{y}^\top \quad \boldsymbol{u}^\top]^\top$ and let $\mathcal{X}_u^{+/-} := E[\mathcal{U}^+ \mid \mathcal{U}^-]$ be the predictor space of the input process. Then $\mathcal{X}^1 = \mathcal{X}_J$ if and only if $\mathcal{X}_u^{+/-} \subseteq \mathcal{X}^1$. □

Proof Since $\mathcal{X}_J$ is minimal it is constructible. Therefore, $\mathcal{S}_J = \bar{\mathcal{S}}_J^{\perp} \vee \mathcal{U}^- \vee \mathcal{Y}^-$, which implies that $\mathcal{S}_J$ is minimal. By observability we have that

$$E\left[\mathcal{U}^+ \vee \mathcal{Y}^+ \mid \mathcal{S}_J\right] = \mathcal{X}_J. \tag{7.78}$$

On the other hand

$$E\left[\mathcal{U}^+ \vee \mathcal{Y}^+ \mid \mathcal{S}_J + \mathcal{U}^+\right] = \mathcal{X}^1 + \mathcal{U}^+. \tag{7.79}$$

Equation (7.78) can be rewritten as

$$\begin{aligned} E\left[E\left[\mathcal{U}^+ \vee \mathcal{Y}^+ \mid \mathcal{S}_J + \mathcal{U}^+\right] \mid \mathcal{S}_J\right] &= E\left[\mathcal{X}^1 + \mathcal{U}^+ \mid \mathcal{S}_J\right] = \\ &= \mathcal{X}^1 \vee E\left[\mathcal{U}^+ \mid \mathcal{S}_J\right] = \\ &= \mathcal{X}^1 \vee E\left[\mathcal{U}^+ \mid \mathcal{U}^-\right] = \\ &= \mathcal{X}^1 \vee \mathcal{X}_u^{+/-}. \end{aligned} \tag{7.80}$$

Therefore $\mathcal{X}_J = \mathcal{X}^1 \vee \mathcal{X}_u^{+/-}$ which implies that $\mathcal{X}_J = \mathcal{X}^1$ if and only if $\mathcal{X}_u^{+/-} \subseteq \mathcal{X}^1$. □

It is natural to ask what kind of situations may lead to such degeneracy. In order to address this problem we shall make a finite dimensionality assumption and work with the spectral representations of these spaces. We shall have to refer the reader to [27] for details.

Let $d\hat{\boldsymbol{z}} := [d\hat{\boldsymbol{u}}^\top \quad d\hat{\boldsymbol{w}}^\top]^\top$ be the spectral measure (i.e. the Fourier transform [30]) of the joint stationary process $\left[\boldsymbol{u}^\top(t)\ \boldsymbol{w}^\top(t)\right]^\top$. Let also (A, B, C, D, K) be a minimal (oblique) realization of $\boldsymbol{y}$ with state space $\mathcal{X}^1$ and let (A_u, K_u, C_u, I) be a minimal innovation representation (with state space $\mathcal{X}_u^{+/-}$) of $\boldsymbol{u}$. The spectral representations, $\hat{\mathcal{X}}^1$, and $\hat{\mathcal{X}}_u^{+/-}$, of $\mathcal{X}^1$ and $\mathcal{X}_u^{+/-}$ with respect to the spectral measure $d\hat{\boldsymbol{z}}$ are given by:

$$\hat{\mathcal{X}}^1 = \overline{\text{row-span}}\ \left\{\left[(zI - A)^{-1}B \qquad (zI - A)^{-1}K\right]\right\} \tag{7.81}$$

and

$$\hat{\mathcal{X}}_u^{+/-} = \overline{\text{row-span}}\ \left\{\left[(zI - (A_u - K_uC_u))^{-1}K_u \qquad 0\right]\right\}. \tag{7.82}$$

We are now able to give precise conditions for $\mathcal{X}^1$ and $\mathcal{X}_J$ to be the same space.

PROPOSITION 7.19
Let (A, B, C, D, K) be a minimal (oblique) realization of $\boldsymbol{y}$ with state space $\mathcal{X}^1$ and let (A_u, K_u, C_u, I) be a minimal innovation representation of $\boldsymbol{u}$ (with state space $\mathcal{X}_u^{+/-}$). Then $\mathcal{X}^1 = \mathcal{X}_J$ if and only if there exist a nonsingular change of basis T such that:

$$TAT^{-1} = \begin{bmatrix} A_u - K_uC_u & 0 \\ * & * \end{bmatrix}$$

and

$$TB = \begin{bmatrix} K_u \\ * \end{bmatrix} \quad , \quad TK = \begin{bmatrix} 0 \\ * \end{bmatrix}$$

□

The proof of this Proposition is a bit lengthy and will be given in another publication.

7.9 Conclusions

In this paper we have presented the basic ideas for a comprehensive theory of stochastic realization in the presence of exogenous inputs. The central concept of *Oblique Markovian Splitting Subspace* leads in principle to state-space construction and to a coordinate-free analysis of stochastic models with inputs. Most of the ideas are applicable to the general case where feedback is present, however there seem to be still some gaps to be filled, in particular we need to understand better how to deal with mixed causality structures as they occur in feedback interconnections where $F(z)$ may be unstable. Some new idea and additional work are needed.

Acknowledgment

This paper is dedicated to Anders Lindquist (Professor Noci) as a token of esteem and friendship over many years.

7.10 References

[1] H. Akaike. Stochastic theory of minimal realization. *IEEE Trans. Automat. Contr.*, 19(6):667–674, 1974.

[2] H. Akaike. Markovian representation of stochastic processes by canonical variables. *SIAM J. Control*, 13:162–173, 1975.

[3] P.E. Caines and C.W. Chan. Estimation, identification and feedback. In R. Mehra and D. Lainiotis, editors, *System Identification: Advances and Case Studies*, pages 349–405. Academic, 1976.

[4] A. Chiuso and G. Picci. On the ill-conditioning of subspace identification with inputs. Tech. Report TRITA/MATH-01-OS5, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, 2001. submitted for pubblication.

[5] Dym H. and McKean H. P. *Gaussian Processes, Function Theory and the Inverse Spectral Problem*, Academic Press, New York, 1976.

[6] M.R. Gevers and B.D.O. Anderson. Representation of jointly stationary feedback free processes. *Intern. Journal of Control*, 33:777–809, 1981.

[7] M.R. Gevers and B.D.O. Anderson. On jointly stationary feedback-free stochastic processes. *IEEE Trans. Automat. Contr.*, 27:431–436, 1982.

[8] C.W.J. Granger. Economic processes involving feedback. *Information and Control*, 6:28–48, 1963.

[9] E.J. Hannan and D.S. Poskitt. Unit canonical correlations between future and past. *The Annals of Statistics*, 16:784–790, 1988.

[10] H. Hotelling. Relations between two sets of variables. *Biometrica*, 28:321–377, 1936.

[11] M. Jansson, *On consistency of subspace methods for system identification*, Automatica **34** (1998), 1507–1519.

[12] T. Katayama and G. Picci. Realization of stochastic systems with exogenous inputs and subspace system identification methods. *Automatica*, 35(10):1635–1652, 1999.

[13] W.E. Larimore. System identification, reduced-order filtering and modeling via canonical variate analysis. In *Proc. American Control Conference*, pages 445–451, 1983.

[14] W.E. Larimore. Canonical variate analysis in identification, filtering, and adaptive control. In *Proc. 29th IEEE Conf. Decision & Control*, pages 596–604, Honolulu, 1990.

[15] A. Lindquist and G. Picci: On the stochastic realization problem *SIAM Journal on Control and Optimization* **17**, No. 3, pp. 365-389, 1979.

[16] A. Lindquist, G. Picci and G. Ruckebush, On minimal splitting subspaces and Markovian representation *Mathematical Systems Theory*, **12**, pp. 271-279, 1979.

[17] A. Lindquist and G. Picci. *Linear Stochastic Systems*. (book, to appear).

[18] A. Lindquist and G. Picci. Realization theory for multivariate stationary gaussian processes. *SIAM J. on control and Optimiz.*, 23(6):809–857, 1985.

[19] A. Lindquist and G. Picci. A geometric approach to modelling and estimation of linear stochastic systems. *Journal of Mathematical Systems, Estimation and Control*, 1:241–333, 1991.

[20] A. Lindquist and G. Picci. Canonical correlation analysis approximate covariance extension and identification of stationary time series. *Automatica*, 32:709–733, 1996.

[21] P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993.

[22] P. Van Overschee and B. De Moor. *N4SID: Subspace algorithms for the identification of combined deterministic– stochastic systems*, Automatica **30** (1994), 75–93.

[23] P. Van Overschee and B. De Moor. *Subspace identification for linear systems*, Kluwer Academic Publications, 1996.

[24] P.D.Lax and R.S.Phillips. *Scattering Theory*. Academic Press, NewYork, 1967.

[25] G. Picci. Stochastic realization of gaussian processes. *Proc. of the IEEE*, 64:112–122, 1976.

[26] G. Picci. Geometric methods in stochastic realization and system identification. In *CWI Quarterly special Issue on System Theory*, volume 9, pages 205–240, 1996.

[27] G. Picci. Oblique splitting susbspaces and stochastic realization with inputs. In D. Prätzel-Wolters U. Helmke and E. Zerz, editors, *Operators, Systems and Linear Algebra*, pages 157–174, Stuttgart, 1997. Teubner,.

[28] G. Picci. Stochastic realization and system identification. In T. Katayama and I. Sugimoto, editors, *Statistical Methods in Control and Signal Processing*, pages 205–240, N.Y., 1997. M. Dekker.

[29] G. Picci and S. Pinzoni Acausal Models and Balanced realizations of stationary processes *Linear Algebra and its Applications* (special issue on Systems Theory), **205-206**, pp. 957-1003, 1994.

[30] Y. A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967.

[31] M. Verhaegen, *Identification of the deterministic part of mimo state space models given in innovations form from input-output data*, Automatica **30** (1994), 61–74.

[32] M. Verhaegen and P. Dewilde, *Subspace model identification, part 1. the output-error state-space model identification class of algorithms; part 2. analysis of the elementary output-error state-space model identification algorithm*, Int. J. Control **56** (1992), 1187–1210 & 1211–1241.

8

Linear Fractional Transformations

Harry Dym

Abstract

A new formula for the linear fractional transformation of the Schur class by a J-inner matrix valued function is presented and applications to bitangential interpolation are outlined. Some of the surveyed results are connected with the role of Riccati equations in the the theory of reproducing kernel Hilbert spaces of the de Branges type.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 127-133, 2003.
© Springer-Verlag Berlin Heidelberg 2003

8.1 A Problem

Let Ω_+ stand for either the open unit disc $\mathbb{D}$, or the open upper half plane $\mathbb{C}_+$, or the open right half plane $\mathbb{IC}_+$. Let $\mathcal{S}^{p\times q}(\Omega_+)$ denote the Schur class of $p\times q$ mvf's (matrix valued functions) that are both holomorphic and contractive in Ω_+ and let

$$j_{pq}=\begin{bmatrix} I_p & 0\\ 0 & -I_q\end{bmatrix},\quad p\ge 1, q\ge 1 \text{ and } p+q=m.$$

It is well known that if

$$W(\lambda)=\begin{bmatrix} w_{11}(\lambda) & w_{12}(\lambda)\\ w_{21}(\lambda) & w_{22}(\lambda)\end{bmatrix}$$

is a j_{pq}-inner mvf with respect to Ω_+, with diagonal blocks $w_{11}(\lambda)$ of size $p\times p$ and $w_{22}(\lambda)$ of size $q\times q$, then the linear fractional transformation

$$T_W[\varepsilon]=(w_{11}\varepsilon+w_{12})(w_{21}\varepsilon+w_{22})^{-1} \tag{8.1}$$

maps every mvf $\varepsilon\in\mathcal{S}^{p\times q}(\Omega_+)$ into $\mathcal{S}^{p\times q}(\Omega_+)$; see e.g., [D1]. A more difficult problem is to furnish a useful description of the set

$$T_W[\mathcal{S}^{p\times q}]=\{T_W[\varepsilon]:\varepsilon\in\mathcal{S}^{p\times q}\}. \tag{8.2}$$

8.2 A Solution

It turns out that the problem formulated above has a neat solution in terms of the RKHS (reproducing kernel Hilbert space) $\mathcal{H}(W)$ that is based on the RK (reproducing kernel)

$$K_\omega(\lambda)=\frac{j_{pq}-W(\lambda)j_{pq}W(\omega)^*}{\rho_\omega(\lambda)},$$

where $\rho_\omega(\lambda)$ depends upon the choice of the region Ω_+:

$$\rho_\omega(\lambda)=\begin{cases} 1-\lambda\overline{\omega} & \text{if } \Omega_+=\mathbb{D}\\ -2\pi i(\lambda-\overline{\omega}) & \text{if } \Omega_+=\mathbb{C}_+\\ 2\pi(\lambda+\overline{\omega}) & \text{if } \Omega_+=\mathbb{IC}_+.\end{cases}$$

The class of RKHS's with kernels of the indicated form were characterized by L. de Branges [Br] for the case $\Omega_+=\mathbb{C}_+$. The case $\Omega_+=\mathbb{D}$, including an important technical improvement by Rovnyak [Ro] was worked out by Ball [Ba]. A unified approach that covers both these cases and more may be found in [AD]. The formulas for $\rho_\omega(\lambda)$ are connected with Cauchy's formula for the Hardy spaces $H_2(\Omega_+)$. In particular the mvf $1/\rho_\omega(\lambda)$ is the RK (reproducing kernel) for $H_2(\Omega_+)$, whereas $I_r/\rho_\omega(\lambda)$ is the RK for the vector Hardy space $H_2^r(\Omega_+)$ of $r\times 1$ vector valued functions with components in $H_2(\Omega_+)$.

The next theorem presents a solution to the problem of interest. The symbol $f^{\#}(\lambda)$ for a mvf $f(\lambda)$ that appears in the statement stands for $f(\lambda^\circ)^*$, where λ° denotes the reflection of λ about the boundary Ω_0 of Ω_+, i.e., $1/\overline{\lambda}$, $\overline{\lambda}$ or $-\overline{\lambda}$ according as Ω_+ is equal to $\mathbb{D}$, $\mathbb{C}_+$ or $\mathbb{IC}_+$, respectively. Also, $\Omega_-=\mathbb{C}\setminus\overline{\Omega_+}$.

THEOREM 8.2.1
Let $s \in \mathcal{S}^{p\times q}(\Omega_+)$ and let $W(\lambda)$ be a j_{pq}-inner mvf with respect to Ω_+. Then $s \in T_W[\mathcal{S}^{p\times q}]$ if and only if the following three conditions are met:

(1) $[I_p \quad -s]f$ belongs to the Hardy space $H_2^p(\Omega_+)$ for every choice of $f \in \mathcal{H}(W)$.

(2) $[-s^{\#} \quad I_q]f$ belongs to the Hardy space $H_2^q(\Omega_-)$ for every choice of $f \in \mathcal{H}(W)$.

(3) $\langle \Delta_s f, f\rangle_{st} \le \langle f, f\rangle_{\mathcal{H}(W)}$ for every choice of $f \in \mathcal{H}(W)$,

where $\Delta_s(\mu)$ denotes the mvf that is defined by the rule

$$\Delta_s(\mu) = \begin{bmatrix} I_p & -s(\mu) \\ -s(\mu)^* & I_q \end{bmatrix} \tag{8.3}$$

for a.e. point μ on the boundary Ω_0 of Ω_+ and $\langle\cdot,\cdot\rangle_{st}$ denotes the standard inner product with respect to Lebesgue measure (normalized by a factor of 2π if $\Omega_+ = \mathbb{D}$). □

The proof of the theorem is lengthy. It will be established in [D5].

8.3 An Application

The utility of this theorem rests on the fact that a number of bitangential interpolation problems can be formulated in terms that bear a striking resemblance to the description of $T_W[\mathcal{S}^{p\times q}]$ that is furnished in the theorem. Thus, for example, bitangential interpolation problems with a finite number of data points in the open right half plane $\mathbb{\Pi}_+$ can be formulated most simply and elegantly in terms of three complex matrices $C \in \mathbb{C}^{m\times n}$, $A \in \mathbb{C}^{n\times n}$, $P \in \mathbb{C}^{n\times n}$ that satisfy the following assumptions:[1]

(A1) $\sigma(A) \cap i\mathbb{R} = \varnothing$.

(A2) P is a positive semidefinite solution of the Lyapunov equation

$$A^*P + PA = -2\pi C^* j_{pq} C. \tag{8.4}$$

The notation

$$F(\lambda) = C(\lambda I_n - A)^{-1}\ ,\ \lambda \in \mathbb{C} \setminus \sigma(A), \quad \text{and} \tag{8.5}$$

$$P_s = \int_{-\infty}^{\infty} F(i\mu)^* \Delta_s(i\mu) F(i\mu) d\mu \tag{8.6}$$

will prove useful.

Let $\widehat{S}(C, A, P)$ denote the set of mvf's $s \in \mathcal{S}^{p\times q}(\mathbb{\Pi}_+)$ that meet the following three conditions.

(C1) $[I_p \quad -s]Fu$ belongs to the Hardy space $H_2^p(\mathbb{\Pi}_+)$ for every $u \in \mathbb{C}^n$.

[1] We shall also assume that A is an upper triangular matrix in Jordan form. This involves no loss of generality.

(C2) $[-s^{\#} \quad I_q]Fu$ belongs to the Hardy space $H_2^q(\mathbb{\Pi}_-)$ for every $u \in \mathbb{C}^n$.

(C3) $P_s \le P$.

For additional clarification and examples, see e.g., [D4].

The description of the set $\widehat{\mathcal{S}}(C,A,P)$ is clearly very close to the description of $T_W[\mathcal{S}^{p\times q}]$ that is furnished in Theorem 8.2.1. To bridge the gap we need to bring the RKHS $\mathcal{H}(W)$ into play. We shall discuss three cases.

P is invertible

Our first objective is to obtain a description of the set $\widehat{\mathcal{S}}(C,A,P)$ under assumptions (A1) and (A2) when P is invertible.

LEMMA 8.3.1
If assumptions (A1) and (A2) are in force and P is invertible, then the pair (C,A) is observable and the space

$$\mathcal{M} = \{F(\lambda)u : u \in \mathbb{C}^n\}$$

endowed with the inner product

$$\langle Fu, Fv\rangle_{\mathcal{M}} = v^*Pu$$

is a de Branges space $\mathcal{H}(W)$, i.e., it is a RKHS with RK

$$K_\omega(\lambda) = \frac{j_{pq} - W(\lambda)j_{pq}W(\omega)^*}{\rho_\omega(\lambda)},$$

where

$$W(\lambda) = I_m - 2\pi C(\lambda I_n - A)^{-1}P^{-1}C^* j_{pq}. \tag{8.7}$$

□

A proof of this lemma (up to minor changes of notation) is provided in [D4]. Next, upon invoking Theorem 8.2.1 in the setting of the lemma, we see that a mvf $s \in T_W[\mathcal{S}^{p\times q}]$ for the j_{pq}-inner mvf $W(\lambda)$ given by formula (8.7) if and only if $f(\lambda) = F(\lambda)u$ meets the conditions (C1), (C2) and

$$\langle \Delta_s f, f\rangle_{st} \le \langle f, f\rangle_{\mathcal{H}(W)}.$$

Thus, as

$$\langle f, f\rangle_{\mathcal{H}(W)} = u^*Pu,$$

these three conditions are seen to be exactly the same as those defining $\widehat{\mathcal{S}}(C,A,P)$. We have thus established the following theorem:

THEOREM 8.3.1
If (A1) and (A2) are in force, and P is invertible, then

$$\widehat{\mathcal{S}}(C,A,P) = T_W[\mathcal{S}^{p\times q}], \tag{8.8}$$

where $W(\lambda)$ is the j_{pq}-inner mvf with respect to $\mathbb{\Pi}_+$ that is defined by formula (8.7). □

P is singular and a related Riccati equation is solvable

Our next objective is to obtain a description of the set $\widehat{\mathcal{S}}(C,A,P)$ when P is singular. We shall first do so under the following extra assumption:
(A3) There exists an $n \times n$ complex matrix X such that:

1. $X \geq 0$.
2. X solves the Riccati equation

$$XA^* + AX = -2\pi X C^* j_{pq} C X \,. \tag{8.9}$$

3. $XPX = X$.
4. $PXP = P$.

If P is invertible, then the choice $X = P^{-1}$ fulfills all four requirements. If P is not invertible, then there is no guarantee that such an X exists. If it does, then in view of (1), (3) and (4), it must be a positive semidefinite pseudoinverse of P that has the same rank as P. But that is not enough, $\mathcal{R}_X$, the range of X, must be invariant under A. But this in turn is equivalent to the statement that X is a solution of the Riccati equation (8.9). A proof of the last assertion and the next lemma may be found in [D4]; for more on Riccati equations and RKHS's, see also [D2] and [D3].

LEMMA 8.3.2
If assumptions (A1)-(A3) are in force, then the space

$$\mathcal{M}_X = \{F(\lambda)Xu : u \in \mathbb{C}^n\}$$

endowed with the inner product

$$\langle FXu, FXv\rangle_{\mathcal{M}_X} = v^* X u$$

is a de Branges space $\mathcal{H}(W_X)$, i.e., it is a RKHS with RK

$$K_\omega(\lambda) = \frac{j_{pq} - W_X(\lambda) j_{pq} W_X(\omega)^*}{\rho_\omega(\lambda)},$$

where

$$W_X(\lambda) = I_m - 2\pi C(\lambda I_n - A)^{-1} X C^* j_{pq}. \tag{8.10}$$

□

Consequently, in the setting of the last lemma, Theorem 8.2.1 implies that $s \in T_{W_X}[\mathcal{S}^{p\times q}]$ if and only if it meets the following three conditions:

(D1) $[I_p \quad -s]FXy$ belongs to the Hardy space $H_2^p(\mathbb{\Pi}_+)$ for every $y \in \mathbb{C}^n$.

(D2) $[-s^\# \quad I_q]FXy$ belongs to the Hardy space $H_2^q(\mathbb{\Pi}_-)$ for every $y \in \mathbb{C}^n$.

(D3) $XP_sX \leq X$.

Thus, as the conditions (C1)–(C3) are more restrictive than the conditions (D1)–(D3), we see that if P is singular and the the extra assumption (A3) is in force, then

$$\widehat{\mathcal{S}}(C,A,P) \subset T_W[\mathcal{S}^{p\times q}].$$

Moreover, the inclusion may be proper.

Additional analysis, that is based in part on the observation that

$$\mathbb{C}^n = \mathcal{N}_P \dot{+} \mathcal{R}_X \tag{8.11}$$

is the direct sum of the null space of P and the range of X, leads to the following conclusion:

THEOREM 8.3.2
If assumptions (A1)-(A3) are in force, then there exist a pair of unitary matrices $U \in \mathbb{C}^{p\times p}$ and $V \in \mathbb{C}^{q\times q}$ such that

$$\widehat{\mathcal{S}}(C,A,P) = \left\{ T_{W_X}\left[U\begin{bmatrix} \widetilde{\varepsilon} & 0 \\ 0 & I_\nu \end{bmatrix} V^*\right] : \widetilde{\varepsilon} \in \mathcal{S}^{(p-\nu)\times(q-\nu)}(\mathbb{\Pi}_+) \right\} . \tag{8.12}$$

Moreover,

$$\nu = \operatorname{rank}\{P + C_1^*C_1\} - \operatorname{rank}P = \operatorname{rank}\{P + C_2^*C_2\} - \operatorname{rank}P, \tag{8.13}$$

where C_1 and C_2 are the top and bottom block components of C of sizes $p \times n$ and $q \times n$, respectively (i.e., $C^* = [C_1^* \;\; C_2^*]$). □

A proof may be found in [D4]. Formula (8.12) is not new, however, the present approach seems eminently natural (at least to the author). Formula (8.13) can be obtained directly, as in [D4], or from the analysis in [BD]. A number of additional references to the literature may also be found there. A particularly comprehensive approach to interpolation theory, that is not as well known as it deserves to be, is the abstract interpolation problem of Katsnelson, Kheifets and Yuditskii; see [Kh] for a recent survey, extensions and references to the earlier papers.

The finishing touch

Not every choice of matrices A, C, P that satisfy assumptions (A1) and (A2) will also satisfy (A3). Nevertheless, this difficulty may be overcome by modifying the off diagonal entries in A "a little" in order to obtain a new upper triangular matrix A_0 such that for this new set of data all three assumptions (A1)-(A3) will hold and

$$\widehat{\mathcal{S}}(C,A,P) = \widehat{\mathcal{S}}(C,A_0,P) \; .$$

The verification of this statement depends upon the analysis of an analogous issue in [BD]. Details for a case that is close to the one at hand are provided in [D4].

Acknowledgment

The author thanks Renee and Jay Weiss for endowing the Chair that supports his research.

8.4 References

[AD] D. Alpay and H. Dym, On a new class of structured reproducing kernel spaces, *J. Funct. Anal.*, **111** (1993), 1-28.

[Ba] J.A. Ball, Models for non contractions, *J. Math. Anal. Apl.*, **52** (1975), 235–254.

[BD] V. Bolotnikov and H. Dym, On degenerate interpolation, entropy and extremal problems for matrix Schur functions, *Integral Equations Operator Theory* **32** (1998), 367–435.

[Br] L. de Branges, Some Hilbert spaces of analytic functions I, *Trans. Amer. Math. Soc.*, **106** (1963), 445-668.

[D1] H. Dym, *J Contractive Matrices, Reproducing Kernel Hilbert Spaces and Interpolation*, CBMS Regional Conference Series, American Mathematical Society, **71** Providence, R.I., 1989.

[D2] H. Dym, On Riccati equations and reproducing kernel spaces, in: *Recent Advances in Operator Theory, Oper. Theory: Adv. Appl.*, **OT124**, Birkhäuser, Basel, 2001, pp. 189-215.

[D3] H. Dym, Reproducing kernels and Riccati equations, *Int. J. Appl. Math. Comp. Science*, **11** (2001), 35-53.

[D4] H. Dym, Riccati equations and interpolation, Contemporary Mathematics (V. Olshevsky, ed.), Amer. Math. Soc., Providence, R.I.,in press.

[D5] H. Dym, Linear fractional transformations, Riccati equations and bitangential interpolation, revisited, in preparation

[Kh] A.Ya. Kheifets, Abstract interpolation scheme for harmonic functions, in: *Interpolation Theory, Systems Theory and Related Topics, Oper. Theory Adv. Appl.*, **134**, Birkhäuser, Basel, 2002, pp. 287-317.

[Ro] J. Rovnyak, Characterizations of spaces $\mathbf{K}(M)$, unpublished manuscript, 1968.

9

Structured Covariances and Related Approximation Questions

Tryphon T. Georgiou

Abstract

Covariance matrices for the state and the output of a linear filter satisfy certain linear constraints dictated by the filter dynamics. Yet, sample covariances often do not. The renewed attention in using state/output statistics for estimating the power spectrum of the input [1, 2] raises the issue of approximating sample statistics so as to abide by the required linear constraints. In the present paper we motivate and formulate relevant questions.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 135-140, 2003.
© Springer-Verlag Berlin Heidelberg 2003

9.1 Introduction

The use of second order (covariance) statistics in spectral estimation was explored extensively during the 1980's and 1990's. This early work centered around the paradigm of uniformly sampled time-series and of linear equi-spaced arrays of sensors for which the covariance data have a Toeplitz structure. Hence, the theory of Toeplitz matrices and of the associated Carathèodory interpolation problem provided the basis for much of the subsequent development and gave rise to a number of alternative methodologies (see [10, 14, 5]).

In recent years, starting with [1, 2], renewed interest has arisen in utilizing state and output covariances of linear filters so as to extract information about the power spectrum of the input process. Such linear filters represent physical or algorithmic devices used in the measurement and, in general, their state covariances have structures which include Toeplitz and Pick matrices as special cases. The structure of state covariances is due to the filter dynamics and relates to an underlying Nevanlinna-Pick-Sarason interpolation problem instead (see [1, 2, 6]).

Whether Toeplitz, Pick, or more general state covariances, such matrices represent a rather "thin" subset of square matrices. At the same time, in practice they are all estimated from finite observation records. As a consequence, sample covariances almost always fail to have the required structure. Therefore, it is natural to ask for approximating sample covariances by nonnegative matrices having the required structure.

In the present paper we summarize basic facts about the structure of state covariances, we discuss the relevant approximation problem, and conclude with two computationally attractive formulations.

9.2 Structured Covariances

Let x_k be the stationary state of a linear filter

$$x_k \quad = \quad Ax_{k-1} + Bu_k, \quad k \in \mathbb{Z}, \tag{9.1}$$

where the input u_k is a stationary, zero-mean, real-valued stochastic process, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, (A, B) is a reachable pair, and A has eigenvalues in the open unit disc. The state covariance

$$\Sigma_{xx} := \mathcal{E}\{x_k x_k'\},$$

($\mathcal{E}$ being the expectation operator) satisfies linear constraints imposed by the system dynamics. More specifically, it can be shown that

$$\Sigma_{xx} - A\Sigma_{xx}A' = BH + H'B' \tag{9.2}$$

with $H \in \mathbb{R}^{m\times n}$ being a suitable matrix which depends on (A, B) as well as on the power spectrum of the input. Conversely, the solvability of (9.2) in terms of H, is sufficient for the nonnegative definite matrix Σ_{xx} to be the state covariance of the linear system (9.1) for some appropriate input (see [6, 7]).

The range of $H \mapsto \Sigma_{xx}$, as specified by (9.2), can be expressed as a linear set of symmetric matrices

$$\mathcal{S}_{A,B} := \{\Sigma_{xx} \ : \ \Sigma_{xx} = \sum_{i=1}^{q} x_i S_i\} \tag{9.3}$$

with $S_i = S_i'$ and $q = mn - m(m-1)/2$. To see this note that the null space of $H \mapsto \Sigma$ consists of matrices $H \in \mathbb{R}^{m\times n}$ of the form $H = M_s B'$ with M_s skew symmetric, and hence, has dimension $m(m-1)/2$.

9.3 Sample Covariances and the Approximation Problem

State covariances are typically estimated from a finite observation record $\{x_k \ : \ k = 1, \ldots, N\}$ as sample covariances

$$\hat{\Sigma}_{xx} := \frac{1}{N} \sum_{k=1}^{N} x_k x_k'.$$

Since the dimension of $\mathcal{S}_{A,B}$ is typically much less than that of of square symmetric matrices of size n, i.e.,

$$mn - m(m-1)/2 < n(n+1)/2,$$

it is expected that $\hat{\Sigma}_{xx}$ will fail to satisfy (9.2) almost always.

Yet, the theory of spectral estimation relies in a rather delicate manner on their satisfying (9.2). Indeed, spectral estimation based on covariance statistics amounts to characterizing all power spectra which are consistent with such statistics. Invariably, a "method" refers to a particular choice amongst the family of admissible spectra (see [14, 5, 6, 7]). Such a selection may be prejudiced (e.g., seeking a maximum entropy spectrum) or dictated by prior information about the nature of the sought spectrum (cf. [9]). But in all cases, the theory requires that the data (A, B) and Σ_{xx} to the problem, be consistent, i.e., satisfy (9.2).

In practice, most methods (such as MUSIC, ESPRIT, Burg's method, etc., see [14, 5]) use sample covariances instead although the effect of inaccuracies is not well understood and has not been analyzed in any detail—except via simulation studies.

The above issues motivate the following approximation problem:

PROBLEM 9.1
Given the filter parameters (A, B), a sampled covariance $\hat{\Sigma}_{xx}$ and a suitable distance measure $\delta(\hat{\Sigma}, \Sigma)$, find $\Sigma \geq 0$ which minimizes $\delta(\hat{\Sigma}, \cdot)$ and is consistent with the linear constraints (9.2). □

Below we discuss possible alternatives.

A convex optimization formulation

An interesting notion of distance has been introduced by von Neumann in order to quantify information and uncertainty in quantum systems. This is known as the quantum relative entropy and is defined (see [11]) by

$$\mathbb{S}(\hat{\Sigma}\|\Sigma) \quad := \quad \text{trace}\left(\hat{\Sigma}\left(\log\hat{\Sigma} - \log\Sigma\right)\right), \tag{9.4}$$

for $\hat{\Sigma}$ and Σ positive matrices with trace 1. When $\hat{\Sigma}$ and Σ commute and $\hat{\lambda}_i$ and λ_i $(i = 1, 2, \ldots, n)$ denote the corresponding eigenvalues, then

$$\mathbb{S}(\hat{\Sigma}\|\Sigma) = \sum_i \left(\hat{\lambda}_i\left(\log\hat{\lambda}_i - \log\lambda_i\right)\right)$$

reduces to the classical notion of relative entropy [3, 12].

The trace condition reflects the property that $\Sigma, \hat{\Sigma}$ are density matrices (or, statistical operators) of quantum systems and can be somewhat relaxed. For our purposes it suffices to assume that

$$\text{trace}(\Sigma) = \text{trace}(\hat{\Sigma}) =: \sigma.$$

The function $\mathbb{S}(\cdot\|\cdot)$, while not a metric, generates a useful topology. It is non-negative, jointly convex in its arguments, and it is equal to zero precisely when the two arguments are identical (see [11, Theorem 11.7]).

Returning to Problem 9.1, if we select $\delta(\cdot,\cdot) \equiv \mathbb{S}(\cdot\|\cdot)$ as our distance measure, then we have the following conclusion.

PROPOSITION 9.2
Given (A, B) as before and $\hat{\Sigma}_{xx} > 0$, then

$$\{\mathbb{S}(\hat{\Sigma}_{xx}\|\Sigma_{xx}) \ : \ \Sigma_{xx} \in \mathcal{S}_{A,B} \text{ and } \text{trace}(\Sigma_{xx}) = \sigma)\}$$

has a minimum attained at a unique minimizer $\Sigma > 0$. □

The proof follows readily from the fact that $\mathbb{S}(\hat{\Sigma}_{xx}\|\cdot)$ is convex and grows unbounded at the boundary of the set of positive matrices. The problem is computationally quite attractive since positivity of Σ is "built-in" and hence, there are no added constraints. A potential drawback is that it does not allow dealing with singular matrices.

A semidefinite programming formulation

Another naturally possibility for Problem 9.1 is to take as distance measure the one induced by the ordinary matrix norm

$$\delta(\hat{\Sigma}_{xx}, \Sigma_{xx}) = \|\hat{\Sigma}_{xx} - \Sigma_{xx}\|.$$

In this case we draw similar conclusions as well:

PROPOSITION 9.3
The set of values

$$\{\|\hat{\Sigma}_{xx} - \Sigma_{xx}\| \ : \ \Sigma_{xx} \in \mathcal{S}_{A,B} \text{ and } \Sigma_{xx} \geq 0\}$$

has a minimum attained at a unique minimizer $\Sigma \geq 0$. □

This is quite standard and the minimal value and minimizer can be computed via the semidefinite programming (e.g., see [15]). Indeed, if

$$S(x) = \hat{\Sigma}_{xx} - \sum_{i=1}^{q} x_i S_i$$

the problem

$$\begin{array}{c} \text{minimize } t \\ \text{subject to} \\ \hat{\Sigma}_{xx} - S(x) \geq 0 \\ tI_n - S(x) \geq 0 \\ tI_n + S(x) \geq 0 \end{array} \tag{9.5}$$

is an SDP in the variables t, x. Note that since $S(x)$ is a symmetric matrix, the condition $t > \|S(x)\|$ is equivalent to the two linear matrix inequalities $tI_n - S(x) \geq 0$ and $tI_n + S(x) \geq 0$ while I_n denotes the $n \times n$ identity matrix.

An analogous formulation for approximating Toeplitz matrices, in the matrix norm as well as the Frobenius norm, have been considered by Tom Luo [13].

9.4 Concluding Remarks

Toeplitz, Pick, or, more generally, state covariances of a given system, represent a rather "thin" subset of nonnegative matrices. Hence, sample state-statistics almost always fail to satisfy linear constraints that are dictated by the underlying dynamics. Yet, consistency with such dynamics is important in addressing the inverse problem of characterizing input power spectra that may have generated the data.

These facts motivate the problem of approximating sample state-statistics so as to enforce the required constraints inherited by the system dynamics. We indicated that such a problem has a computable solution for certain alternative choices of distance measure. Yet other choices need to be considered as well (trace distance, fidelity [11], etc.) while the final test of relevance needs a detailed sensitivity analysis in going from state covariances to families of consistent power spectra.

The present work has been motivated by collaborative work with A. Lindquist [9] where a relative entropy type of distance was introduced as a distance measure between power spectra.

9.5 References

[1] C. Byrnes, T.T. Georgiou, and A. Lindquist, A generalized entropy criterion for Nevanlinna-Pick interpolation: A convex optimization approach to certain problems in systems and control, *IEEE Trans. on Automatic Control*, **45(6):** 822-839, June 2001.

[2] C. I. Byrnes, T.T. Georgiou, and A. Lindquist, A new approach to spectral estimation: A tunable high-resolution spectral estimator, *IEEE Trans. on Signal Proc.*, **48(11):** 3189-3206, November 2000.

[3] T.M. Cover and J.A. Thomas, **Elements of Information Theory**, Wiley, 1991.

[4] T.T. Georgiou, Signal Estimation via Selective Harmonic Amplification: MUSIC, Redux, *IEEE Trans. on Signal Processing*, March 2000, **48(3)**: 780-790.

[5] T.T. Georgiou, Spectral Estimation via Selective Harmonic Amplification, *IEEE Trans. on Automatic Contr.*, January 2001, **46(1)**: 29-42.

[6] T.T. Georgiou, The structure of state covariances and its relation to the power spectrum of the input, *IEEE Trans. on Automatic Control*, **47(7):** 1056-1066, July 2002.

[7] T.T. Georgiou, Spectral analysis based on the state covariance: the maximum entropy spectrum and linear fractional parameterization, *IEEE Trans. on Automatic Control*, to appear.

[8] T.T. Georgiou, Toeplitz covariance matrices and the von Neumann relative entropy, preprint, May 2002.

[9] T.T. Georgiou and A. Lindquist, Kullback-Leibler approximation of spectral density functions, preprint, May 2002.

[10] S. Haykin, **Nonlinear Methods of Spectral Analysis,** Springer-Verlag, New York, 247 pages, 1979.

[11] M.A. Nielsen and I.L. Chuang, **Quantum Computation and Quantum Information,** Cambridge University Press, 2000.

[12] S. Kullback, **Information Theory and Statistics**, 2nd edition, New York: Dover Books, 1968 (1st ed. New York: John Wiley, 1959).

[13] T. Luo, notes and personal communication, August 2002.

[14] P. Stoica and R. Moses, **Introduction to Spectral Analysis**, Prentice Hall, 1997.

[15] L. Vandenberghe and S. Boyd, Semidefinite Programming, *SIAM Review,* **38(1):** 49-95, March 1996.

10

Risk Sensitive Identification of ARMA Processes

László Gerencsér *György Michaletzky*

Abstract

In this paper we consider the problem of recursive identification of ARMA processes. This recursive procedure is parameterized by a weight-matrix acting on the stochastic gradient. The optimal weight-matrix will be defined using a risk-sensitive identification criterion. First the cost function will be expressed using LEQG-theory. Then, applying stochastic realization theory and the bounded real lemma we derive alternative expressions for the cost function. We prove among others, that the LQG functional of a properly augmented system gives the LEQG cost function of the original system. Furthermore, we point out that this cost function can be interpreted as mutual information between two stochastic processes. The optimal weight-matrix will be computed first as the optimum of a multi-dimensional constrained minimization, then a direct approach for solving the optimization problem will be presented. Finally, we briefly indicate that the results above can be extended to multivariate stochastic systems.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 141-157, 2003.
© Springer-Verlag Berlin Heidelberg 2003

10.1 Weighted Recursive Prediction Error Identification

Let (y_n) be a wide-sense stationary ARMA (p,q) process satisfying the difference equation

$$A^* y = C^* e,$$

where A^* and C^* are polynomials of the shift operator of degree p and q respectively. We assume, that A^*, C^* are stable, relative prime, and the leading coefficients of A^* and C^* are equal to 1. The remaining coefficients of A^* and C^* are collected in a parameter vector θ^*. The noise process is assumed to fulfill the following standard conditions, in particular there exists an increasing family of σ-algebras $(\mathcal{F}_n)$ such that e_n is $\mathcal{F}_n$-measurable and

$$\mathrm{E}(e_n|\mathcal{F}_{n-1}) = 0, \quad \mathrm{E}(e_n^2|\mathcal{F}_{n-1}) = \sigma^2.$$

This is a minimal technical condition that is sufficient to derive the results of the next paragraph.

Let $D \subset \mathbb{R}^{p+q}$ denote the set of system parameters such that the corresponding polynomials A and C are stable. For fixed $\theta \in D$ define the process $\overline{\varepsilon}(\theta) = (\overline{\varepsilon}_n(\theta))$ by the difference equation $C\overline{\varepsilon} = Ay$, i.e.

$$\overline{\varepsilon} = (A/C)(C^*/A^*)e,$$

with $\overline{\varepsilon}_n = y_n = 0$ for $n \leq 0$. The asymptotic cost function is defined by

$$W(\theta) = \lim_{n\to\infty} \frac{1}{2\sigma^2} \mathrm{E}\overline{\varepsilon}_n^2(\theta).$$

It is well known and easily shown , that

$$\frac{\partial}{\partial\theta} W(\theta)\bigg|_{\theta=\theta^*} = 0 \quad \text{and} \quad R^* \stackrel{\Delta}{=} \frac{\partial^2}{\partial\theta^2} W(\theta)\bigg|_{\theta=\theta^*} > 0.$$

For a weighted recursive prediction error identification method (cf. [7]) let $\widehat{\theta}_n$ denote the estimator of θ^* at time n and the on-line estimate of $\overline{\varepsilon}_n(\widehat{\theta}_{n-1})$ is denoted by ε_n. They are constructed as follows. $\widehat{\theta}_0 \in D$ is an arbitrary initial guess and set $\varepsilon_n = y_n = 0$ for $n \leq 0$. Now, if $\widehat{\theta}_n$ and ε_n have already been generated for $n \leq N-1$ then define ε_N by the equation:

$$\left(\widehat{C}_{N-1}\varepsilon\right)_N = \left(\widehat{A}_{N-1}y\right)_N, \tag{10.1}$$

where $\widehat{A}_{N-1}$, $\widehat{C}_{N-1}$ denote the polynomials corresponding to $\widehat{\theta}_{N-1}$ and $(\cdot)_N$ denotes evaluation at time N.

Similarly, we define the on-line estimate of the gradient of the process $\overline{\varepsilon}(\theta)$, denoted by ε_θ, by

$$\left(\widehat{C}_{N-1}\varepsilon_\theta\right)_N = -\phi_{N-1}, \tag{10.2}$$

where $\phi_{N-1} = (-y_{N-1}, \cdots - y_{N-p}, \varepsilon_{N-1}, \ldots \varepsilon_{N-q})^T$. Then the weighted recursive prediction error estimate of θ^* at time N is defined by the recursion

$$\widehat{\theta}_N = \widehat{\theta}_{N-1} - \frac{1}{N} K \varepsilon_{\theta N} \varepsilon_N, \tag{10.3}$$

where K is a weighting matrix.

The asymptotic properties of $\widehat{\theta}_N$ have been rigorously analyzed in [1] and [3] under various technical conditions. In [1] it is required that $\hat{\theta}_N \in D_0 \subset D$, where D_0 is a prescribed compact domain, otherwise the process is stopped. In [3] the boundedness condition above is inforced by a resetting mechanism: if $\hat{\theta}_N \notin D_0$ then we redefine it to be $\widehat{\theta}_0$ again.

To sketch a key result of [1] define the estimation sequence $\widehat{\theta}_{N,k}$ using the recursion (10.3) but changing the stepsizes from $1/N$ to $1/(N+k)$ and consider a piecewise constant embedding defined by

$$\widehat{\theta}_{s,k} = \widehat{\theta}_{N,k} \quad \text{for} \quad s \in [N, N+1), \quad N \geq 1.$$

Introduce a normalized and rescaled process by first normalizing $\widehat{\theta}_{s,k} - \theta^*$ by $s^{1/2}$, followed by an exponential change of time-scale $s = e^t$. This yields a new process $\psi_k = (\psi_{s,k})$:

$$\psi_{t,k} = e^{t/2}(\widehat{\theta}_{e^t,k} - \theta^*), \quad t \geq 0.$$

It is claimed in Theorem 12 Chapter 4.5, Part II of [1] that $(\psi_{t,k})$ converges weakly, for $k \to \infty$, to the process $(\widetilde{x}(t))$ defined by

$$d\widetilde{x}(t) = (-KR^* + I/2)\widetilde{x}(t)dt + Gdw(t), \tag{10.4}$$

where $(w(t))$ is a standard $\mathbb{R}^{p+q}$-valued Wiener-process and G is the symmetric positive semidefinite square-root of KR^*K^T, thus

$$GG = KR^*K^T,$$

assuming that

$$F = -KR^* + I/2$$

is asymptotically stable.

For the recursive estimation procedure with enforced boundedness, that has been rigorously analyzed in [3], a corresponding result has not yet been fully derived, but a significant part of the analysis has been completed in [4] and [5].

A direct corollary of the cited result of [1] is that the asymptotic covariance matrix $S = S(K)$ of the estimator process $\widehat{\theta}_N$ exists and it is obviously given by

$$S = \mathrm{E}\widetilde{x}(t)\widetilde{x}(t)^T \ .$$

It is well-known that S is the solution of the Lyapunov-equation

$$(-KR^* + I/2)S + S(-KR^* + I/2)^T + KR^*K^T = 0.$$

Using the partial ordering for symmetric matrices. i.e. $A \leq B$ if and only if $B - A$ is positive semidefinite, where A, B are symmetric matrices, it is also well-known that $S = S(K)$ is minimized with respect this ordering for the choice $K = (R^*)^{-1}$. Then $F = -I/2$ and for the asymptotic covariance matrix of the error-process we have $S = (R^*)^{-1}$.

10.2 A Risk-Sensitive Criterion

A risk-sensitive approach to system identification has been first proposed in [8] for AR processes. The purpose of this section is to define a risk-sensitive performance index in a different way using the asymptotic theory of recursive estimation. The new criterion is applicable both for AR and for ARMA-systems, and even for multivariable linear stochastic systems. Expression for the new criterion will be given using LEQG-theory. We are going to consider recursive estimation procedures of the form (10.3) with a fixed weighting matrix K. Minimizing a risk-sensitive performance index with respect to K we arrive at a procedure which is very close to the procedure of [8].

Consider the criterion

$$\frac{2}{c}\mathrm{E}\left(\exp\left\{\frac{c}{2}\sum_{n=1}^{N-1}(\theta-\widehat{\theta}_n)^T(\theta-\widehat{\theta}_n)_n\right\}\right),$$

with a positive c. Using the piecewise constant embedding, followed by an exponential change of time-scale $s=e^t$, $N=e^T$ the above sum will become

$$\frac{2}{c}\mathrm{E}\left(\exp\{\frac{c}{2}\int_0^T \psi(t)^T\psi(t)dt\}\right). \tag{10.5}$$

Motivated by the above calculations and Theorem 12 Chapter 4.5, Part II of [1], the new risk-sensitive identification criterion is defined in terms of the stationary Gaussian process $\widetilde{x}(t)$ as

$$J(K)=\lim_{T\to\infty}\frac{2}{T\,c}\log\mathrm{E}\left(\exp\{\frac{c}{2}\int_0^T \widetilde{x}(t)^T H^T H\widetilde{x}(t)dt\}\right), \tag{10.6}$$

where $c>0$ and H^TH is some non-singular weighting-matrix. The study of this criterion and its optimization with respect to K is the subject matter of this paper. Functionals of the form similar to $J(K)$ are well-known in LEQG control (cf. e.g. [2, 6, 8]) and a number of useful expressions for $J(K)$ have been found. The following proposition is given as Proposition 6.3.1 in [6]. Recall that $KR^*K^T=GG$ with G symmetric and positive semidefinite. Let

$$\mathcal{G}(s)=H(sI-F)^{-1}G. \tag{10.7}$$

Proposition 10.1
Assume that $F=-KR^*+I/2$ is asymptotically stable.

(i) If the H_∞ norm of $c^{1/2}\mathcal{G}(s)$ is strictly less than 1, then $J(K)$ is finite and moreover

$$J(K)=\lim_{s_0\to\infty}-\frac{1}{2\pi}\int_{-\infty}^{\infty}\ln|\det\left(I-c\mathcal{G}(i\omega)\mathcal{G}^*(i\omega)\right)|\left[\frac{s_0^2}{s_0^2+\omega^2}\right]d\omega. \tag{10.8}$$

(ii) If the H_∞ norm of $c^{1/2}\mathcal{G}(s)$ is greater than 1, then the limit in $J(K)$ exists but it is infinite.

□

The set of K-s for which the H_∞-norm of $c^{1/2}\mathcal{G}$ is strictly less than 1 will be denoted by E_K:

$$E_K = \{K : ||c^{1/2}\mathcal{G}(s)||_\infty = ||c^{1/2}H(sI + KR^* - I/2)^{-1}(KR^*K^T)^{1/2}||_\infty < 1\} \quad (10.9)$$

where $||.||_\infty$ denotes the H_∞-norm. It is taken for granted that $\mathcal{G}(s)$ is analytic on the closed right-half plane, and thus $K \in E_K$ implies that $F = -KR^* + I/2$ is asymptotically stable. According to (i) of the previous proposition $J(K)$ is finite on E_K.

The set of K-s for which the H_∞-norm of $c^{1/2}\mathcal{G}$ is less than or equal to 1 and the functional $J(K)$ is finite will be denoted by E_K°:

$$E_K^\circ = \{K : ||c^{1/2}\mathcal{G}(s)||_\infty \leq 1 \quad \text{and } J(K) < \infty\} . \quad (10.10)$$

Here again it is taken for granted that $\mathcal{G}(s)$ is analytic on the closed right-half plane, and thus $K \in E_K^\circ$ implies that $F = -KR^* + I/2$ is asymptotically stable. Note that according to the above cited Proposition 6.3.1 in [6], if $\|c^{1/2}\mathcal{G}(s)\| > 1$ then $J(K)$ cannot be finite, thus the extra requirement in the definition of the set E_K° concerning the $\|H\|_\infty$ norm of $c^{1/2}\mathcal{G}$ was put for easier comparison with the definition of E_K. The sets E_K and E_K° obviously depend on c. In such cases when this is important the notations $E_K(c)$ and $E_K^\circ(c)$ will be used.

Based on Lemma 5 and Theorem 5 of [9] and Proposition 5.3.2 of [6] the following statement can be formulated.

PROPOSITION 10.2
Assume that $F = -KR^* + I/2$ is asymptotically stable. Then

(i) the H_∞ norm of $c^{1/2}\mathcal{G}(s)$ is less than or equal to 1 if and only if the control Riccati-equation:

$$F^TQ + QF + H^TH + cQGGQ = 0 \quad (10.11)$$

has a real symmetric solution Q for which the matrix $F + cGGQ$ is stable. Moreover, this solution is unique and positive definite.

(ii) the H_∞ norm of $c^{1/2}\mathcal{G}(s)$ is strictly less then 1 if and only if the control Riccati equation above has a real symmetric solution, for which the matrix $F + cGGQ$ is asymptotically stable.

Furthermore, in case (ii) the functional $J(K)$ is finite and

$$J(K) = \text{tr } GQG. \quad (10.12)$$

□

In the next section using equation (10.12) we outline the minimization problem of $J(K)$.

10.3 The Minimization of $J(K)$

First we prove the following simple proposition:

PROPOSITION 10.3
Assume that there exists a weighting matrix K such that $F = -KR^* + I/2$ is asymptotically stable $\|c^{1/2}\mathcal{G}(s)\|_\infty \leq 1$. Then

$$R^* > cH^TH\ .$$

□

PROOF. According to Proposition 10.2 there exists a symmetric, positive definite solution Q of the Riccati-equation (10.11). Substituting $F = -KR^* + I/2$ into (10.11), multiplying by c and completing to squares the terms containing K, we arrive at the equation

$$cQ + (cQK - I)R^*(cK^TQ - I) = R^* - cH^TH,$$

implying that $R^* - cH^TH$ is positive definite.

Consider now the problem of minimizing $J(K)$ with respect to K subject to the condition that $K \in E_K$. The first main result of the paper is the following:

THEOREM 10.1
Assume that $R^* - cH^TH$ is positive definite. Then the set E_K defined under (10.9) is non-empty, and $J(K)$ achieves its minimum over E_K at the unique minimizing K given by

$$K^* = (R^* - cHH^T)^{-1}.$$

The optimal cost is

$$J^* = J(K^*) = \text{tr } H^TH(R^* - cHH^T)^{-1}\ .$$

□

REMARK. Observe that Theorem 10.1 implies the following result which might be surprising at first sight. Consider the set

$$E_c = \left\{c \mid \text{there exists a } K \text{ such that } \|c^{\frac{1}{2}}\mathcal{G}\| \leq 1\right\}\ .$$

Then by the continuity of the H_∞-norm with respect to c and K the set E_c is an open interval.

We need the following notations: the set of K-s for which $F = -KR^* + I/2$ is asymptotically stable will be denoted by D_K:

$$D_K = \{K \in \mathbb{R}^{r\times r} :\ F = -KR^* + I/2 \text{ is asymptotically stable}\}.$$

Obviously D_K is an open domain in the set of $r \times r$ matrices, where $r = p + q$. The set of $r \times r$ positive definite symmetric matrices will be denoted by D_Q.

Consider the extended variable (K, Q) and, motivated by Proposition 10.2 define the relaxed constrained minimization problem. Set

$$J(K, Q) = \text{tr } QKR^*K^T ,$$

and

$$\begin{aligned} &\text{minimize} && J(K, Q) && (10.13)\\ &\text{subject to} && K \in D_K, \quad Q \in D_Q, \\ & && F^TQ + QF + HH^T + cQKR^*K^TQ = 0. && (10.14) \end{aligned}$$

By virtue of Proposition 10.2, if we add the condition that $F^T + cQKR^*K^T$ is asymptotically stable, then the constrained optimization problem is equivalent to minimizing $J(K)$ over E_K. The solution of this *relaxed* problem is given by the following theorem.

THEOREM 10.2
Assume that $R^* - cH^TH$ is positive definite. Then the constrained minimization problem (10.13), (10.14) has a unique solution

$$K^* = (R^* - cHH^T)^{-1}, \qquad Q^* = (R^* - cHH^T)(R^*)^{-1}H^TH. \tag{10.15}$$

□

It will be easy to show that Q^* is in fact a stabilizing solution in the sense that $F^T + cQKR^*K^T$ is asymptotically stable, and therefore K^* is a solution of the original problem.

PROOF. We are going to give the outline of two proofs to this theorem. The first one is based on applying Lagrange multipliers, while the second one is based on some algebraic manipulation of the Riccati-equation (10.14).

The following lemma shows that if K goes towards the boundary of the region D_K then the value of the functional $J(K, Q)$ evaluated under the constraint (10.14) goes to infinity.

Let us denote by h the smallest eigenvalue of H^TH and by r that of R^*. Then $r > 0$ and due to the assumption that H is nonsingular $h > 0$, as well. Furthermore, $\lambda_{\max}(F)$ denotes the eigenvalue of F with maximal real part. (Since its real part will be used in the following lemma, the possible ambiguity causes no problem.)

LEMMA 10.1
Assume that $K \in D_K$, i.e. $F = -KR^* + \frac{I}{2}$ is asymptotically stable and let $Q = Q^T$ be a solution of the Riccati equation (10.14). Then

$$\begin{aligned} J(K, Q) &\geq \frac{hr}{2} \frac{\text{tr } (KK^T)}{|\, \Re\lambda_{\max}(F)\,|} \geq \frac{hr \text{ tr } (KK^T)}{2\|K\|\|R\| - 1}, && (10.16)\\ \|K\| &> \frac{1}{2\|R^*\|} . && (10.17) \end{aligned}$$

□

Now since $J(K,Q)$ is continuous the previous lemma implies that its infimum is attained inside the feasible set. To find it the Lagrange-multipliers rule can be applied.

First, since the equation (10.14) is symmetric, the Lagrange-multipliers corresponding to the constraints (10.14) can be assumed to be represented by an $r \times r$ symmetric matrix Λ. The Lagrangian of the constrained optimization problem is as follows:

$$L(K,Q,\Lambda) = \mathrm{tr}\ (GQG + \Lambda(F^TQ + QF + H^TH + cQGGQ)). \tag{10.18}$$

We get after a sequence of elementary equations of matrix-analysis and writing shortly $R = R^*$, the following proposition:

PROPOSITION 10.4
For the Lagrangian-function defined under (10.18) we have

$$\frac{\partial}{\partial K}L(K,Q,\Lambda) = 2QKR^* - 2Q\Lambda R^* + 2cQ\Lambda QKR^*\,, \tag{10.19}$$

and

$$\frac{\partial}{\partial Q}L(K,Q,\Lambda) = GG + \Lambda F^T + F\Lambda + cGGQ\Lambda + c\Lambda QGG. \tag{10.20}$$

□

We have assumed that Q is non-singular. Setting

$$\frac{\partial}{\partial K}\mathrm{tr}\ L(K,Q,\Lambda) = 0 \tag{10.21}$$

and multiplying by Q^{-1} from the left, and by $(R^*)^{-1}$ from the right we get

$$K - \Lambda + c\Lambda QK = 0. \tag{10.22}$$

From this it is easy to see that Λ is nonsingular.

Multiply (10.22) by K^{-1} from the right and by Λ^{-1} from the left. Then we get

$$\Lambda^{-1} - K^{-1} + cQ = 0. \tag{10.23}$$

Since Λ^{-1} and Q are symmetric, it follows that K^{-1} and hence K is symmetric. From (10.23) we get that

$$Q = \frac{1}{c}(K^{-1} - \Lambda^{-1}). \tag{10.24}$$

Substituting this Q into (10.20) setting

$$\frac{\partial}{\partial Q}L(K,Q,\Lambda) = 0 \tag{10.25}$$

and working out the left hand side we get the following equation:

$$\Lambda = KR^*K. \tag{10.26}$$

Finally, substituting the value of Q given by (10.24) into the Riccati-equation (10.11) and using (10.26) in the last term we arrive at the conclusion that equations (10.21) and (10.25) together with the Riccati-equation uniquely determine the matrix K:

$$K^{-1} = R^* - cH^T H. \tag{10.27}$$

Then we get for Λ from (10.26):

$$\Lambda = (R^* - cH^T H)^{-1} R^* (R^* - cH^T H)^{-1} ,$$

and for Q from (10.24) and (10.26) :

$$Q = (R^* - cH^T H)(R^*)^{-1} H^T H$$

The unique solutions obtained above will be denoted by K^*, Q^* and Λ^*. Using again Lemma 10.1 we obtain that they determine the unique minimum point.

It is worth noting that the Hessian-matrix of $J(K)$ at $K = K^*$ generates the quadratic form

$$K'(0) \mapsto 2\text{tr } H^* H K'(0) R K'(0)^T ,$$

where $K'(0)$ is an arbitrary $r \times r$ matrix.

The optimal cost is

$$J^{**} = J(K^*, Q^*) = \text{tr } H^T H (R^* - cHH^T)^{-1}. \tag{10.28}$$

Proof of theorem 10.1. To prove Theorem 10.1 we need to show only that Q^* is a stabilizing solution. We have

$$F^T + cQ^* GG = (-R^* K^* + I/2) + cQ^* K^* R^* K^* = -I/2,$$

thus Q^* is indeed a stabilizing solution, concluding the proof of Theorem 10.1.

10.4 Alternative Expressions for $J(K)$

An alternative expression for $J(K)$ is given in the following proposition, which is immediate for example in view of Proposition 2.3.1 of [6] or Lemma 8 of [9].

Proposition 10.5
Assume that $F = -KR^* + I/2$ is asymptotically stable. Then the H_∞ norm of $c^{1/2}\mathcal{G}(s)$ is strictly less than 1 if and only if the filter Riccati-equation

$$FP + PF^T + cPH^T HP + GG = 0. \tag{10.29}$$

has a unique symmetric asymptotically stabilizing solution P. Then $J(K)$ is finite and

$$J(K) = \text{tr } (HPH^T) \tag{10.30}$$

□

Recall that P is an asymptotically stabilizing solution if $F+cPH^TH$ is asymptotically stable.

Remark. It follows immediately that, assuming that $J(K)$ is finite for sufficiently small c, for $c \searrow 0$ $J(K)$ converges to

$$\mathrm{E}\widetilde{x}^T(s)HH^T\widetilde{x}(s) = \mathrm{tr}\ S(K)HH^T.$$

A new expression for $J(K)$ is then obtained by factorizing $I-c\mathcal{G}\mathcal{G}^*$, see [6]. This is achieved using the bounded real lemma by constructing an auxiliary stochastic process. To this aim observe that using the solution of the equation (10.29) an all-pass extension of the original system can be defined.

In fact, consider the state-space equation

$$dx(t) = Fx(t)dt + c^{1/2}Gdw(t), \tag{10.31}$$

where $w(t)$ is a standard Wiener process. Note that $x(t) = c^{1/2}\widetilde{x}(t)$. Extend this system with

$$d\xi(t) = F\xi(t)dt - cPH^Tdb(t) \tag{10.32}$$

where $b(t)$ is a standard Wiener process, independent of $w(t)$ and $\xi(t)$ is the stationary solution of (10.32). Then the process y defined by

$$dy(t) = H(x(t)+\xi(t))dt + db(t) \tag{10.33}$$

is a standard Wiener process, i.e. the transfer function

$$\mathcal{N}(s) = I - H(sI-F)^{-1}cPH^T$$

provides an all-pass "extension" of the transfer function $c^{1/2}\mathcal{G}(s)$. In other words $\left[c^{1/2}\mathcal{G}(s), \mathcal{N}(s)\right]$ is all-pass (in particular it is inner, due to the stability of F), mapping the Wiener process $\begin{bmatrix} w(t) \\ b(t) \end{bmatrix}$ into the Wiener process $y(t)$. Indeed, it is easy to see, that a factorization of $I - c\mathcal{G}\mathcal{G}^*$ is given by $\mathcal{N}\mathcal{N}^*$, i.e.

$$I - c\mathcal{G}(i\omega)\mathcal{G}^*(i\omega) = \mathcal{N}(i\omega)\mathcal{N}^*(i\omega).$$

(Cf. Mustafa and Glover [6] Lemma 5.3.2.)

We might consider the "partial" inverse of the system (10.31), (10.32) and (10.33) taking w and y as inputs and b as the output process. This partial inverse system can be realized by the following state space equation with $x(t)+\xi(t)$ being the state-vector:

$$\begin{aligned} d\left(x(t)+\xi(t)\right) &= F\left(x(t)+\xi(t)\right) + \\ &\quad + c^{1/2}Gdw(t) - cPH^Tdb(t) \\ &= \left(F+cPH^TH\right)\left(x(t)+\xi(t)\right)dt + \\ &\quad + c^{1/2}Gdw(t) - cPH^Tdy(t) \end{aligned}$$

$$db(t) = -H\left(x(t)+\xi(t)\right)dt + dy(t)\ .$$

Observe that the solution of (10.29) can be expressed using the covariance matrix of $x(t) + \xi(t)$.

$$\mathrm{cov}(x(t) + \xi(t)) = cP.$$

In other words the asymptotic LEQG functional for the *original* system can be considered as the asymptotic LQG functional for this *augmented* system.

The following proposition provides another interesting characterization of the LEQG functional $J(K)$ via the mutual information rate between the processes y and w.

PROPOSITION 10.6
Assume that the $\|c^{1/2}\mathcal{G}\|_\infty \leq 1$. Consider the auxiliary process y defined by equations (10.32), (10.33) and let $I(y,w)$ denote the mutual information rate between the processes y and w. Then

$$J(K) = \frac{2}{c} I(y,w) = \mathrm{tr}\ PH^T H\ . \tag{10.34}$$

□

For the proof recall that the mutual information rate between two processes is defined as follows. Consider a finite value T and denote by $Q_{y,T}$ and $Q_{w,T}$ the distribution of the processes $y(s), 0 \leq s \leq T$ and $w(s), 0 \leq s \leq T$ defined on the space of continuous vector-valued functions and let $Q_{y,w,T}$ denote their joint distribution. Then

$$I(y,w) = -\lim_{T\to\infty} \frac{1}{T} E\left(\ln \frac{d\left(Q_{y,T} \times Q_{w,T}\right)}{dQ_{y,w,T}} \left(y(s), w(s), 0 \leq s \leq T\right)\right)\ , \tag{10.35}$$

assuming that the limit exists. It can be shown that, introducing the notation $\zeta = x + \xi$, we have

$$\frac{dQ_{y,T} \times dQ_{w,T}}{dQ_{y,w,T}} \left(y(s), w(s), 0 \leq s \leq T\right)$$
$$= \exp\left\{ -\int_0^T \zeta(t)^T \left[0, H^T\right] d \begin{bmatrix} dw(t) \\ db(b) \end{bmatrix} - \frac{1}{2}\int_0^T \zeta^T(t) \left[0, H^T\right] \begin{bmatrix} 0 \\ H \end{bmatrix} \zeta(t) dt \right\}.$$

Taking the logarithm and using that the first term is a square-integrable martingale with zero expectation we get that

$$I(y,w) = \lim_{T\to\infty} \frac{1}{T} E\left(\frac{1}{2}\int_0^T \zeta^T(t) \left[0, H^T\right] \begin{bmatrix} 0 \\ H \end{bmatrix} \zeta(t) dt \right)\ .$$

The covariance matrix of the stationary process ζ is cP, and thus

$$I(y,w) = \frac{c}{2} \mathrm{tr}\ \left(PH^T H\right)\ ,$$

proving that the mutual information rate between y and w exists. Taking into account the equation (10.30) the proposition follows.

Thus the minimization of $J(K)$ is equivalent to minimizing the mutual information rate between the fixed process w and the K-dependent process y.

An alternative proof of Theorem 10.1 can be obtained using the alternative form of the asymptotic cost functional given in (10.30) and applying the Lagrange multipliers method. Not going into details we remark that the value of the corresponding Lagrange-multipliers at the optimal solution is $H^T H$, and $P^* = K^*$

A genuinely new, direct, computational proof of Theorem 10.2 can be obtained using the filter Riccati-equation (10.29) representation of $J(K)$.

An outline of the second proof of theorem 10.2. First recall that if F is asymptotically stable then P is positive definite. The key idea is to rewrite the filter Riccati-equation can be written in the form

$$\left(-K\left(R^* - cH^T H\right) + \frac{I}{2}\right) P + P\left(-K\left(R^* - cH^T H\right) + \frac{I}{2}\right)^T + \\ + K\left(R^* - cH^T H\right) K^T + c(K-P)H^T H(K-P)^T = 0\,. \tag{10.36}$$

It is relatively easily shown using standard Lyapunov-equation techniques that the matrix

$$-K\left(R^* - cH^T H\right) + \frac{I}{2}$$

is asymptotically stable.

Next compare the matrix P in (10.36) with $\left(R^* - cH^T H\right)^{-1}$. Straightforward calculation gives that (10.36) can be written as:

$$\left[-K\left(R^* - cH^T H\right) + \frac{I}{2}\right]\left[P - \left(R^* - cH^T H\right)^{-1}\right] \\ + \left[P - \left(R^* - cH^T H\right)^{-1}\right]\left[-K\left(R^* - cH^T H\right) + \frac{I}{2}\right]^T \\ + \left[-K\left(R^* - cH^T H\right) + I\right]\left(R^* - cH^T H\right)^{-1}\left[-K\left(R^* - cH^T H\right) + I\right]^T \\ + c(K-P)H^T H(K-P)^T = 0\,. \tag{10.37}$$

Since the last two terms in this equation is positive semidefinite the asymptotic stability of the matrix $\frac{I}{2} - K\left(R^* - cH^T H\right)$ implies that

$$P \geq (R^* - cH^T H)^{-1} \tag{10.38}$$

Since the functional to be minimized can be written in the form $\operatorname{tr} PH^T H$ the argument above implies that

$$J(K) = \operatorname{tr}\, PH^T H \geq \operatorname{tr}\, \left(R^* - cH^T H\right)^{-1} H^T H\,,$$

and equality holds if and only if $P = \left(R^* - cH^T H\right)^{-1}$ corresponding to

$$K = \left(R^* - cH^T H\right)^{-1}\,,$$

concluding the proof of the theorem.

Now consider the original problem of minimizing $J(K)$ over the set where it is well-defined and finite, i.e. on E_K°. We have the following basic theorem:

THEOREM 10.3
Assume that $R^* - cH^TH$ is positive definite. Then the set E_K° defined under (10.10) is non-empty, and $J(K)$ achieves its minimum over E_K° at the unique minimizing K given by

$$K^* = (R^* - cHH^T)^{-1}.$$

The optimal cost is

$$J^* = J(K^*) = \text{tr } H^TH(R^* - cHH^T)^{-1}$$

□

PROOF. In view of Theorem 10.1 it is enough to prove that if $K \in E_K$ but $K \notin E_K^\circ$ then $J(K) > J(K^*)$, where $K^* = \left(R^* - cH^TH\right)^{-1}$ is the optimizing value in E_K.

Since in the course of this proof the normalizing parameter c will not be kept constant it is important to express its actual value in the notations. Thus we shall write $J(K,c), K^*(c), P^*(c), J^*(c)$ with their obvious meaning.

Now observe that $\frac{c}{2}J(K,c)$ is a monotonically increasing function of c and if $K \in E_K^\circ(c)$ then $K \in E_K(c')$ if $c' < c$. On the other hand Theorem 10.1 immediately implies that

$$\lim_{c' \nearrow c} J(K^*(c'), c') = J(K^*(c), c)\ , \tag{10.39}$$

as long as the inequality $R^* > cH^TH$ holds. Now let $K \in E_K^\circ(c)$, then by Propositions 10.1 and 10.3 $R^* > cHH^T$ hence (10.39) applies. Thus

$$\begin{aligned} \frac{c}{2}J(K,c) &\geq \lim_{c' \nearrow c} \frac{c'}{2}J(K,c') \geq \lim_{c' \nearrow c} \frac{c'}{2}J(K^*(c'),c') \qquad (10.40)\\ &= \frac{c}{2}J(K^*(c),c)\ , \end{aligned}$$

proving that the optimal value is given by $K = K^*$. Uniqueness is a consequence of the following inequality obtained by a more precise analyis. Introduce the notation

$$\Delta J = J(K,c) - J\left(K^*(c),c\right)\ .$$

Then there exists a constant d depending only on the matrices H and R such that if $K \in E_K(c)$ then

$$\|K - K^*(c)\| \leq d \max\left[(\Delta J)^{\frac{1}{2}}, \Delta J\right] \tag{10.41}$$

10.5 Multivariable Systems

Consider the multivariable linear stochastic system given by

$$y = H(\theta^*)e$$

where

$$H(\theta) = I + C(\theta)\left(q^{-1}I - A(\theta)\right)^{-1}B(\theta)\,,$$

is an $m \times m$ square transfer function and q^{-1} denotes the backward shift operator assuming that the following condtions are satisfied:

CONDITION 10.1
$H(\theta)$, $\theta \in D \subset \mathbb{R}^p$, stable, inverse-stable, $A(\theta), B(\theta), C(\theta)$ twice continuously differentiable; □

CONDITION 10.2
(e_n) is an i.i.d. sequence with $E(e_n) = 0$, and $E\left(e_n e_n^T\right) = \Lambda^* > 0$. □

The set of symmetric, positive definite $m \times m$ matrices is denoted by D_Λ. Define for $\theta \in D$

$$\bar{\varepsilon}(\theta) = H(\theta)^{-1}y\,.$$

Then the asymptotic cost function is defined for $\theta \in D$, $\Lambda \in D_\Lambda$ by

$$W(\theta, \Lambda) = \frac{1}{2}E\left(\bar{\varepsilon}_n^T(\theta)\Lambda^{-1}\bar{\varepsilon}_n(\theta)\right) + \frac{1}{2}\log\det\Lambda\,.$$

Note that if (e_n) is an i.i.d. sequence of Gaussian random vectors with distribution $N(0, \Lambda^*)$, then $W(\theta, \Lambda)$ is the asymptotic negative log-likelihood function, except for an additive constant.

The gradient of $W(\theta, \Lambda)$ with respect to θ and Λ^{-1} is given by

$$\begin{aligned} W_\theta(\theta, \Lambda) &= \mathrm{E}\ \bar{\varepsilon}_{\theta n}^T(\theta)\Lambda^{-1}\bar{\varepsilon}_n(\theta) \\ W_{\Lambda^{-1}}(\theta, \Lambda) &= \frac{1}{2}\ (\mathrm{E}\ \bar{\varepsilon}_n(\theta)\bar{\varepsilon}_n^T(\theta) - \Lambda). \end{aligned}$$

Here the gradients of the components of $\bar{\varepsilon}_n(\theta)$ are represented as column-vectors.

Define

$$R_1^* = W_{\theta,\theta}(\theta^*, \Lambda^*)\,.$$

Then the Hessian of $W(\theta, \Lambda)$ at (θ^*, Λ^*) is

$$R^* = \begin{pmatrix} R_1^* & 0 \\ 0 & (\Lambda^*)^{-1} \otimes (\Lambda^*)^{-1} \end{pmatrix}$$

CONDITION 10.3
We assume that for any fixed $\Lambda > 0$ the equation

$$W_\theta(\theta, \Lambda) = 0$$

has a unique solution $\theta = \theta^*$, and

$$R_1^* = W_{\theta,\theta}(\theta^*, \Lambda^*) > 0 \, .$$

□

A weighted recursive prediction estimation of θ^* and Λ^* is obtained as follows. First define the correction terms

$$\begin{aligned} H_{1n}(\theta, \Lambda^{-1}) &= \bar{\varepsilon}_{\theta n}^T(\theta)\Lambda^{-1}\bar{\varepsilon}_n(\theta) \\ H_{2n}(\theta, \Lambda^{-1}) &= \frac{1}{2}(\, \bar{\varepsilon}_n(\theta)\bar{\varepsilon}_n^T(\theta) - \Lambda_{n-1}) \end{aligned}$$

and then consider the recursion

$$\begin{bmatrix} \theta_n \\ \Lambda_n^{-1} \end{bmatrix} = \begin{bmatrix} \theta_{n-1} \\ \Lambda_{n-1}^{-1} \end{bmatrix} - \frac{1}{n} K \begin{bmatrix} H_{1n}(\theta, \Lambda^{-1}) \\ H_{2n}(\theta, \Lambda^{-1}) \end{bmatrix} \, .$$

It is easily seen that at $\theta = \theta^*, \Lambda = \Lambda^*$ the sample covariance-matrix of the process (H_n) is given by

$$S^* = \begin{pmatrix} R_1^* & 0 \\ 0 & \frac{1}{4}\,\mathrm{E}\,(ee^T \otimes ee^T - \Lambda \otimes \Lambda) \end{pmatrix} \, .$$

The recursion above is a frozen-parameter recursion from which a genuinely rcursive estimation is obtained in a standard way. Omitting technical details and conditions we note that the shifts of the transformed error process

$$\psi_t = \left(e^{t/2} \left(\hat{\theta}_{e^t} - \theta^* \right), e^{t/2} \left(\hat{\Lambda}_{e^t}^{-1} - (\Lambda^*)^{-1} \right) \right)$$

converge weakly to the diffusion process defined by

$$d\tilde{x}(t) = \left(-KR^* + \frac{I}{2} \right) \tilde{x}(t)dt + d\tilde{w}(t) \, , \tag{10.42}$$

where $\tilde{w}(t)$ is a Wiener-process with covariance-matrix KS^*K^T, assuming that

$$-KR^* + \frac{I}{2}$$

asymptotically stable.

Equation (10.42) looks very much the same as equation (10.4) for the single variable case, but there is a major difference: the covariance matrix of $d\tilde{w}(t)$ is KS^*K^T instead of KR^*K^T and generically $S^* \neq R^*$.

A risk sensitive identification criterion is given by the following functional:

$$J(K) = \frac{2}{c} \lim_{T\to\infty} \frac{1}{T} \log \mathrm{E} \left(\exp \left\{ \frac{c}{2} \int_0^T \|H\tilde{x}(t)\|^2 \, dt \right\} \right) ,$$

where $c > 0$, H is nonsingular. Note that this is mixed criterion in the sense that $J(K)$ is effected by $\hat{\theta}_n - \theta^*$ and $\left(\hat{\Lambda}_n\right)^{-1} - (\Lambda^*)^{-1}$. Similarly to the one-dimensional case set

$$\mathcal{G}(s) = H \left(sI - \left(-KR^* + \frac{I}{2} \right) \right)^{-1} \left(KS^*K^T \right)^{\frac{1}{2}} .$$

PROPOSITION 10.7
The inequality $S^* > cS^*R^{*-1}H^TH(R^*)^{-1}S^*$ holds if and only if there exists a matrix K, for which $-KR^* + \frac{I}{2}$ is asymptotically stable and $\|c^{1/2}\mathcal{G}\|_{H^\infty} < 1$. □

Note that for $S^* = R^*$ the latter inequality reduces to

$$R^* > cH^TH$$

which case has already been established in Proposition 10.3.

Finally, let

$$E_K = \left\{ K \mid K \in D_K, \|c^{\frac{1}{2}}\mathcal{G}\|_\infty < 1 \right\} .$$

THEOREM 10.4
The risk sensitive criterion $J(K)$ is minimized over E_K at the unique point

$$K^* = \left(R^* - cS^*(R^*)^{-1}H^TH \right)^{-1} ,$$

and

$$J(K^*) = \mathrm{tr} \left(R^* \left(S^*\right)^{-1} R^* - cH^TH \right)^{-1} H^TH .$$

□

Note that if H block diagonal, say

$$H = \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix} ,$$

where H_1 and H_2 are $p \times p$ and $m \times m$ matrices, respectively, then K^* is also block-diagonal and in particular

$$K_1^* = \left(R_1^* - cH_1^TH_1 \right)^{-1} ,$$

a familiar expression. It is easy to see that if the problem of minimizing $J(K)$ is considered on the set of block-diagonal matrices K, say

$$K = \begin{pmatrix} K_1 & 0 \\ 0 & K_2 \end{pmatrix}$$

then we can write

$$J(K) = J_1(K_1) + J_2(K_2) \,,$$

thus the minimization of $J(K)$ can be reduced to the separate minimizations with respect to K_1 and K_2. But the fact that the optimal K in the general minimization problem is block-diagonal can not be easily shown directly.

Acknowledgments

The first author expresses his thanks to Jan van Schuppen for introducing him to LEQG-control and valuable discussions.

This research was supported in part by grants from the Swedish Research Council for Engineering Sciences (TFR), the Göran Gustafsson Foundation, the National Research Foundation of Hungary (OTKA) under Grants T015668, T16665, T 020984 and T032932.

10.6 Bibliography

[1] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*. Springer-Verlag, Berlin, 1990.

[2] A. Bensoussan and J.H. van Schuppen. Optimal control of partially observable stochastic systems with an exponential-of-integral performance index. *SIAM J. Control Optim.*, 23:599–613, 1985.

[3] L. Gerencsér. Rate of convergence of recursive estimators. *SIAM J. Control and Optimization*, 30(5):1200–1227, 1992.

[4] L. Gerencsér. A representation theorem for the error of recursive estimators. In *Proc. of the 31st IEEE Conference on Decision and Control, Tucson*, pages 2251–2256, 1992.

[5] L. Gerencsér. A representation theorem for the error of recursive estimators. *Submitted to SIAM J. Control and Optimization*, 2000.

[6] D. Mustafa and K. Glover, *Minimum Entropy H_∞ Control*. Lecture Notes in Control and Information Sciences, Vol. 146. Springer Verlag, 1990.

[7] T. Söderström and P. Stoica. *System identification*. Prentice Hall, 1989.

[8] A. Stoorvogel and J.H. van Schuppen. *System identification with information theoretic criteria*. Report BS-R9513, CWI, Amsterdam, 1995.

[9] J. C. Willems, Least squares stationary optimal control and the algebraic Riccati equation, *IEEE Transactions of Automat. Contr.* 18(6):621–634, 1971.

11

Input Tracking and Output Fusion for Linear Systems

Xiaoming Hu *Ulf Jönsson* *Clyde F. Martin*

Abstract

In this paper, the input-output behavior of a linear stable system is studied from a geometric point of view. Based on these results, it is discussed how to choose an output and how to fuse a number of outputs, to best track the input in stationarity.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 159-172, 2003.
© Springer-Verlag Berlin Heidelberg 2003

11.1 Introduction

In this paper we first review in some detail how the input of a linear stable system is tracked by the output. Our motivation to further study this classical problem lies in that the results we obtain can be applied to other important problems such as optimal input tracking and sensor fusion. It is natural that before we can develop a procedure to choose an output or a combination of outputs (sensors) that best tracks an input (in the case that the input is at least partially unknown), we need to understand when a given output tracks the input.

In the second part of the paper, we will discuss a sensor fusion problem. There has been a vast literature on sensor fusion, see for example, the papers in [1] and the references therein. However, treatment of the problem from the input tracking point of view has to our knowledge not been addressed.

This paper is organized as follows. In section 2, we discuss the problem of how a given output tracks an input in stationarity. In sections 3 and 4, we discuss the problem of how to choose an output, or a combination of sensors, to optimally track an input in stationarity. Finally, we use an example to illustrate our methods.

11.2 Autonomous Linear Systems

In this section we review some classical results on asymptotic input tracking. Consider a stable, controllable and observable SISO linear system:

$$\begin{aligned} \dot{x} &= Ax + bu \\ y &= cx \end{aligned} \tag{11.1}$$

where $x \in R^n$ and $\sigma(A) \in C^-$.

We will consider the case when the input u is generated by the following exogenous system:

$$\begin{aligned} \dot{w} &= \Gamma w \\ u &= qw \end{aligned} \tag{11.2}$$

where $w \in R^m$ and $\sigma(\Gamma) \in \bar{C}^+$. This exo-system can generally have a block diagonal Jordan realization

$$\begin{aligned} q &= \begin{pmatrix} q_1 & q_2 & \dots q_M \end{pmatrix} \\ \Gamma &= \mathrm{diag}(\Gamma_1, \Gamma_2, \dots, \Gamma_M) \end{aligned} \tag{11.3}$$

where each $q_m = \begin{pmatrix} 1 & 0 \dots 0 \end{pmatrix}$ is a first unit vector of length dim (Γ_m) and each Jordan block corresponds either to polynomial, exponential, or sinusoidal functions. The output of the exo-system becomes

$$u(t) = \sum_{m=1}^{M} q_m e^{\Gamma_m t} w_{0_m}.$$

Such exo-systems can generate, for example, step functions, ramp functions, polynomials, exponentials, sinusoidals, and combinations of such functions.

PROPOSITION 11.2.1
Suppose A is a stable matrix, then all trajectories of $(x(t), w(t))$ tend asymptotically to the invariant subspace $S := \{(x, w) : \ x = \Pi w\}$, where Π is the solution of

$$A\Pi - \Pi\Gamma = -bq.$$

On the invariant subspace, we have

$$y(t) = c\Pi w(t).$$

□

The proof of this proposition can be found, for example, in [5], and there is also a vast literature on the nonlinear case [3, 6].

Using the matrix Π we have that the output of the linear system in the steady-state can be represented as

$$y = c\Pi w.$$

PROPOSITION 11.2.2
Let the system

$$\dot{w} = \Gamma w, \quad u = qw$$

be observable and no eigenvalue of Γ is a transmission zero of (11.1). Then the system on the invariant subspace

$$\dot{w} = \Gamma w, \quad y = c\Pi w$$

is also observable.

□

Proof: A similar proof can be found in [5]. Since this result will be used several times later on, we give a full proof here.

We first need to establish that under the hypotheses, the composite system

$$\begin{array}{rcl} \begin{pmatrix} \dot{x} \\ \dot{w} \end{pmatrix} & = & \begin{pmatrix} A & bq \\ 0 & \Gamma \end{pmatrix} \begin{pmatrix} x \\ w \end{pmatrix} \\ y & = & cx \end{array} \tag{11.4}$$

is observable. Methods for proving similar results can be found, for example, in [2].

Define

$$H(s) = \begin{pmatrix} sI - A & -bq \\ 0 & sI - \Gamma \\ c & 0 \end{pmatrix}.$$

By Hautus test we know that the system is observable if and only if

$$rank(H(s)) = n + m \quad \forall s.$$

If s is not an eigenvalue of Γ, it is easy to see that $rank(H(s)) = n + m$ since (c, A) is observable. Now suppose s is an eigenvalue of Γ,

$$H(s) = \begin{pmatrix} sI - A & b & 0 \\ 0 & 0 & I_m \\ c & 0 & 0 \end{pmatrix} \begin{pmatrix} I_n & 0 \\ 0 & -q \\ 0 & sI - \Gamma \end{pmatrix}.$$

If s is not a transmission zero of (11.1), then and only then the first matrix on the right-hand side has rank $n + 1 + m$. The second has rank $n + m$ since (q, Γ) is observable. By Sylvester's inequality, we have

$$rank(H(s)) \geq n + 1 + m + n + m - (n + m + 1) = n + m.$$

Therefore $rank(H(s)) = n + m$.

Now we do a coordinate change $\bar{x} = x - \Pi w$. Then (11.4) becomes

$$\begin{pmatrix} \dot{\bar{x}} \\ \dot{w} \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & \Gamma \end{pmatrix} \begin{pmatrix} \bar{x} \\ w \end{pmatrix}$$
$$y = c\bar{x} + c\Pi w$$

It is straight forward to see that

$$((c, c\Pi), \begin{pmatrix} A & 0 \\ 0 & \Gamma \end{pmatrix})$$

is observable implies $(c\Pi, \Gamma)$ is so too. Q.E.D.

An observer for the input Based on this result it follows that the input u can be reconstructed by the observer

$$\begin{aligned} \dot{\hat{w}} &= \Gamma\hat{w} + L(y - c\Pi\hat{w}) \\ \hat{u} &= q\hat{w} \end{aligned} \tag{11.5}$$

where the vector L can be designed such that the eigenvalues of $\Gamma - Lc\Pi$ can be placed anywhere we desire in the complex plane. This is a reduced order observer since the dynamics corresponding to A is not included in the observer equation. One of the prices we pay for this is that the convergence of the observer is restricted by the transients corresponding to the eigenvalues of A. Indeed, if $\bar{x} = x - \Pi w$, $\bar{w} = w - \hat{w}$, and $\bar{u} = q(w - \hat{w})$ then the error dynamics becomes

$$\begin{pmatrix} \dot{\bar{x}} \\ \dot{\bar{w}} \end{pmatrix} = \begin{pmatrix} A & 0 \\ -Lc & \Gamma - Lc\Pi \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{w} \end{pmatrix}$$
$$\bar{u} = q\bar{w}$$

We will in the next section see that under special conditions it is possible to design the output such that the input is reconstructed in stationarity. This will give a memoryless observer, which is preferable from a computational point of view.

11.3 Exact Input Tracking

We will now discuss how Proposition 11.2.1 can be used to determine an appropriate output in order to track the input exactly in stationarity. It follows obviously that the output tracks the input if the vector c is chosen such that

$$(c\Pi - q)e^{\Gamma t}w_0 = 0 \tag{11.6}$$

where w_0 is the initial state of (11.2) that generates the input. This is clearly the case if $c\Pi = q$. If Π, for example, has full column rank, then it is possible to design an output c for perfect input tracking in stationarity. We will show below that this is the case if (A, b) is controllable, (q, Γ) is observable and dim $(A) \geq$ dim (Γ).

In some way one may view this problem as a dual one to the output regulation problem discussed in [4]. In this section, we discuss some necessary and sufficient conditions.

THEOREM 11.3.1
Suppose (q, Γ) is observable and (A, b) controllable. Then a necessary and sufficient condition for the existence of a c, such that $c\Pi = q$, is that the dimension of A is greater or equal to that of Γ. □

Proof: We can rewrite $\dot{x} = Ax + bu$ in the canonical form:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ &\vdots \\ \dot{x}_{n-1} &= x_n \\ \dot{x}_n &= -\sum_{i=1}^{n} a_i x_i + ku, \end{aligned} \tag{11.7}$$

where $k \neq 0$ and $\rho(s) = s^n + \sum_{i=1}^{n} a_i s^{i-1}$ is Hurwitz. In the steady state, by Proposition 11.2.1 we have

$$x_1 = \pi_1 w,$$

where π_1 is the first row of Π. Since $x_i = \pi_1 \Gamma^{i-1} w$, for $i = 1, ..., n$, we have

$$\Pi = \begin{pmatrix} \pi_1 \\ \pi_1\Gamma \\ \vdots \\ \pi_1\Gamma^{n-1} \end{pmatrix}. \tag{11.8}$$

Thus,

$$\pi_1\Gamma^n = -\sum_{i=1}^{n} a_i \pi_1 \Gamma^{i-1} + kq.$$

Since by assumption Γ does not have any eigenvalue in the open left-half plane, we have

$$\pi_1 = kq\rho(\Gamma)^{-1}. \tag{11.9}$$

If there exists a c, such that

$$q = c\Pi = \sum_{i=1}^{n} c_i \pi_1 \Gamma^{i-1},$$

then

$$q = kq\rho(\Gamma)^{-1} \sum_{i=1}^{n} c_i \Gamma^{i-1},$$

or,

$$q(I - k\rho(\Gamma)^{-1} \sum_{i=1}^{n} c_i \Gamma^{i-1}) = 0.$$

Denote $\Delta = I - k\rho(\Gamma)^{-1} \sum_{i=1}^{n} c_i \Gamma^{i-1}$. It is easy to show that

$$\Delta = (\Gamma^n + \sum_{i=1}^{n} (a_i - kc_i)\Gamma^{i-1})\rho(\Gamma)^{-1}$$

thus $q\Delta = 0$ if and only if

$$q\Gamma^n + \sum_{i=1}^{n} (a_i - kc_i) q\Gamma^{i-1} = 0. \tag{11.10}$$

Since (q, Γ) is observable, (11.10) has a solution if and only if n is greater or equal to the dimension of Γ. Q.E.D.

COROLLARY 11.3.1
If $dim(A) \geq dim(\Gamma)$, then Π has full column rank, and thus there exists a c, such that,

$$c\Pi = q.$$

Moreover, if $dim(A) = dim(\Gamma)$, then such c is unique. □

Proof: It follows from (11.8), (11.9), (11.10), and observability of the exo-system (11.2). Indeed,

$$\Pi = \begin{pmatrix} \pi_1 \\ \pi_1\Gamma \\ \vdots \\ \pi_1\Gamma^{n-1} \end{pmatrix} = k \begin{pmatrix} q \\ q\Gamma \\ \vdots \\ q\Gamma^{n-1} \end{pmatrix} \rho(\Gamma)^{-1}$$

which has full rank since (q, Γ) is observable. Q.E.D.

COROLLARY 11.3.2
Suppose $dim(A) = n \geq dim(\Gamma) = m$, then there exists a c such that $c\Pi = q$ and the resulting system (11.4) is observable and (11.1) does not have any transmission zero that is also an eigenvalue of Γ. □

Proof: Consider the canonical form (11.7). Suppose the characteristic polynomial for Γ is $\rho_\Gamma(s) = s^m + \sum_{i=1}^{m} \gamma_i s^{i-1}$. It follows from (11.10) and Cayley-Hamilton that

$$c_i = \frac{1}{k}(a_i - \bar{\gamma}_i) \quad i = 1, ..., n,$$

where $\bar{\gamma}_i = 0 \ \forall i < n - m + 1$ and $\bar{\gamma}_i = \gamma_{i-n+m}$ otherwise, is a solution such that $c\Pi = q$. It then follows from the fact that A and Γ do not share any eigenvalue, that no eigenvalue s_0 of A or Γ is a root of

$$\sum_{i=1}^{n} c_i s_0^{i-1} = \frac{1}{k} \sum_{i=1}^{n} a_i s_0^{i-1} - \frac{s_0^{n-m}}{k} \sum_{i=1}^{m} \gamma_i s_0^{i-1}.$$

Indeed, if s_0 is for example a root of the characteristic polynomial of A, the above expression reduces to

$$-\frac{s_0^{n-m}}{k} \rho_\Gamma(s_0),$$

which must be nonzero. Thus, no transmission zero of the corresponding (11.1) is an eigenvalue of Γ and the pair (c, A) is observable. From the proof of Proposition 11.2.2 we derive that (11.4) is observable. Q.E.D.

We have shown under the assumptions of Corollary 11.3.1 that the input u can be reconstructed simply as

$$\hat{u} = cx,$$

where, e.g. $c = q\Pi^\dagger$, where $\Pi^\dagger$ is the pseudo inverse. The tracking error satisfies (where $\bar{x} = x - \Pi w$)

$$\bar{u} = u - \hat{u} = cx - qw = c\bar{x} + c\Pi w - qw = c\bar{x}$$

Hence, the error dynamics in this case becomes

$$\dot{\bar{x}} = A\bar{x}$$
$$\bar{u} = c\bar{x}$$

which, as for the observer in (11.5), has its rate of convergence limited only by the eigenvalues of A.

The exo-system will in many applications have significantly larger dimension than the linear system (11.1) and then there only could exist a solution to (11.6) for special choices of initial condition w_0 of the exo-system. In the next section we discuss a strategy for fusing the output of a number of outputs in order to minimize the steady state tracking error.

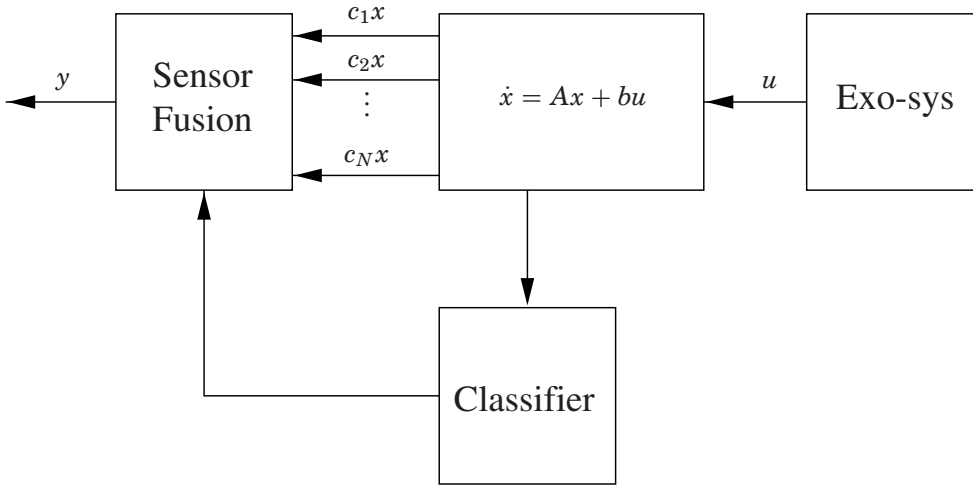


Figure 11.1 Sensor fusion set-up.

11.4 Output Fusion for Input Tracking

We will here consider a special sensor fusion problem where we try to minimize the tracking error by appropriately combining the outputs of a number of sensors. A sensor in our terminology means a particular choice of c_k matrix. If the state space model represents physical variables, then typically each c_k corresponds to one state variable. The idea is that the sensor fusion block should determine a linear combination of the sensor signals such that the output

$$y = \sum_{k=1}^{N} \alpha_k c_k x \tag{11.11}$$

tracks the input u in stationarity. We will here discuss how this sensor fusion idea works for the case when the input is generated by an observable exo-system of the form (11.2).

The Classifier Block

We first need a classifier block in order to determine what Jordan blocks are active in the generation of the input u, i.e. it determines a set $\mathcal{M} \subset \{1, \ldots, M\}$ of indices such that the input can be represented as

$$u(t) = \sum_{m \in \mathcal{M}} q_m e^{\Gamma_m t} w_{0_m}.$$

We will see below that this information sometimes is enough to obtain perfect tracking. However, it is generally important to use as much information on the vector w_0 as possible in order to obtain better tracking. For example, if we in addition to $\mathcal{M}$ also obtain an estimate $\hat{w}_{0_{\mathcal{M}}} = [\hat{w}_{0_{m_1}}, \ldots, \hat{w}_{0_{m_n}}]$ of the initial condition then our ability to reconstruct the input improves. Even qualitative information such as the relative amplitude of the various blocks is useful.

The Sensor Fusion Block

This block takes as input the classification $\mathcal{M}$ and maps it to a vector α that minimizes the steady state tracking error for the output (11.11) according to some cost criterion. We will discuss this in more detail below where we also give necessary and sufficient conditions for obtaining perfect tracking. In more sophisticated schemes we may also use an estimate $\hat{w}_{0_{\mathcal{M}}}$ of the initial condition of the exo-system. This may give better tracking, however at the price of more complex classifier and sensor fusion blocks. Note that this scheme will be independent of the state space realization and the convergence to the steady state solution depends on the spectrum of A.

A main practical motivation for our sensor fusion scheme is due to the limited communication and computation resources in many embedded systems, such as mobile robotic systems. There is a need to develop "cheap" sensing algorithms. The central idea in our scheme is to optimally combine the existing sensors for state variables to measure the external signals. This optimization can be done off-line and then the sensor fusion block only needs to use a table look-up to decide the parameter vector α. The only remaining issue is how to design the online classifier.

The most natural way from a systems point of view is perhaps to use a dynamical observer (in discrete time) to identify qualitatively the initial condition (or the active Γ blocks) and then shut it down. This is possible since the state of the exo-system is observable from the output from any sensor that satisfies the conditions of Proposition 11.2.2. However, this approach could be computationally expensive even if we only run it once in a while. A more refined scheme is discussed for a special case below.

For many practical systems, it is perhaps more realistic to design the classifier based on other sensors that sense the interaction of the system with the environment (such as laser scanners and video cameras), or/and on the nature of application the system is operated for. In this way, typically only a range of the Γ blocks (such as a frequency range) can be identified.

An Example Classifier

Let us consider the case when only one Jordan block of Γ is active at a time. Then we can construct a classifier consisting of a bank of discrete time observers. Each observer in the classifier corresponds to a particular Jordan block. The idea is to initially activate all observers in order to decide which block is active. Once this is done, all observers are shut down except the one corresponding to the active Jordan block. This observer is used to detect changes in the exo-system. By running the monitor on low sampling frequency a minimum computational effort is spent in the classifier.

Let us discuss how to construct such an observer for block Γ_m. Zero order hold sampling of the dynamics gives

$$\begin{pmatrix} x_{k+1} \\ w_{k+1} \end{pmatrix} = \begin{pmatrix} A_d & b_{m,d}q_{m,d} \\ 0 & \Gamma_{m,d} \end{pmatrix} \begin{pmatrix} x_k \\ w_k \end{pmatrix}$$

where h is the sampling time, $x_k = x(kh)$, $w_k = w(kh)$, $A_d = e^{Ah}$, $\Gamma_{m,d} = e^{\Gamma_m h}$,

and $b_{m,d}, q_{m,d}$ are matrices of full rank such that

$$b_{m,d}q_{m,d} = \int_0^h e^{A(h-s)}bqe^{\Gamma_m s}ds.$$

We can use exactly the same arguments as in the proof of Proposition 11.2.1 to show that the functions (x_k, w_k) converges to an invariant subspace $S_m = \{(x,w) : x = \Pi_m w\}$. It should be intuitively clear that the invariant subspace is the same as in continuous time. The following proposition gives a formal mathematical proof.

PROPOSITION 11.4.1
The sampled functions converge to the invariant subspace $S_m = \{(x,w) : x = \Pi_m w\}$, where

$$A\Pi_m - \Pi_m\Gamma_m = bq_m$$

□

Proof: Let $u_k = qw_k$ and assume $x_k = \Pi_m w_k$. It is straight forward to derive that $x_k = \Pi_m w_k$ is invariant if and only if

$$A_d\Pi_m - \Pi_m\Gamma_m = b_{m,d}q_{m,d}.$$

This can be written

$$e^{Ah}\Pi_m - \Pi_m e^{\Gamma_m h} = -\int_0^h e^{A(h-s)}bq_m e^{\Gamma_m s}ds$$

$$\Leftrightarrow \quad e^{Ah}\Pi_m e^{-\Gamma_m h} - \Pi_m = -\int_0^h e^{A(h-s)}bq_m e^{-\Gamma_m(h-s)}ds$$

This is a discrete time Lyapunov equation so we get

$$\begin{aligned}\Pi_m &= \sum_{i=0}^{\infty} e^{Aih}\int_0^h e^{A(h-s)}bq_m e^{-\Gamma_m(h-s)}ds e^{-\Gamma_m ih} \\ &= \sum_{i=0}^{\infty}\int_{ih}^{(i+1)h} e^{A\tau}bq_m e^{-\Gamma_m\tau}d\tau = \int_0^{\infty} e^{A\tau}bq_m e^{-\Gamma_m\tau}d\tau\end{aligned}$$

which is the solution of the Lyapunov equation $A\Pi_m - \Pi_m\Gamma_m = bq_m$. The convergence to the subspace can be proven similarly to that in the proof of Proposition 11.2.1.
Q.E.D.

We will next consider observability of the pairs $(c\Pi_m, \Gamma_{md})$ on the invariant subspace $S_m = \{(x,w) : x = \Pi_m w\}$. It is well known that the sampled system $(c\Pi_m, \Gamma_{md})$ is observable if and only if the continuous time system $(c\Pi_m, \Gamma_m)$ is observable. Hence, from Proposition 11.2.2 we know that a sufficient condition is that the exo-system (q_m, Γ_m) is observable and that no eigenvalue of Γ_m is a transmission zero of the system (11.1). The next proposition shows that it is possible to design one sensor that works for all Jordan blocks.

PROPOSITION 11.4.2
Suppose all pairs (q_m, Γ_m) are observable. Then there exists a sensor combination, c, such that all pairs $(c\Pi_{md}, \Gamma_{md})$ are observable on the corresponding invariant subspace. □

Proof: From the above discussion and Proposition 11.2.2 it follows that it is enough to find c such that (c, A) is observable and (11.1) does not have any transmission zero at the eigenvalues of the Γ_m. Such a c is always possible to design since only a finite number of constraints are imposed. Q.E.D.

Assume now that we have sensors c_o, which corresponds to the output z, such that the pairs $(c_o\Pi_m, \Gamma_{md})$ are observable. We can then use the following block of observers

$$\hat{w}_{k+1} = \Gamma_{m,d}\hat{w}_k + L_m(z_k - c_o\Pi_m\hat{w}_k)$$

where the observer gains L_m are designed such that the eigenvalues of $\Gamma_{m,d} - L_m c_o \Pi_m$ are stable. The observer that corresponds to the active Jordan block has error dynamics

$$\begin{pmatrix} \bar{x}_{k+1} \\ \bar{w}_{k+1} \end{pmatrix} = \begin{pmatrix} A_d & 0 \\ -L_m c_o & \Gamma_{m,d} + L_m c_o \Pi_m \end{pmatrix} \begin{pmatrix} \bar{x}_k \\ \bar{w}_k \end{pmatrix}$$

which convergences to zero. By proper design of L_m the convergence rate is determined by the eigenvalues of A_d. All other observers have the error dynamics

$$\bar{w}_{k+1} = (\Gamma_{m,d} - L_m c_o \Pi_m)\bar{w}_k + (\Gamma_{n,d} - \Gamma_{m,d} - L_m c_o(\Pi_n - \Pi_m))w_k$$

The matrix $\Lambda = \Gamma_{n,d} - \Gamma_{m,d} - L_m c_o(\Pi_n - \Pi_m))$ can without loss of generality be assumed to be nonzero, which implies that $\bar{w}_k$ does not converge to zero. To see that Λ generically is nonzero, let us suppose $\Lambda = 0$. This implies

$$(\Gamma_{m,d} - L_m c_o \Pi_m)e^{-\Gamma_n h} - I = -L_m c_o \Pi_n e^{-\Gamma_n h}$$

On the other hand,

$$(\Gamma_{m,d} - L_m c_o \Pi_m)\Psi e^{-\Gamma_n h} - \Psi = -L_m c_o \Pi_n e^{-\Gamma_n h}$$

has the unique solution

$$\Psi = \sum_{i=0}^{\infty} (\Gamma_{m,d} - L_m c_o \Pi_m)^i L_m c_o \Pi_n e^{-\Gamma_n h(i+1)}.$$

Generically, we have $\Psi \neq I$, thus we can design L_m such that $\Lambda \neq 0$.

In the decision logic block we simply need to compare the magnitude of the error signals $\bar{w}_m$ from observers. Under our hypothesis that only one Jordan block at a time is active it follows that only one error signal will converge to zero.

Once it is decided which Jordan block is active the classifier goes into the monitoring mode, where the observer corresponding to the active Jordan block

is used to detect a change in the input signals. The sampling frequency in the monitoring mode can often be chosen significantly lower than in the detection mode. Suppose, for example, that the Jordan blocks correspond to sinusoidals with distinct frequencies. Then by choosing the sampling frequency such that these frequencies fold into distinct locations, we can detect a change of frequency with a significantly lower sampling rate than the Nyquist frequency.

Perfect Steady-state Tracking

Assume we are given K sensor combinations $c_1, \ldots, c_K$. We derive a sufficient (and in a sense also necessary) condition for obtaining perfect steady state tracking using these sensors. We have the following result.

PROPOSITION 11.4.3
Suppose $\mathcal{M} = \{m_1, \ldots, m_n\} \subset \{1, \ldots, M\}$ are the active Jordan blocks. Then we can obtain perfect tracking if

$$q_{\mathcal{M}}^T \in \mathrm{Im}(\Pi_{\mathcal{M}}^T C^T)$$

where $A\Pi_{\mathcal{M}} - \Pi_{\mathcal{M}}\Gamma_{\mathcal{M}} = -bq_{\mathcal{M}}$ and

$$\begin{aligned} \Gamma_{\mathcal{M}} &= \mathrm{diag}(\Gamma_{m_1}, \ldots, \Gamma_{m_n}) \\ q_{\mathcal{M}} &= (q_{m_1}, \ldots, q_{m_n}) \\ C^T &= [\, c_1^T \quad \ldots \quad c_K^T \,] \end{aligned} \tag{11.12}$$

□

Proof: The steady state output will be $y = \alpha C\Pi_{\mathcal{M}} w_{\mathcal{M}}(t)$, where $w_{\mathcal{M}}(t) = e^{\Gamma_{\mathcal{M}} t} w_{0_{\mathcal{M}}}$. Hence, we obtain perfect tracking since our assumption implies that there exists a solution to $\alpha C\Pi_{\mathcal{M}} = q_{\mathcal{M}}$. Note that this condition is necessary if $w_{0_{\mathcal{M}}}$ is allowed to take any value. Q.E.D.

If the condition of the proposition holds then we normally want to find the vector α with the minimum nonzero coefficients such that

$$\alpha C\Pi_{\mathcal{M}} = q_{\mathcal{M}}$$

in order to minimize the number of sensors used. This can be done off-line and then the sensor fusion block only need to use a table look-up to decide the vector α.

Approximate Tracking

It will often happen that we have too few sensors or too poor knowledge of the exo-system to obtain perfect tracking. In such cases we need to optimize the sensor fusion in order to get best tracking in some average sense.

Suppose $\Gamma(\delta),\ q(\delta)$ is an uncertain parameterization of the exo-system, where $\delta \subset \Delta$ is the uncertain parameters. We let $\Delta = \{0\}$ correspond to the case when we have perfect knowledge of the exo-system. If $\Delta \neq \{0\}$ or if the condition in

Proposition 11.4.3 does not hold then we let the sensor fusion be determined by the solution to some optimization problem

$$\min_{\alpha} \mathcal{L}_{\Delta}(|\alpha C\Pi(\delta) - q(\delta)|)$$

where $\mathcal{L}_{\Delta}$ is some functional over Δ and $\Pi(\delta)$ is the solution to $A\Pi(\Gamma(\delta)) - \Pi(\delta)\Gamma(\delta) = -bq(\delta)$. Some examples are given in [5].

An Example of optimal output design

As an example to demonstrate our methodology, we consider a car-like base and a robot arm mounted on it. By using the homogeneous representation of rigid body motions, we can easily compute the position of the end-effector, relative to the base, r_A^B, and thus the kinematic model as

$$\dot{x}_A = f(x_A, u).$$

Under the assumptions that the velocity of the car is constant, and that the side slip angle is small, we can get a simplified model of the base vehicle as follows:

$$\dot{x} = v\cos_B(\psi + \beta) \tag{11.13}$$

$$\dot{y} = v\sin_B(\psi + \beta) \tag{11.14}$$

$$\dot{\beta} + r = \frac{f_f + f_r}{mv} = a_{11}\beta + a_{12}r + b_{11}\delta_f \tag{11.15}$$

$$\dot{\psi} = r \tag{11.16}$$

$$\dot{r} = \frac{f_f l_f - f_r l_r}{J} = a_{21}\beta + a_{22}r + b_{21}\delta_f + d(t), \tag{11.17}$$

where (x, y) is the center of gravity of the vehicle, m is the vehicle mass and J is the moment of inertia, and the a and b coefficients depend on the physical parameters of the vehicle.

The disturbance $d(t)$ could be for example, side wind or roughness on the road surface.

This example shows that by measuring the orientation (ψ) and yaw rate (r), it is possible to recover $d(t)$ in some cases, which shall be useful to know for the control of the arm.

The advantage of using the output fusion method, in comparison to the observer method, is that the lateral dynamics of the car is typically very fast. Thus this method converges fast, while not having the transient peakings that might be induced by the pole placements of an observer.

Figure 11.2 shows a simulation where by combining $r(t)$ and $\psi(t)$ we can track a sinusoidal disturbance $d(t)$ in stationarity. The upper diagram shows the convergence of the output and the lower bound shows the error $y - u$.

11.5 Concluding Remarks

We have discussed conditions for steady state input tracking for stable linear systems. It has been shown that these results can be used to devise a sensor fusion scheme for input tracking in some situations. We need to further study the robustness issue of such methods (for example, when there are uncertainties in Γ and/or in the measurements) and ways to generalize them.

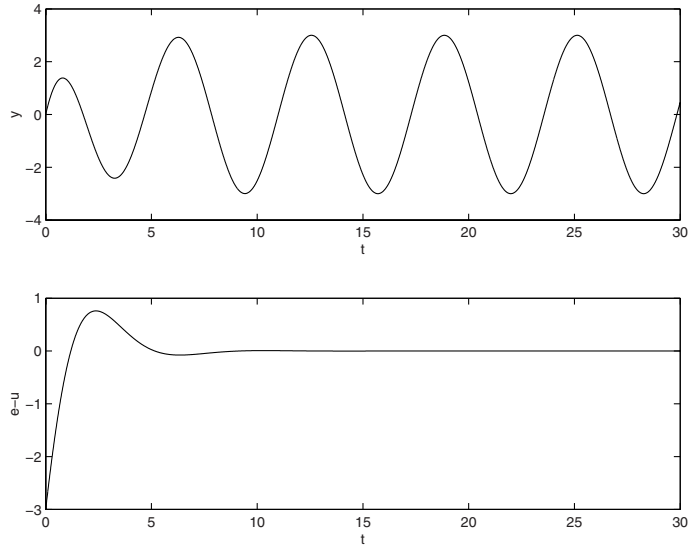


Figure 11.2 Simulation of the sensor fusion.

11.6 References

[1] *Proceedings of the IEEE*, volume 85, Jan 1997. Special issue on sensor fusion.

[2] C.T. Chen. *Linear system theory and design*. HRW, New York, 1984.

[3] E. Coddington and N. Levinson. *Theory of ordinary differential equations.* McGraw-Hill, New York, 1984.

[4] B. Francis. The linear multivariable regulator problem. *SIAM J. Control. Optim.*, 15:486–505, 1977.

[5] X. Hu, U. Jönsson, and C. Martin. Input tracking for stable linear systems. In *Proceedings of the 2002 IFAC Congress*, Barcelona, July 2002.

[6] A. Kelly. The stable, center-stable, center, center-unstable, and unstable manifolds. *J.Diff.Eqns*, 3:546–570, 1967.

12

The Convergence of the Extended Kalman Filter

Arthur J. Krener

Abstract

We demonstrate that the extended Kalman filter converges locally for a broad class of nonlinear systems. If the initial estimation error of the filter is not too large then the error goes to zero exponentially as time goes to infinity. To demonstrate this, we require that the system be C^2 and uniformly observable with bounded second partial derivatives.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 173-182, 2003.
© Springer-Verlag Berlin Heidelberg 2003

12.1 Introduction

The extended Kalman filter is a widely used method for estimating the state $x(t) \in \mathbb{R}^n$ of a partially observed nonlinear dynamical system,

$$\begin{aligned} \dot{x} &= f(x,u) \\ y &= h(x,u) \\ x(0) &= x^0 \end{aligned} \tag{12.1}$$

from the past controls and observations $u(s) \in U \subset \mathbb{R}^m,\ y(s) \in \mathbb{R}^p, 0 \le s \le t$ and some information about the initial condition x^0. The functions f, h are known and assumed to be C^2.

An extended Kalman filter is derived by replacing (12.1) by its linear approximation around the trajectory $\hat{x}(t)$ and adding standard white Gaussian driving noise $w(t) \in \mathbb{R}^l$ and independent, standard white Gaussian observation noise $v(t) \in \mathbb{R}^l$,

$$\begin{aligned} \dot{z} &= f(\hat{x}(t),u(t)) + A(t)z + Gw \\ y &= h(\hat{x}(t),u(t)) + C(t)z + v \\ z(0) &= z^0 \end{aligned} \tag{12.2}$$

where G is a $n \times l$ matrix chosen by the designer,

$$A(t) = \frac{\partial f}{\partial x}(\hat{x}(t),u(t)), \qquad C(t) = \frac{\partial h}{\partial x}(\hat{x}(t),u(t)), \tag{12.3}$$

and z^0 is a Gaussian random vector independent of the noises with mean $\hat{x}^0$ and variance P^0 that are chosen by the designer.

The Kalman filter for (12.2) is

$$\begin{aligned} \dot{\hat{z}}(t) &= f(\hat{x}(t),u(t)) + A(t)\hat{z}(t) + P(t)C'\left(y(t) - h(\hat{x}(t),u(t)) - C(t)\hat{z}(t)\right) \\ \dot{P}(t) &= A(t)P(t) + P(t)A'(t) + \Gamma - P(t)C'(t)C(t)P(t) \\ \hat{z}(0) &= \hat{x}^0 \\ P(0) &= P^0 \end{aligned} \tag{12.4}$$

where $\Gamma = GG'$.

The extended Kalman filter for (12.1) is given by

$$\begin{aligned} \dot{\hat{x}}(t) &= f(\hat{x}(t),u(t)) + P(t)C'(t)\left(y(t) - h(\hat{x}(t),u(t))\right) \\ \dot{P}(t) &= A(t)P(t) + P(t)A'(t) + \Gamma - P(t)C'(t)C(t)P(t) \\ \hat{x}(0) &= \hat{x}^0 \\ P(0) &= P^0. \end{aligned} \tag{12.5}$$

Actually there are many extended Kalman filters for (12.1), depending on the choice of the design parameters $G,\ \hat{x}^0,\ P^0$. We could also broaden the class of extended Kalman filters for (12.1) by allowing $G = G(\hat{x}(t))$ and putting a similar coefficient in front of the observation noise in (12.2). We chose not to do

so to simplify the discussion. For similar reasons we omit the discussion of time varying systems. We expect that our main theorem can be generalized to cover such systems. For more on the derivation of the extended Kalman filter, see Gelb [3].

Baras, Bensoussan and James [1] have shown that under suitable conditions, the extended Kalman filter converges locally, i.e., if the initial error $\tilde{x}(0) = x(0) - \hat{x}(0)$ is sufficiently small then $\tilde{x}(t) = x(t) - \hat{x}(t) \to 0$ as $t \to \infty$. Unfortunately their conditions are difficult to verify and may not be satisfied even by an observable linear system. Krener and Duarte have given a simple example where any extended Kalman filter fails to converge. More on these points later.

By modifying the techniques of [1] and incorporating techniques of the high gain observer of Gauthier, Hammouri and Othman [2] we shall show that under verifiable conditions that the extended Kalman filter converges locally. To state the main result we need a definition.

DEFINITION 12.1
[2] The system

$$\begin{aligned} \dot{\xi} &= f(\xi, u) \\ y &= h(\xi, u) \end{aligned} \tag{12.6}$$

is uniformly observable for any input if there exist coordinates

$$\{x_{ij} : i = 1, \ldots, p,\ j = 1, \ldots, l_i\}$$

where $1 \le l_1 \le \ldots \le l_p$ and $\sum l_i = n$ such that in these coordinates the system takes the form

$$\begin{aligned} y_i &= x_{i1} + h_i(u) \\ \dot{x}_{i1} &= x_{i2} + f_{i1}(\underline{x}_1, u) \\ &\vdots \\ \dot{x}_{ij} &= x_{ij+1} + f_{ij}(\underline{x}_j, u) \\ &\vdots \\ \dot{x}_{il_i-1} &= x_{il_i} + f_{il_i-1}(\underline{x}_{l_i-1}, u) \\ \dot{x}_{il_i} &= f_{il_i}(\underline{x}_{l_i}, u) \end{aligned} \tag{12.7}$$

for $i = 1, \ldots, p$ where $\underline{x}_j$ is defined by

$$\underline{x}_j = (x_{11}, \ldots, x_{1,j\wedge l_1}, x_{21}, \ldots, x_{pj}). \tag{12.8}$$

Notice that in $\underline{x}_j$ the indices range over $i = 1, \ldots, p;\ k = 1, \ldots, \min\{j, l_i\}$ and the coordinates are ordered so that second index moves faster than the first.

We also require that each f_{ij} be Lipschitz continuous, there exists an L such that for all $x, \xi \in \mathbb{R}^n, u \in U$,

$$|f_i(\underline{x}_j, u) - f_i(\underline{\xi}_j, u)| \le L|\underline{x}_j - \underline{\xi}_j|. \tag{12.9}$$

The symbol $|\cdot|$ denotes the Euclidean norm. □

Let

$$\bar{A}_i = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}^{l_i \times l_i} \qquad \bar{A} = \begin{bmatrix} \bar{A}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \bar{A}_p \end{bmatrix}^{n \times n}$$

$$\bar{C}_i = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \end{bmatrix}^{1 \times l_i} \qquad \bar{C} = \begin{bmatrix} \bar{C}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \bar{C}_p \end{bmatrix}^{p \times n}$$

$$\bar{f}_i(x,u) = \begin{bmatrix} f_{i1}(\underline{x}_1, u) \\ \vdots \\ f_{il_i}(\underline{x}_{l_i}, u) \end{bmatrix}^{l_i \times 1} \qquad \bar{f}(x,u) = \begin{bmatrix} \bar{f}_1(x,u) \\ \vdots \\ \bar{f}_p(x,u) \end{bmatrix}^{n \times 1}$$

$$\bar{h}(u) = \begin{bmatrix} h_1(u) \\ \vdots \\ h_p(u) \end{bmatrix}^{p \times 1}$$

then (12.7) becomes

$$\begin{aligned} \dot{x} &= \bar{A}x + \bar{f}(x,u) \\ y &= \bar{C}x + \bar{h}(u) \end{aligned} \tag{12.10}$$

A system such as (12.7) or, equivalently (12.10), is said to be in observable form [4].

We shall also require that the second derivative of $\bar{f}$ is bounded, i.e., for any $x, \xi \in \mathbb{R}^n, u \in U$,

$$\left| \frac{\partial^2 \bar{f}}{\partial x_i \partial x_j}(x,u) \xi_i \xi_j \right| \le L|\xi|^2. \tag{12.11}$$

On the left we employ the convention of summing on repeated indices.

THEOREM 12.1
(Main Theorem) Suppose

- the system (12.1) is uniformly observable for any input and so without loss of generality we can assume that is in the form (12.10) and satisfies the Lipschitz conditions (12.9),
- the second derivative of $\bar{f}$ is bounded (12.11),
- $x(t), y(t)$ are any state and output trajectories generated by (12.10),

- G has been chosen to be invertible,
- $\hat{x}(t)$ and $P(t)$ are a solution of the extended Kalman filter (12.5) where $P(0)$ is positive definite and $\tilde{x}(0) = x(0) - \hat{x}(0)$ is sufficiently small,

Then $|x(t) - \hat{x}(t)| \to 0$ exponentially as $t \to \infty$. □

12.2 Proof of the Main Theorem

We extend the method of proof of [1]. Since the system is in observable form

$$\begin{aligned} A(t) &= \bar{A} + \tilde{A}(t) \\ C(t) &= \bar{C} \end{aligned} \tag{12.12}$$

where

$$\tilde{A}(t) = \frac{\partial \bar{f}}{\partial x}(\hat{x}(t), u(t)),$$

and

$$\frac{\partial \bar{f}_{ir}}{\partial x_{jk}}(\hat{x}(t)) = 0$$

if $k > r$.

First we show that there exists $m_1 > 0$ such that for all $t \geq 0$

$$P(t) \leq m_1 I^{n \times n}. \tag{12.13}$$

Consider the optimal control problem of minimizing

$$\xi'(0)P(0)\xi(0) + \int_0^t \xi'(s)\Gamma\xi(s) + \mu'(s)\mu(s)\, ds$$

subject to

$$\begin{aligned} \dot{\xi}(s) &= -A'(s)\xi(s) - \bar{C}'\mu(s) \\ \xi(t) &= \zeta. \end{aligned}$$

It is well-known that the optimal cost is

$$\zeta' P(t) \zeta$$

where $P(t)$ is the solution of (12.5).

Following [2] for $\theta > 0$ we define $S(\theta)$ as the solution of

$$\bar{A}'S(\theta) + S(\theta)\bar{A} - \bar{C}'\bar{C} = -\theta S(\theta). \tag{12.14}$$

It is not hard to see that $S(\theta)$ is positive definite for $\theta > 0$ as it satisfies the Lyapunov equation

$$\left(-\bar{A} - \frac{\theta}{2} I\right)' S(\theta) + S(\theta)\left(-\bar{A} - \frac{\theta}{2} I\right) = -\bar{C}'\bar{C}$$

where $\bar{C}, (-\bar{A} - \frac{\theta}{2} I)$ is an observable pair and $(-\bar{A} - \frac{\theta}{2} I)$ has all eigenvalues equal to $-\frac{\theta}{2}$. It follows from (12.14) that

$$S_{ij,\rho\sigma}(\theta) = \frac{S_{ij,\rho\sigma}(1)}{\theta^{j+\sigma-1}} = \frac{(-1)^{j+\sigma}}{\theta^{j+\sigma-1}} \begin{pmatrix} j+\sigma-2 \\ j-1 \end{pmatrix}.$$

Let $T(\theta) = S^{-1}(\theta) > 0$ then

$$T_{ij,\rho\sigma}(\theta) = \theta^{j+\sigma-1} T_{ij,\rho\sigma}(1)$$

and $T(\theta)$ satisfies the Riccati equation

$$-\bar{A}T(\theta) - T(\theta)\bar{A}' + T(\theta)\bar{C}'\bar{C}T(\theta) = \theta T(\theta).$$

We apply the suboptimal control $\mu = -\bar{C}T(\theta)\xi$ to the above optimal control problem and conclude that

$$\zeta' P(t) \zeta \leq \xi'(0)P(0)\xi(0) + \int_0^t \xi'(s) \left(\Gamma + T(\theta)\bar{C}'\bar{C}T(\theta)\right) \xi(s)\, ds \quad (12.15)$$

where

$$\begin{aligned} \dot{\xi}(s) &= \left(-A'(s) + \bar{C}'\bar{C}T(\theta)\right) \xi(s) \\ \xi(t) &= \zeta. \end{aligned}$$

Now

$$\begin{aligned} \frac{d}{ds} \xi'(s) T(\theta) \xi(s) &= \xi'(s) \left(\theta T(\theta) + T(\theta)\bar{C}'\bar{C}T(\theta)\right) \xi(s) \\ &\quad -\xi'(s) \left(\tilde{A}(s)T(\theta) + T(\theta)\tilde{A}'(s)\right) \xi(s). \end{aligned}$$

Because of the Lipschitz condition (12.9) we conclude that

$$|A(s)| \leq L$$

and

$$|\tilde{A}(s)| \leq L + |\bar{A}|.$$

From the form of $\tilde{A}(s)$ and $T(\theta)$ we conclude that

$$\left(\tilde{A}(s)T(\theta)\right)_{ij,\rho\sigma} = O(\theta)^{j+\sigma-1}$$

while on the other hand

$$\theta T_{ij,\rho\sigma}(\theta) = \theta^{j+\sigma} T_{ij,\rho\sigma}(1).$$

Hence we conclude that for any $\alpha > 0$ there exists θ sufficiently large so that

$$\theta T(\theta) + T(\theta)\bar{C}'\bar{C}T(\theta) - \tilde{A}(s)T(\theta) - T(\theta)\tilde{A}'(s) \geq \alpha I^{n \times n}.$$

Therefore for $0 \leq s \leq t$

$$\xi'(s)T(\theta)\xi(s) \quad \leq \quad e^{\alpha(s-t)}\zeta'\zeta$$

Now there exists $m_2(\theta) > 0$ such that

$$\xi'(s)\xi(s) \leq m_2(\theta)\xi'(s)T(\theta)\xi(s)$$

so we conclude that

$$\xi'(s)\xi(s) \leq m_2(\theta)e^{\alpha(s-t)}\zeta'\zeta.$$

There exist constants $m_3 > 0, m_4(\theta) > 0$ such that

$$\begin{aligned} P(0) &\leq m_3 I^{n\times n} \\ \Gamma + T(\theta)\bar{C}'\bar{C}T(\theta) &\leq m_4(\theta)I^{n\times n} \end{aligned}$$

From (12.15) we obtain the desired conclusion,

$$\begin{aligned} \zeta'P(t)\zeta &\leq m_3 e^{-\alpha t}\zeta'\zeta + m_4(\theta)\int_0^t e^{\alpha(s-t)}\zeta'\zeta \; ds \\ \zeta'P(t)\zeta &\leq m_3\zeta'\zeta + m_4(\theta)\int_{-\infty}^t e^{\alpha(s-t)}\zeta'\zeta \; ds \\ \zeta'P(t)\zeta &\leq \frac{m_3 + m_4(\theta)}{\alpha}\zeta'\zeta \end{aligned}$$

Define

$$Q(t) = P^{-1}(t)$$

then Q satisfies

$$\begin{aligned} \dot{Q}(t) &= -A'(t)Q(t) - Q(t)A(t) - Q(t)\Gamma Q(t) + \bar{C}'\bar{C} \\ Q(0) &= P^{-1}(0) > 0 \end{aligned} \tag{12.16}$$

Next we show that there exists $m_5 > 0$ such that for all $t \geq 0$

$$Q(t) \quad \leq \quad m_5 I^{n\times n}.$$

This will imply that

$$P(t) \quad \geq \quad \frac{1}{m_5}I^{n\times n}. \tag{12.17}$$

Consider the optimal control problem of minimizing

$$\xi'(0)Q(0)\xi(0) + \int_0^t \xi'(s)\bar{C}'\bar{C}\xi(s) + \mu'(s)\mu(s) \; ds$$

subject to

$$\begin{aligned} \dot{\xi}(s) &= A(s)\xi(s) + G\mu(s) \\ \xi(t) &= \zeta. \end{aligned}$$

It is well-known that the optimal cost is

$$\zeta' Q(t)\zeta$$

where $Q(t)$ is the solution of (12.5).

We use the suboptimal control

$$\mu(s) = G^{-1}(\alpha I^{n\times n} - A(s))\xi(s)$$

so that the closed loop dynamics is

$$\begin{aligned}\dot{\xi}(s) &= \alpha\xi(s)\\ \xi(t) &= \zeta\\ \xi(s) &= e^{\alpha(s-t)}\zeta.\end{aligned}$$

From this we obtain the desired bound

$$\begin{aligned}\zeta' Q(t)\zeta &\le \xi'(0)Q(0)\xi(0)\\ &\quad + \int_0^t \xi'(s)\left(\bar{C}'\bar{C} + (\alpha I^{n\times n} - A'(s))\Gamma^{-1}(\alpha I^{n\times n} - A(s))\right)\xi(s)\, ds\\ \zeta' Q(t)\zeta &\le e^{-2\alpha t}\zeta' Q(0)\zeta + \int_0^t e^{2\alpha(s-t)}\zeta'\left(\bar{C}'\bar{C} + (\alpha+L)^2\Gamma^{-1}\right)\zeta\, ds\\ \zeta' Q(t)\zeta &\le \left(m_6 + \frac{m_7}{2\alpha}\right)\zeta'\zeta\end{aligned}$$

where

$$\begin{aligned}Q(0) &\le m_6 I^{n\times n}\\ \bar{C}'\bar{C} + (\alpha+L)^2\Gamma^{-1} &\le m_6 I^{n\times n}.\end{aligned}$$

Now let $x(t), u(t), y(t)$ be a trajectory of the system (12.10) starting at x^0. Let $\hat{x}(t)$ be the trajectory of the extended Kalman filter (12.5) starting at $\hat{x}^0$ and $\tilde{x}(t) = x(t) - \hat{x}(t)$, $\tilde{x}^0 = x^0 - \hat{x}^0$. Then

$$\begin{aligned}\frac{d}{dt}\tilde{x}'(t)P(t)\tilde{x}(t) &= -\tilde{x}'(t)\left(\bar{C}'\bar{C} + P(t)\Gamma P(t)\right)\tilde{x}(t)\\ &\quad + 2\tilde{x}'(t)P(t)\left(\bar{f}(x(t),u(t)) - \bar{f}(\hat{x}(t),u(t)) - A(t)\tilde{x}(t)\right).\end{aligned}$$

Now following [1]

$$\begin{aligned}&\bar{f}(x(t),u(t)) - \bar{f}(\hat{x}(t),u(t)) - A(t)\tilde{x}(t)\\ &\quad = \int_0^1\int_0^1 r\frac{\partial^2 \bar{f}}{\partial x_i \partial x_j}(\hat{x}(t) + rs\tilde{x}(t), u(t))\tilde{x}_i(t)\tilde{x}_j(t)\, ds\, dr\\ &\quad \le L|\tilde{x}(t)|^2.\end{aligned}$$

Since G is invertible there exists $m_7 > 0$ such that

$$\Gamma \ge m_7 I^{n\times n}$$

and so

$$\begin{aligned}\frac{d}{dt}\tilde{x}'(t)P(t)\tilde{x}(t) &\leq -\frac{m_7}{m_5^2}|\tilde{x}(t)|^2 + m_1 L|\tilde{x}(t)|^3 \\ &\leq \left(-\frac{m_7}{m_5^2} + m_1 L\left(m_5\tilde{x}'(t)P(t)\tilde{x}(t)\right)^{\frac{1}{2}}\right) m_5\tilde{x}'(t)P(t)\tilde{x}(t).\end{aligned}$$

Therefore if

$$m_1 L\left(m_5\tilde{x}'(0)P(0)\tilde{x}(0)\right)^{\frac{1}{2}} < \frac{m_7}{m_5^2}$$

then $|\tilde{x}(t)| \to 0$ exponentially as $t \to \infty$.

12.3 Conclusions

The above result does not follow from that of Baras, Bensoussan and James [1]. To show local convergence of the extended Kalman filter they required "uniform detectability". They define this as follows. The system

$$\begin{aligned}\dot{x} &= f(x) \\ y &= Cx\end{aligned} \tag{12.18}$$

is uniformly detectable if there exists a bounded Borel matrix-valued function $\Lambda(x)$ and a constant $\alpha > 0$ such that for all $x, \xi \in \mathbb{R}^n$

$$\xi'\left(\frac{\partial f}{\partial x}(x) + \Lambda(x)C\right)\xi \leq -\alpha|\xi|^2.$$

This is a fairly restrictive condition as not all observable linear systems are uniformly detectable. Consider

$$\begin{aligned}\dot{x} &= \begin{bmatrix} 0 & 1 \\ a_1 & a_2 \end{bmatrix} x \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} x.\end{aligned}$$

Suppose

$$\Lambda(x) = \begin{bmatrix} \lambda_1(x) \\ \lambda_2(x) \end{bmatrix}$$

then

$$\frac{\partial f}{\partial x}(x) + \Lambda(x)C = \begin{bmatrix} \lambda_1(x) & 1 \\ \lambda_2(x) + a_1 & a_2 \end{bmatrix}.$$

If $a_2 > 0$ and $\xi' = \begin{bmatrix} 0 & 1 \end{bmatrix}$ then

$$\xi'\left(\frac{\partial f}{\partial x}(x) + \Lambda(x)C\right)\xi = a_2 > 0$$

so the system is not uniformly detectable. This system does satisfies the conditions of Theorem 12.1 so an extended Kalman filter would converge locally. Since the system is linear, an extended Kalman filter is also a Kalman filter that converges globally

An example [5] of a highly nonlinear problem where an EKF may fail to converge is

$$\begin{aligned} \dot{x} &= f(x) = x(1-x^2) \\ y &= h(x) = x^2 - x/2 \end{aligned} \tag{12.19}$$

where $x, y \in \mathbb{R}$. The system is observable as $y, \dot{y}, \ddot{y}$ separate points but it is not uniformly observable. The dynamics has two stable equilibria at $x = \pm 1$ and an unstable equilibrium at $x = 0$. Under certain initial conditions, the extended Kalman filter fails to converge. Suppose the $x^0 = 1$ so $x(t) = 1$ and $y(t) = 1/2$ for all $t \geq 0$. But $h(-1/2) = 1/2$ so if $\hat{x}^0 \leq -1/2$ the extended Kalman filter will not converge. To see this notice that when $\hat{x}(t) = -1/2$, the term $y(t) - h(\hat{x}(t)) = 0$ so $\dot{\hat{x}} = f(\hat{x}(t)) = f(-1/2) = -3/8$. Therefore $\hat{x}(t) \leq -1/2$ for all $t \geq 0$.

Acknowledgment

Research supported in part by NSF DMS-0204390 and AFOSR F49620-01-1-0202.

12.4 References

[1] J. S. Baras, A. Bensoussan and M. R. James, *Dynamic observers as asymptotic limits of recursive filters: special cases,* SIAM J. on Applied Mathematics, 48 (1988), pp. 1147-1158.

[2] J. P. Gauthier, H. Hammouri and S. Othman, *A simple observer for nonlinear systems with applications to bioreactors,* IEEE Trans. on Automatic Control, 37 (1992), pp. 875-880.

[3] A. Gelb, *Applied Optimal Estimation,* MIT Press, Cambridge, MA, 1974.

[4] A. J. Krener, *Normal forms for linear and nonlinear systems,* Contemporary Mathematics, 68, *Differential Geometry, the Interface between Pure and Applied Mathematics*, American Mathematical Society, Providence, RI, (1987), pp. 157-189.

[5] A. J. Krener and A. Duarte, *A hybrid computational approach to nonlinear estimation,* in Proc. of 35th Conference on Decision and Control, Kobe, Japan, (1996) pp. 1815-1819.

13

On the Separation of Two Degree of Freedom Controller and Its Application to H_∞ Control for Systems with Time Delay

Yohei Kuroiwa *Hidenori Kimura*

Abstract

An advantage of the two degree of freedom control scheme lies in the fact that the feedback controller and the feedforward controller can be designed separately. The former is designed to meet feedback properties such as sensitivity, robust stability and disturbance rejection, while the latter is designed to improve tracking performances. However, it is not completely clear whether the designs of feedback controller and feedforward controller can be done independently. The two kinds of specifications may interfere each other. In this note, a design scheme of the two degree of freedom control is established that guarantees the independence of each controller. This result is applied to the tracking H_∞ control for systems with time delay based on previous researches [1][2]. It is shown that the problem is reduced to two independent mixed sensitivity problems with time delay.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 183-192, 2003.
© Springer-Verlag Berlin Heidelberg 2003

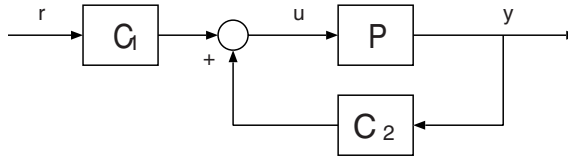


Figure 13.1 A standard form of the two degree of freedom control systems

13.1 Introduction

There are a number of important specifications for control system designs, that is, internal stability, reference tracking and feedback properties such as sensitivity, robust stability,etc. In conventional unity feedback control systems, as the plant input is generated by the controller based on single information, that is, the error between the actual plant outputs and the reference signals, there are fundamental limitations which need some sort of trade-off among the specifications. It is known that if one wishes to achieve the above specifications simultaneously, the two degree of freedom controller which is initially discussed by Horowitz [3] is necessary, instead of the usual unity feedback controller.

In the two degree of freedom control systems, the control inputs are generated by the reference signals and the measurements. Fig.1 illustrates a standard form of the two degree of freedom controller. Here the plant input is given by u,

$$u = C_1 r + C_2 y, \tag{13.1}$$

where r and y are reference signals and measurements, respectively, and C_1 and C_2 denote feedforward controller and feedback controller which guarantees internal stability, respectively. The two degree of freedom controller of this type may be parametrized in terms of two stable but otherwise free Q parameters [4][5].

The basic idea of the two degree of freedom control scheme is that the role of the feedback controller is to meet the specifications of internal stability, disturbance rejection, measurement noise attenuation and sensitivity minimization, while that of the feedforward controller is to improve the tracking performances. In this sense the objectives of the feedback controller and the feedforward controller are clear. However, it is not completely clear whether designs of the feedback controller and the feedforward controller are independent to each other.

Design problems of the two degree of freedom controller have been discussed by many researchers in the area of H_∞ control [1][4][6]. The approach of the two degree of freedom problems can be extended in the general multivariable case, if a separation is achieved, where the feedforward controller and the feedback controller are obtained independently. In the paper [1], the two degree of freedom H_∞ minimization problems are solved by transforming the original problems into two standard problems. Also the feedback error learning scheme which was proposed as an architecture of the brain motor control is formulated as a two degree of freedom controller equipped with the adaptive capability in the feedforward controller [7]. The salient feature of the feedback error learning lies in its use of feedback error for learning the inverse model.

In this note, we shall show that each controller can be designed independently in the framework of H_∞ control. By the separation [1], the H_∞ control problem of feedforward part for systems with time delay becomes equivalent to the mixed sensitivity H_∞ control problem for systems with time delay. Recently, this problem is solved by J-spectral factorization approach [2], so we can use the solution for the problem directly. We apply this separation to the H_∞ tracking control problem for systems with time delay. More precisely, this problem is reduced to the H_∞-norm minimization of the transfer function from reference, disturbance and measurement noise to tracking error, the plant and controller inputs.

13.2 System Description and Compensator Structure

We consider a two degree of freedom control system shown in Fig.2, where P is a linear time invariant multivariable plant which can be infinite dimensional, C_1 is a feedforward controller, and C_2 is a feedback controller. The reason why the block PC_1 is inserted from r to v will be explained later. The following notation is used to denote the dimension of the vector, for example, $\dim(\ r\ ) = n_r$. The tracking error is defined as

$$e = y - r. \tag{13.2}$$

The controlled outputs are defined by

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} := \begin{bmatrix} W_1 v \\ W_2 u \\ W_3 e \end{bmatrix}, \tag{13.3}$$

where W_1, W_2 and W_3 are weighted functions, which are stable transfer functions. And we obtain the transfer function $\Sigma(C_1, C_2)$ from external inputs to controlled outputs

$$z = \Sigma(C_1, C_2)w, \tag{13.4}$$

$$\Sigma(C_1, C_2) = \begin{bmatrix} 0 & -W_1 S & -W_1 S \\ W_2 C_1 & -W_2 Q & -W_2 Q \\ W_3(I - PC_1) & -W_3 S & W_3 T \end{bmatrix}, \tag{13.5}$$

where the sensitivity function S for the aforementioned feedback system shown in Fig.2 is

$$S = (I + PC_2)^{-1}, \tag{13.6}$$

the complementary sensitivity functions T is

$$T = PC_2(I + PC_2)^{-1}, \tag{13.7}$$

and the Q operator is defined as follows:

$$Q = C_2(I + PC_2)^{-1}. \tag{13.8}$$

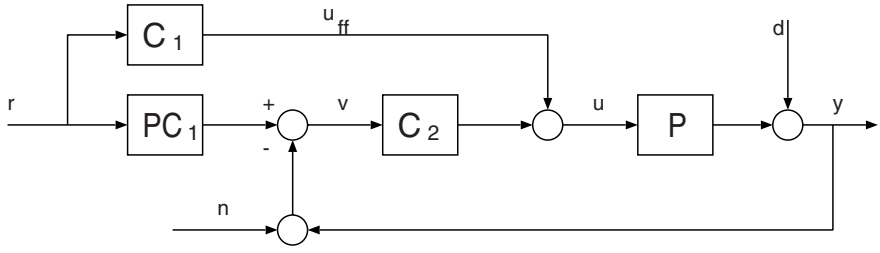


Figure 13.2 The two degree of freedom controller

Obviously, $\Sigma(C_1, C_2)$ can be written as

$$\Sigma(C_1, C_2) = \left[\begin{array}{cc} \Sigma_1(C_1) & \Sigma_2(C_2) \end{array} \right], \tag{13.9}$$

where

$$\Sigma_1(C_1) = \left[\begin{array}{c} 0 \\ W_2 C_1 \\ W_3(I - PC_1) \end{array} \right], \tag{13.10}$$

and

$$\Sigma_2(C_2) = \left[\begin{array}{cc} -W_1 S & -W_1 S \\ -W_2 Q & -W_2 Q \\ -W_3 S & W_3 T \end{array} \right]. \tag{13.11}$$

The representation (13.5) implies that the feedforward controller C_1 and the feedback controller C_2 appear separately, and this separated form is due to insertion of PC_1 from r to v. It allows the separation between the tracking properties included in $\Sigma_1(C_1)$ which is a function of C_1 only and feedback properties such as sensitivity, robust stability and disturbance attenuation included in $\Sigma_2(C_2)$ which is a function of the C_2 only. Normally in the scheme of Fig.2, C_1 is parameterized by $P^{-1}W$ where W is a filter with appropriate relative degree that makes $P^{-1}W$ proper. This is well known two degree of freedom control scheme.

The system shown in Fig.2 is said to be internally stable if all the transfer functions from external inputs to each controller and plant outputs are stable. We note, at this point, that the feedback part of the two degree of freedom controller, namely C_2, must stabilize the closed loop. The feedforward part of the two degree of freedom controller, namely PC_1 and C_1, have to be stable since they are open loop.

13.3 Application of H_∞ Control to Systems with Time Delay

The characteristic of systems with time delay is that the action of control inputs takes a certain time before it affects the measured outputs. We consider the typical systems with time delay,

$$P = e^{-sh} P_r, \tag{13.12}$$

where P_r is some rational function, and h is a delay time. Models like this appear frequently in applications. It can represent a number of type of delay systems in the real world such as transport delay, sensor delay and actuator delay. They often serve as a simple yet adequate model for otherwise complicated high-order or infinite dimensional systems. Though it is not trivial whether infinite dimensional system is Bezout domain, the systems with time delay like (13.12) have a matrix fraction representation over H_∞ which is coprime in the sense that a matrix Bezout identity can be satisfied. [8][9].

We consider the sub-optimal H_∞ minimization problem of $\Sigma(C_1, C_2)$ which is defined in the previous section, that is,

$$||\Sigma(C_1, C_2)||_\infty < \gamma. \tag{13.13}$$

From (13.9), it follows that

$$||\Sigma(C_1, C_2)||_\infty < ||\Sigma_1(C_1)||_\infty + ||\Sigma_2(C_2)||_\infty, \tag{13.14}$$

if we select the feedforward controller C_1 and the feedback controller C_2 respectively to satisfy

$$||\Sigma_1(C_1)||_\infty < \gamma_1, ||\Sigma_2(C_2)||_\infty < \gamma_2 \tag{13.15}$$

such that $\gamma > \gamma_1 + \gamma_2$, then (13.13) holds.

Fundamentally, the solution of the minimization problems of $\Sigma_1(C_1)$ and $\Sigma_2(C_2)$ are based on that for the mixed sensitivity H_∞ problem shown in Fig.3 based on the J-spectral factorization approach [2]. Like the Smith predictor, it solves the problem by transforming original problem, which is infinite dimensional, into the finite dimensional one. The essence of its solution is that the irrational J-spectral factorization is reduced to the rational J-spectral factorization via a partial fraction expansion of an irrational transfer function, and the controller corresponding to central solution is a feedback interconnection of a finite dimensional system and a finite memory system, see Appendix. The computations needed to construct these controllers are all matrix computations involving only a finite dimensional Riccati equation, so the method is easy to implement.

Now we concentrate on solving H_∞ control problem of (13.15), where $\Sigma_2(C_2)$ is different from the mixed sensitivity setting. Nevertheless, J-spectral factorization approach given in [2] which is discussed in Appendix is applicable. $\Sigma_2(C_2)$ can be derived by

$$\Sigma_2(C_2) = HM(G_2; C_2), \tag{13.16}$$

where G_2 is defined as follows:

$$G_2 := \begin{bmatrix} 0 & 0 & -W_1 & -W_1 \\ -W_2 & -W_2 & 0 & 0 \\ 0 & W_3P & -W_3 & 0 \\ P & 0 & I & 0 \\ 0 & P & 0 & I \end{bmatrix}, \tag{13.17}$$

and HM represents HoMographic transformation defined in [10]. The fundamental computations shown in Appendix to reduce the infinite dimensional problem to a finite dimensional problem are as follows:

$$\tilde{G_2} J_\gamma G_2 := \begin{bmatrix} T_{11} & e^{sh} T_{12} \\ e^{-sh} T_{21} & T_{22} \end{bmatrix}, \tag{13.18}$$

$$\begin{aligned} T_{11} &= \begin{bmatrix} \tilde{W_2} W_2 - \gamma^2 \tilde{P_r} P_r & \tilde{W_2} W_2 \\ W_2 \tilde{W_2} & \tilde{W_2} W_2 + \tilde{P_r} \tilde{W_3} W_3 P_r - \gamma^2 \tilde{P_r} P_r \end{bmatrix} \\ T_{12} &= \begin{bmatrix} -\gamma^2 \tilde{P_r} & 0 \\ -\tilde{P_r} \tilde{W_3} W_3 & -\gamma^2 \tilde{P_r} \end{bmatrix}, T_{21} = \begin{bmatrix} -\gamma^2 P_r & -\tilde{W_3} W_3 P_r \\ 0 & -\gamma^2 P_r \end{bmatrix} \\ T_{22} &= \begin{bmatrix} \tilde{W_1} W_1 + \tilde{W_3} W_3 - \gamma^2 I & \tilde{W_1} W_1 \\ \tilde{W_1} W_1 & \tilde{W_1} W_1 - \gamma^2 I \end{bmatrix}, \end{aligned}$$

where its components of $T_{ij} (i, j = 1, 2)$ are finite dimensional transfer functions. The form of (13.18) is identical to that of $(A.5)$ which is derived in [2], so we apply this method.

As before, the plant P is assumed to be the form $e^{-sh} P_r$ with the P_r rational, so the rational part of the plant has the standard coprime factorization [4], that is,

$$P_r = P_d^{-1} P_n, \tag{13.19}$$

where P_d and P_n are in RH_∞, and P_d can be chosen inner. So we can rewrite the H_∞-norm minimization problems,

$$||\Sigma_1(C_1) P_d^{-1}||_\infty < \gamma_1, ||\Sigma_2(C_2) P_d^{-1}||_\infty < \gamma_2. \tag{13.20}$$

As a result, the multiplication of P_d^{-1} makes (13.20) equivalent to a mixed sensitivity setting [2].

Next, we focus on the the sub-optimal H_∞-norm minimization of $\Sigma_1(C_1)$, whose components are linear functions of the feedforward controller C_1. The H_∞-norm minimization problem for systems with time delay is given by

$$\left\| \begin{bmatrix} 0 \\ W_2 C_1 \\ W_3 (I - P C_1) \end{bmatrix} P_d^{-1} \right\|_\infty \tag{13.21}$$

with constraints on stability of the open loop transfer functions C_1 and PC_1, which is equivalent to the mixed sensitivity H_∞ problem for systems with time delay, that is,

$$\left\| \begin{bmatrix} 0 \\ W_2 C_0 (I + P C_0)^{-1} \\ W_3 (I + P C_0)^{-1} \end{bmatrix} P_d^{-1} \right\|_\infty . \tag{13.22}$$

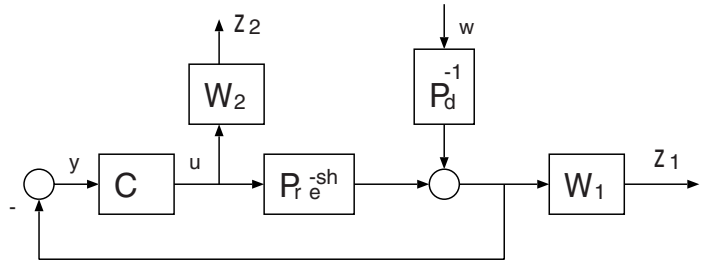


Figure 13.3 The mixed sensitivity configuration

Here we represent C_1 as

$$C_1 = C_0(I + PC_0)^{-1}. \tag{13.23}$$

As above mentioned, the mixed sensitivity H_∞ problem for systems with time delay shown in Fig.3 is solved [2].

So we can use that solution directly. This two-block H_∞-norm minimization problem is solved, and the closed loop transfer function is a stable transfer function, that is,

$$(I + PC_0)^{-1} \tag{13.24}$$

is stable. This means that C_1 and PC_1 are both stable transfer functions. The controller C_0 corresponding to central solution is the feedback interconnection of the finite dimensional part K and the infinite dimensional part F_{stab}, where the definition of K and F_{stab} are given in $(A.9)$ and $(A.6)$,respectively,

$$C_0 = (I - KF_{stab})^{-1}K, \tag{13.25}$$

which is derived by the solution for the mixed sensitivity H_∞ problem for systems with time delay. As a result, the controller C_1 corresponding to central solution shown in Fig.4 is given by

$$C_1 = (I - K(F_{stab} - P))^{-1}K. \tag{13.26}$$

At this point, it follows that the further feedback interconnection incorporating the plant must be appended. This is the essential difference from the controller of the mixed sensitivity setting.

13.4 Conclusion

In this note, it has been shown that a separation of the two degree of freedom controller is established, which enables to design the feedforward controller and the feedback controller independently. The result is applied to the tracking H_∞ control problem for systems with time delay. As a result, the tracking H_∞ control problem is reduced to two H_∞-norm minimization problems. One is the design of feedback controller. The other is concerned with the design of feedforward controller with stability constraints on the open loop transfer functions. However, the H_∞-norm minimization problem of feedforward controller can be transformed

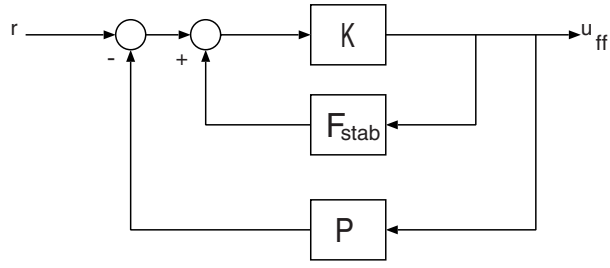


Figure 13.4 The controller C_1 corresponding to central solution

to an equivalent mixed sensitivity problem which was already solved in [2], and its solution achieves the stability of open loop transfer functions automatically, that is, the stability of the closed loop means the stability of open loop.

Although in the previous research [1] the separation of the feedback controller and the feedforward controller is regarded as a tool which enables to transform the two degree of freedom H_∞ minimization problem into two standard problems, it has more potential beyond the only tool to separately design two controllers. For example, it can be used for various adaptive scheme including feedback error learning one. These are topics for further research.

[1] I. Yaesh and U. Shaked(1991). Two-Degree-of-Freedom H_∞-Optimization of Multivariable Feedback Systems, IEEE Trans. Automat. Contr., vol. 36, pp. 1272-1276.

[2] G. Meinsma and H. Zwart(2000). On H_∞ Control for Dead-Time Systems, IEEE Trans. Automat. Contr., vol. 45, pp. 272-285.

[3] I. Horowitz(1963). Synthesis of Feedback Systems. New York: Academic.

[4] M. Vidyasagar(1985). Control System synthesis: A Factorization Approach. Cambridge, MA: M.I.T. Press.

[5] D. C. Youla and J. J. Bongiorno(1985). A feedback theory of two-degree-of-freedom optimal Wiener-Hopf design, IEEE Trans. Automat. Contr., vol. 29, pp. 652-665.

[6] D. J. N. Limebeer, E. M. Kasenally and J. D. Perkins(1993). On the Design of Robust Two Degree of Freedom Controllers, Automatica, vol. 29, pp. 157-168.

[7] A. Miyamura and H. Kimura(2002). Stability of feedback error learning scheme, Syst. Contr. Lett., vol. 45, pp. 303-316.

[8] M. C. Smith(1989). On Stabilization and the Existence of Coprime Factorizations, IEEE Trans. Automat. Contr., vol. 34, pp. 1005-1007.

[9] R. F. Curtain and H. J. Zwart(1995). An Introduction to Infinite-dimensional Linear Systems Theory. New York: Springer-Verlag.

[10] H. Kimura(1996). Chain-Scattering Approach to H_∞ Control. Birkh$\ddot{a}$user

Appendix

We briefly review the results of [2] which enable to reduce the mixed sensitivity problem for systems with time delay to usual mixed sensitivity one. Fig.3 shows the closed loop configuration corresponding to the mixed sensitivity problem for systems with time delay, that is, $||H||_\infty < \gamma$, where H is defined as follows:

$$H := \begin{bmatrix} W_1(I+PC)^{-1}P_d^{-1} \\ W_2C(I+PC)^{-1}P_d^{-1} \end{bmatrix}. \tag{A.1}$$

Here P is the typical systems with time delay,

$$P = e^{-sh}P_r, \tag{A.2}$$

where the P_r is rational, and has the standard coprime factorization, that is, $P_r = P_d^{-1}P_n$, where P_n and P_d^{-1} are in RH_∞, and P_d^{-1} can be chosen to be inner. The open loop(i.e., the controller taken away) can be expressed as a map from (u, y) to (w, z_1, z_2) as

$$\begin{bmatrix} -z_1 \\ z_2 \\ -w \end{bmatrix} = \begin{bmatrix} 0 & W_1 \\ W_2 & 0 \\ e^{-sh}P_n & P_d \end{bmatrix} \begin{bmatrix} u \\ y \end{bmatrix}, \tag{A.3}$$

where we define as follows:

$$G := \begin{bmatrix} 0 & W_1 \\ W_2 & 0 \\ e^{-sh}P_n & P_d \end{bmatrix} \in H_\infty^{(n_{z_1}+n_{z_2}+n_w)\times(n_u+n_y)}, \tag{A.4}$$

and

$$J_\gamma := \begin{bmatrix} I_{n_{z_1}} & 0 & 0 \\ 0 & I_{n_{z_2}} & 0 \\ 0 & 0 & -\gamma^2 I_{n_w} \end{bmatrix}, J := \begin{bmatrix} I_{n_u} & 0 \\ 0 & -I_{n_y} \end{bmatrix}.$$

From the algebraic manipulations, it follows that

$$\tilde{G}J_\gamma G := \begin{bmatrix} S_{11} & e^{sh}S_{12} \\ e^{-sh}S_{21} & S_{22} \end{bmatrix}, \tag{A.5}$$

where S_{ij} $(i, j = 1, 2)$ are the rational transfer functions. Here we write $e^{-sh}S_{22}^{-1}S_{21}$ as a sum of a irrational, but stable, part F_{stab}, and a proper rational part R

$$e^{-sh}S_{22}^{-1}S_{21} = F_{stab} + R, \tag{A.6}$$

then write

$$\Theta := \begin{bmatrix} I & -\tilde{F_{stab}} \\ 0 & I \end{bmatrix} \tilde{G} J_\gamma G \begin{bmatrix} I & 0 \\ -F_{stab} & I \end{bmatrix}$$
$$= \begin{bmatrix} S_{11} - S_{12}S_{22}^{-1}S_{21} + \tilde{R}S_{22}R & \tilde{R}S_{22} \\ S_{22}R & S_{22} \end{bmatrix}. \quad (A.7)$$

The point is that Θ defined here is rational and proper, and that the factor $\begin{bmatrix} I & 0 \\ -F_{stab} & I \end{bmatrix}$ is bistable. A stabilizing controller exists such that $||H||_\infty < \gamma$ if and only if the following three conditions hold [2].
(1) $\Theta(\infty)$ is nonsingular and has the same number of positive and negative eigenvalues as J, and $\Theta(j\omega)$ is nonsingular on the imaginary axis.
(2) A bistable Q_r exists such that $\Theta = \tilde{Q_r} J Q_r$, and the lower right $n_y \times n_y$-block M_{22} of $M := G \begin{bmatrix} I & 0 \\ -F_{stab} & I \end{bmatrix} Q_r^{-1}$ is bistable.
(3) $\Theta_{22}(\infty) < 0$ if $h > 0$.
In this case, all stabilizing controllers are parameterized by

$$C = (Z_{11}U + Z_{12})(Z_{21}U + Z_{22})^{-1}, \quad (A.8)$$

where $Z := \begin{bmatrix} I & 0 \\ -F_{stab} & I \end{bmatrix} Q_r^{-1} \in H_\infty^{(n_u+n_y)\times(n_u+n_y)}$ and $U \in H_\infty^{n_u \times n_y}$ with $||U||_\infty < 1$. Moreover, Q_r can be chosen such that its upper-right $n_u \times n_y$ block $Q_{r,12}$ satisfies $Q_{r,12}(\infty) = 0$, and for this choice of Q_r, the controller is causal for any strictly proper U. The controller corresponding to central solution is given by $C = (I - KF_{stab})^{-1}K$, where

$$K = Z_{r,12}Z_{r,22}^{-1} \quad (A.9)$$

is finite dimensional, and Q_r^{-1} is defined as follows:

$$Q_r^{-1} := \begin{bmatrix} Z_{r,11} & Z_{r,12} \\ Z_{r,21} & Z_{r,22} \end{bmatrix}. \quad (A.10)$$

14

The Principle of Optimality in Measurement Feedback Control for Linear Systems

Alexander B. Kurzhanski

Abstract

This paper discusses a Dynamic Programming approach to the solution of the measurement feedback target control problem in finite time for a linear system with unknown but bounded disturbances in the system inputs and measurement channel. The control parameters and unknown items are subjected to hard (instantaneous) bounds. The paper also emphasizes the feasibility of the techniques of duality theory of nonlinear analysis for calculating the solutions.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 193-202, 2003.
© Springer-Verlag Berlin Heidelberg 2003

14.1 Introduction

This paper deals with the problem of measurement feedback target control for linear systems subjected to unknown but bounded disturbances with hard bounds on the controls and the uncertain items. The basic problem is split into two coupled problems - the one of guaranteed (set-membership) state estimation and the other of feedback control under incomplete information. The solution to the first one is described by level sets for the solution of a related HJB (Hamilton-Jacobi-Bellman) equation and the solution to the second one through an HJBI (Hamilton-Jacobi-Bellman-Isaacs) equation in the space of value functions for the first equation. This is a complicated scheme. Nevertheless, in the case of linear systems the overall problem is solvable through an array of finite-dimensional optimization problems.

14.2 The Basic Problem and the Cost Functional

We start with the formulation of the basic problem and of the general approach to its solution.

Consider the system

$$dx/dt = A(t)x + B(t)u + C(t)v(t), \tag{14.1}$$

with continuous matrix coefficients $A(t), B(t), C(t)$. The realizations of the controls $u(t)$, and disturbances $v(t)$ are taken to be bounded by hard (geometric) bounds

$$u(t) \in \mathcal{P}(t), \quad v(t) \in Q(t), \tag{14.2}$$

for all $t \in [t_0, t_1]$. Here $\mathcal{P}(t), Q(t)$ - are set-valued functions with values in the variety of convex compact sets of the Euclidean spaces $\mathbb{R}^p, \mathbb{R}^q$ respectively, continuous in the Hausdorff metric.

The online information on the vector x arrives through available observations due to the measurement equation

$$y(t) = H(t)x + w, \tag{14.3}$$

where $y(t) \in \mathbb{R}^m$ is the given measurement, $w(t)$ is the unknown but bounded measurement "noise"

$$w(t) \in \mathcal{R}(t), \quad t \in [t_0, t_1], \tag{14.4}$$

with function $\mathcal{R}(t)$ similar to $\mathcal{P}(t), Q(t)$ and H(t) continuous.

The initial condition is given by the inclusion

$$x(t_0) \in X^0, \tag{14.5}$$

where X^0 - is a given convex compact set in $\mathbb{R}^n$. The pair $\{t_0, X^0\}$ will be referred to as the *initial position* of the system.

Given the initial position $\{t_0, X^0\}$, the functions $A(t), B(t), C(t), H(t)$, the realization of the "used" control $u[s], s \in [t_0, t]$, as well as the set-valued functions $Q(t), \mathcal{R}(t)$ and the available measurements $y_t(s) = y(t+s), s \in [t_0 - t, 0]$, one may apply the theory of guaranteed (set-membership) estimation, [5], [7], [14]

to calculate the *information sets* $\mathcal{X}(t, y_t(\cdot)) = \mathcal{X}(t, \cdot)$ of system (14.1)- (14.4), consistent with its parameters and with the measurements obtained within time $[t_0, t]$.

Therefore the state space variable (or the"current position") of the system may be taken as the pair $\{t, \mathcal{X}(t, \cdot\}$.

In a loose formulation the problem under consideration consists in specifying such a feedbak control strategy $U(t, \mathcal{X}(t, \cdot))$,(possibly multivalued), that would steer the system (namely, transfer the state space variable) from any initial position $\{\tau, \mathcal{X}(\tau, \cdot)\}, \tau \in [t_0, t_1]$, to a preassigned neighborhood of a given target set $\mathcal{M}$ at time t_1 *despite the unknown disturbances and the incomplete information*. At the same time the class of admissible strategies $\mathcal{U}_V = \{U(t, \mathcal{X}(t, \cdot))\}$ must ensure the existence and prolongability of solutions to the equation (differential inclusion) (14.1)) with $u = U(t, \mathcal{X}(t, \cdot)), \ \ t \in [t_0, t_1]$.

In the sequel we will formulate a more rigorous setting of the basic problem. However, even the loose setting indicates that the problem may be split into two - Problem (E) of set-membership estimation and Problem (C) of feedback control. The overall problem may be described through the following cost functional.

$$\mathcal{V}(t_0, X^0) = \min_U \max\{-d^2(x(t_0), X^0) + \int_{t_0}^{\tau} (d^2(u(t), \mathcal{P}(t)) - d^2(v(t), Q(t))dt$$

$$- \int_{t_0}^{\tau} d^2(y^*(t) - H(t)x(t), \mathcal{R}(t))dt + d^2(x(t_1), \mathcal{M}) \ |x(t_0); v(t) \in Q(t), t \in [t_0, \tau]\} \tag{14.6}$$

Here the minimum is to be taken over strategies of the type $U = U(t, \mathcal{X}(t, \cdot))$, where $\mathcal{X}(t, \cdot)$ is the information (consistency) domain which determines the current state $\{t, \mathcal{X}(t, \cdot))\}$ of the system. The details of this problem are explained in the next sections.

14.3 Guaranteed (Set-Membership) Estimation

As mentioned above, the problem of calculating the functional $\mathcal{V}(\tau, \mathcal{X}(\tau, \cdot))$ can be split into two problems, starting with Problem (E) - the one of guaranteed (set-membership) estimation. We consider Problem (E) in two possible settings – E_1 and E_2.

Problem E_1

At instant τ given are system (14.1) - (14.2) and measurement equation (14.3)), (14.4), as well as initial position $\{t_0, \mathcal{X}^0\}$, available observation (measurement) $y = y^*(t), \ t \in [t_0, \tau]$, and the realization $u = u^*(t)$ of the control that had already operated throughout the interval $t \in [t_0, \tau]$.

One is to specify the *information set* $\mathcal{X}(\tau, \cdot))$ of solutions $x(\tau)$ to system (14.1) consistent with measurement $y^*(t), \ t \in [t_0, \tau]$ under constraints (14.1) - (14.5)) and given realization $u^*(t), t \in [t_0, \tau]$ **!**

The "information set" $\mathcal{X}(\tau, \cdot))$ is a guaranteed estimate of the realized solution of equation (14.1) which contains the unknown actual actual vector $x(\tau)$ of this equation. The calculation of set $\mathcal{X}(\tau, \cdot))$ is the topic of guaranteed estimation theory [5], [14]. In order to treat the problem of this paper it is however necessary

to decribe not only $\mathcal{X}(\tau,\cdot)$ itself, but also its evolution in time. This can be done through the following auxiliary problem whose form is borrowed from the theory of H_∞ control [1], [4].

Problem E_2

Given initial position $\{t_0,\mathcal{X}^0\}$, measured values $y^*(t)$, and control realization $u^*(t)$ for $t\in[t_0,\tau]$, specify

$$-V(t,x)=\max\{-d^2(x(t_0),X^0)+\int_{t_0}^{\tau}(-d^2(v(t),Q(t))dt$$

$$-\int_{t_0}^{\tau}d^2(y^*(t)-H(t)x,\mathcal{R}(t))dt\ |x(t_0);v(t)\in Q(t),t\in[t_0,\tau]\} \tag{14.7}$$

under condition $x(\tau)=x$, due to system (14.1)**!**

Function $V(\tau,x)$ will be referred to as the "information state" of the overall system (14.1) - (14.5).

LEMMA 14.1

The information set $\mathcal{X}(\tau,\cdot)$ is the level set of function (the information state) $V(\tau,x)$:

$$\mathcal{X}(\tau,\cdot)=\{x:V(\tau,x)\le 0\}.$$

□

From the last assertion it follows that *the current position (state space variable)* of system (14.1) - (14.5)) may be selected not only as $\{\tau,\mathcal{X}(\tau,\cdot)\}$, but also as the pair $\{\tau,V(\tau,\cdot)\}$. We will further accept the last option.

For the function $V(\tau,x)$ we introduce the notation $V(\tau,x)=V(\tau,x|V(t_0,\cdot))$, emphasizing the dependence of $V(\tau,x)$ on $V(t_0,x)$.

LEMMA 14.2

The following property is true

$$V(\tau,x|V(t_0,\cdot))=V(\tau,x|V(t,\cdot|V(t_0,\cdot))),\ t_0\le t\le\tau. \tag{14.8}$$

□

REMARK 14.1

Formula (14.8) expresses the "Principle of Optimality" for Problem (E) taken in form E_2 **!**

The latter formula yields a partial differential equation for $V(t,x)$, whose formal derivation follows the standatd routines of Dynamic Programming.

Thus we have

$$\frac{\partial V}{\partial t}+\max_v\{(\frac{\partial V}{\partial x},A(t)x+B(t)u+C(t)v)\} \tag{14.9}$$

$$-d^2(v,Q(t))-d^2(y^*(t)-H(t)x),\mathcal{R}(t))=0,$$

under boundary condition

$$V(t_0,x)=d^2(x,X^0). \tag{14.10}$$

□

Equation (14.9), (14.10) may have no classical solution as $V(t,x)$ may not be smooth. Then the solution to this equation should be understood in the generalized sense, as "viscosity", [3] or "minmax" [16] solution. In the general case it may be defined trough Dini subdifferentials or through equivalent notions [2].

THEOREM 14.1
If with $u = u^*(t),\ \ y = y^*(t),\ t \in [t_0, \tau]$ given equation (14.9), (14.10) does have a generalized (viscosity) solution $V(t,x)$, then

$$\mathcal{X}(\tau, \cdot)) = \{x : V(\tau, x) \leq 0\}.$$

□

The existence of a generalized solution to equation (14.9), (14.10) is necessary and sufficient for the solvability of Problem (E). In this case $V(t,\cdot) = V(t, \cdot | V(t_0, \cdot))$ satisfies the *evolution equation*

$$\frac{\partial V}{\partial t} = \Phi(t, V(t,\cdot), u^*(t), y^*(t)), \tag{14.11}$$

in the Hilbert space of elements $\{V(t,\cdot)\}$, which is nothing else than equation (14.9), (14.10).

A question arises whether does the tube $\mathcal{X}[\tau] = \mathcal{X}(\tau, \cdot)$ satisfy any evolution equation in the space of set-valued functions? The answer known at this time is given in paper [7], where such an equation is given in the form of a funnel equation for a differential inclusion of type

$$dx/dt \in A(t)x + B(t)u^*(t) + C(t)Q(t), \tag{14.12}$$

under the state constraint

$$H(t)x \in Y(t) = y^*(t) - \mathcal{R}(t). \tag{14.13}$$

Here is one of such equations

$$\lim_{\sigma \to +0} \sigma^{-1} h_+(\mathcal{Z}[t+\sigma], (I + \sigma A(t))(\mathcal{Z}(t) \cap Y(t)) + \\ + B(t)u^*(t) + C(t)Q(t)) = 0, \tag{14.14}$$

where - h_+ - is the Hausdorff semidistance:

$$h_+(\mathcal{X}', \mathcal{X}") = \min\{\varepsilon : \mathcal{X}' \subseteq \mathcal{X}" + \varepsilon\mathcal{B}\},$$

$\mathcal{B}$ is a unit ball in $\mathbb{R}^n$. The solution $\mathcal{Z}[t]$; of equation (14.14) – is a set-valued function with $\mathcal{Z}[t_0] = \mathcal{X}^0$. This solution is nonunique.

The desired solution $\mathcal{X}[t]$, which coincides with the realization of the set-valued function $\mathcal{X}(t,\cdot)$, is the solution of (14.14) which is *maximal with respect to inclusion*, namely, $\mathcal{X}[t] \supseteq \mathcal{Z}[t]$.

Note that equation (14.14) makes sense for piecewise-continuous functions $u(t)$, $y(t)$ which in this paper are taken to be right-continuous.

We now pass to Problem (C) of control synthesis for system (14.9), (14.10).

14.4 Control Synthesis for the Set-Valued Control System

Consider equations (14.9))-(14.10), which describe the dynamics of value function $V(t,x)$, whose level sets are the estimates $\mathcal{X}(t,\cdot)$ of the phase vector $x(t)$ of the original system (14.1)-(14.2).

Problem (C)

With $V(\tau,x)$ given, find value functional

$$\mathcal{V}(\tau,V(\tau,\cdot)|\mathcal{V}(t_1,\cdot)) = \tag{14.15}$$

$$= \min_{U}\max_{x,v,w}\{-V(\tau,x)+$$

$$+\int_{\tau}^{t_1}(d^2(U,\mathcal{P}(t))-d^2(v,Q(t))-d^2(y(t)-G(t)x(t),\mathcal{R}(t))dt+h^2_+(\mathcal{X}(t_1,\cdot)\mathcal{M})\},$$

along the solution $V(t,x)$ of equation

$$\frac{\partial V}{\partial t}+\max_{v}\{(\frac{\partial V}{\partial x},A(t)x+B(t)u)+C(t)v))\}-$$

$$-d^2(v,Q(t))-d^2(y(t)-G(t)x(t),\mathcal{R}(t))=0, \tag{14.16}$$

with starting position $\{\tau,V(\tau,\cdot)\}$ given, being obtained from equation (14.9), (14.10)**!**

Thus the state space variable ("the position") of the overall system with incomplete measurements is the pair $\{t,V(t,\cdot))\}$ and its "trajectory" in time, described by equation (14.16), is $V(t,\cdot)$ - a function of t with values in the space of functions of x which, with t fixed, are $V(t,x),\ x\in\mathbb{R}^n$. We therefore have to minimaximalize (14.15) over the indicated "trajectories" $V(t,\cdot)$.

The control $U=U(t,\cdot)$ in (14.15) is to be selected as a set-valued functional of the state space variable $\{t,V(t,\cdot)\}$, namely, as $U=U(t,V(t,\cdot))$, where the latter functional is continuous in t and upper semicontinuous in $V(t,\cdot)$ in the metric given by an $\mathcal{L}^n_2(\mathcal{B}(r))$ - norm. (In this paper we presume that functions $V(t^*,x)$ are elements of a ball with center at the origin, of radius ρ in the Hilbert space $L^n_2(\mathcal{B}(r))$ of n-dimensional functions square-integrable in the domain $\mathcal{B}(r)$ - a ball of radius r in $\mathbb{R}^n$, where ρ,r are sufficiently large, $\mathcal{B}(0)=\{0\}$.)

The function $x(t)=x(t,\tau,x^*)$ in (14.15), (14.16) is the trajectory of system (14.1) which starts at $\{\tau,x^*\},\ x^*\in\mathcal{X}(\tau,\cdot)=\{x:V(\tau,x)\le 0\}$ and the term

$$d^2(\mathcal{X}(t_1,\cdot),\mathcal{M})=\max_{x}\{d^2(x,\mathcal{M})|V(t_1,x)\le 0\}.$$

The boundary condition for (14.16) is then given by condition

$$\mathcal{V}(t_1,V(\cdot))=\max_{x}\{d^2(x,\mathcal{M})|V(t_1,x)\le 0\}, \tag{14.17}$$

We may now specify the underlying "principle of optimality".

LEMMA 14.1
The following Principle of Optimality in the class of state space variables $\{t, V(t,\cdot)\}$ is true $(\tau \le t \le t_1):$.

$$\mathcal{V}(\tau, V(\tau,\cdot)|t_1, \mathcal{V}(t_1,\cdot)) = \mathcal{V}(\tau, V(\tau,\cdot)|t, \mathcal{V}(t,\cdot|t_1, \mathcal{V}(t_1,\cdot)),$$

where $\mathcal{V}(t_1,\cdot))$ is determined by the boundary condition (14.17). □

The last assertion yields a generalized HJBI equation which may formally be written as

$$\min_u \max_{v,y}\{\frac{d\mathcal{V}(\tau, V(\tau,\cdot))}{dt} + d^2(u, \mathcal{P}(t)) - d^2(v, Q(t)) - d^2(y - H(t)x(t), \mathcal{R}(t))\} = 0, \tag{14.18}$$

where $y \in Y(t)$ and $d\mathcal{V}(\tau, V(\tau,\cdot))/dt$ is the complete Dini-type derivative of functional $\mathcal{V}(\tau, V(\tau,\cdot))$ along the trajectories of the evolution equation (14.9), (14.10). The previous equation actually reduces to

$$\min_u \max_v\{\frac{d\mathcal{V}(\tau, V(\tau,\cdot))}{dt} + d^2(u, \mathcal{P}(t)) - d^2(v, Q(t))\} = 0.$$

The detailed interpretation of the last equations will be the topic of a separate paper, being important for the analysis of systems when the original equations (14.1), (14.3) are nonlinear. In the linear case however the function $V(t,x)$ and the functional $\mathcal{V}(t, V(t,\cdot))$ may be calculated through the techniques of duality theory of nonlinear analysis.

14.5 Solution through Duality Techniques

In this section we demonstrate the application of duality techniques of nonlinear analysis ([15], [2]) to the problems of this paper, assuming $C(t) \equiv 0$. The more general case of $C(t) \neq 0$ may be treated along the lines of papers [6], [11].

Starting from Problem (E), we note that in our case function $V(t,x)$ may be found as

$$V(t,x) = \min\{d^2(x(t_0), X^0) + \\ + \int_{t_0}^{\tau} d^2(y^*(t) - H(t)x(t), \mathcal{R}(t))dt \ |x(t_0)\}, \quad x(\tau) = x. \tag{14.19}$$

This function may also be calculated as

$$\min\{d^2(x(t_0), X^0)|x(t_0)\} = \min\{\max_l\{(l, x(t_0)) - \rho(l|\mathcal{X}^0) - \frac{1}{4}(l,l)\} \mid x(t_0)\}$$

under restrictions

$$x(\tau) = x,$$
$$d^2(y^*(t) - H(t)x(t), \mathcal{R}(t))^2 \le 0, \quad t \in [t_0, \tau].$$

Solving this problem through the techniques of nonlinear analysis [5], [7], [15], [13], we have

$$V(t,x) = \max_l \sup_{\lambda(\cdot)}\{(s(t), x) +$$

$$+\int_{t_0}^{\tau}((\lambda(t),y^*(t))-(s(t),B(t)u^*(t))-\rho(\lambda(t)|\mathcal{R}(t))dt-$$

$$-\rho(l|\mathcal{X}^0)-\frac{1}{4}(l,l)-\frac{1}{4}\int_{t_0}^{t}(\lambda(t),\lambda(t))dt, \tag{14.20}$$

where the row $s(t)$ is the solution of the adjoint system for Problem (E)

$$\dot{s}=-sA(t)+\lambda(t)'H(t),\ \ s(t_0)=l.$$

Given $V(\tau,x)$, we may now calculate the support function

$$\rho(l|\mathcal{X}[\tau])=\max\{(l,x)|V(\tau,x)\le 0\}$$

of set $\mathcal{X}(\tau,\cdot)$.

Note that set $\mathcal{X}(t,\cdot)=\{x:V(t,x)\le 0\},\ \ t\in[t_0,\tau]$ evolves due to equation

$$\frac{\partial V}{\partial t}+(\frac{\partial V}{\partial x},A(t)x+B(t)u^*(t))+d^2(y^*(t)-H(t)x,\mathcal{R}(t))=0.$$

$$V(t_0,x)=d^2(x,\mathcal{X}^0),$$

where $u=u^*(t),y=y^*(t)$ are given realizations of the control u and the measurement y for the interval $[t_0,\tau]$.

We may now pass Problem (C) which is

$$\mathcal{V}(\tau,\mathcal{X}(\tau,\cdot))=\min_{U}\max_{x,w}\{\{d^2(x,\mathcal{M})|V(t_1,x)\le 0\}|U\in\mathcal{U}_V\}, \tag{14.21}$$

Here $\mathcal{U}_V$ - is the class of feedback control strategies $U(t,V(\cdot))$ mentioned in the previous Section. Recall that these strategies are convex compact-valued in $\mathbb{R}^p$, continuous in t and upper semicontinuous in $V(\cdot)\in\mathcal{L}_2[\mathcal{B}(\mu)]$. One may observe that by maximizing over x,w in (14.21) we have to maximize over *all possible future realizations* of the information function $V(t_1,x)$, or in other words, over all the possible future realizations of the information set $\mathcal{X}(t_1,\cdot)$.

For each given realization of $u(t),\ t\in[\tau,t_1]$ the possible information set $\mathcal{X}(t_1,\cdot)$ depends on $u(t)$ and also on the possible realization $y(t),\ t\in[\tau,t_1]$ which in its turn depends on the pair $x\in\mathcal{X}(\tau,\cdot),w(t)\in\mathcal{R}(t),t\in(\tau,t_1]$. In other words, with $u(t)$ given, each possible set $\mathcal{X}(t_1,\cdot)$ is the cross-section ("the cut") of trajectories $x(t,\tau,x)$ that satisfy *for a fixed* $u(t)$ the constraints

$$y(t)-H(t)x(t,\tau,x)\in\mathcal{R}(t),\ \ x\in\mathcal{X}(\tau,\cdot)$$

generated through $y(t)$ by some pair $x\in\mathcal{X}(\tau,\cdot),\xi(t)\in\mathcal{R}(t),t\in[\tau,t_1]$. Denoting all the possible future realizations of $y(t),\ t\in[\tau,t_1]$ as $\mathcal{Y}(t)$, we further denote $\mathcal{X}(\tau,\cdot)=\mathcal{X}(\tau,\cdot|u(\cdot),y(\cdot)))$ for all future realizations of set $\mathcal{X}(\tau,\cdot)$, emphasizing their dependence on $u(t),y(t),\ t\in(\tau,t_1]$.

Also denote

$$X(t_1,\tau,\mathcal{X}(\tau,\cdot)|u(\cdot)))=G(t_1,\tau)\mathcal{X}(\tau,\cdot)+\int_{\tau}^{t_1}G(t_1,t)B(t)u(t)dt,$$

where $G(t,\tau)$ is the transition matrix of the homogeneous equation (14.1).

A standard calculation indicates the next property.

LEMMA 14.1
The following equality is true:

$$X(t_1,\tau,\mathcal{X}(\tau,\cdot)|u(\cdot))) = \cup\{\mathcal{X}(\tau,\cdot|u(\cdot),y(\cdot)|\ y(t)\in Y(t), t\in(\tau,t_1]\}.$$

□

Therefore Problem (C) is now reduced to the minimization of the distance

$$\varepsilon(\tau,\mathcal{X}(\tau,\cdot)) = \min_u\{d^2(X(t_1,\tau,\mathcal{X}(\tau,\cdot)|u(\cdot))),\mathcal{M}) \tag{14.22}$$

over all the controls $u(t)\in\mathcal{P}(t),\ t\in(\tau,t_1]$. (In the absence of disturbance $v(t)$ the open and closed-loop solutions to this problem are the same).

Obviously we have

$$\varepsilon(\tau,\mathcal{X}(\tau,\cdot)) = \mathcal{V}(\tau,\mathcal{X}(\tau,\cdot)).$$

The latter value may now be calculated through duality techniques of nonlinear analysis. This gives

$$\mathcal{V}(\tau,\mathcal{X}(\tau,\cdot)) = \max\{-(-h^{**}(\tau,l) - \int_\tau^{t_1}\rho(-l|\mathcal{P}(t))dt\ |l\in\mathbb{R}^n\} \tag{14.23}$$

where

$$h(t,l) = \rho(l|\mathcal{X}(\tau,\cdot)|u(\cdot))) - \rho(l|\mathcal{M}) - \frac{1}{4}(l,l).$$

and the second conjugate $h^{**}(\tau,l)$ is taken only in the second variable.

THEOREM 14.2
The value functional $\mathcal{V}(\tau,\mathcal{X}(\tau,\cdot))$ is given by formula (14.23) as a result of a finite-dimensional optimization procedure. □

(The finite-dimensionality of the related optimization problem (14.23) also holds for the general case $C(t)\neq 0$).

The functional (14.23) produces a unique maximizer $l^0 = l^0(\tau,V(\tau,\cdot))$ which further yields a guaranteed solution strategy $U^0(\tau,V(\tau,\cdot))$ which belongs to class $\mathcal{U}_V$ described above.

THEOREM 14.3
The solution strategy for the basic problem is given by formula

$$U^0(\tau,V(\tau,\cdot)) = arg\min\{(l^0(\tau,V(\tau,\cdot)),B(t)u))|u\in\mathcal{P}(\tau)\},$$

where $l^0(\tau,V(\tau,\cdot)$ is the maximizer for problem (14.23), whilst $V(\tau,x)$ is given by formula (14.20).

Strategy $U^0(\tau,V(\tau,\cdot))$ ensures the inequality

$$d^2(x(t_1),\mathcal{M}) \leq \mathcal{V}(t_0,\mathcal{X}^0),$$

whatever be the disturbances $v(\cdot),w(\cdot)$ at the system inputs and the initial vector $x(t_0)=x^0$ subjected to constraints (14.2), (14.4), (14.5). □

The numerical realization of the presented scheme may be achieved through the techniques of ellipsoidal calculus which effectively allows to deal with set-valued functions along the lines of publications [8], [9], [10]. But this a subject for a separate publication.

14.6 References

[1] Basar T., Bernhard P.,(1995)H^∞ *Optimal Control and Related Minmax Design Problems*, Birkhäuser, Boston, 2-nd. ed.

[2] Clarke F.H., Ledyaev Yu.S., Stern R.J., Wolenski P.R.,(1998) *Nonsmooth Analysis and Control Theory*, Springer-Verlag.

[3] Fleming W.H., Soner H.M., (1993)*Controlled Markov Processes and Viscosity Solutions*, *Springer, NY*.

[4] James M.R., Baras J.S. (1996) Partially onserved differential games, infinite-dimensional HJB equations and nonlinear H^∞ control. *SIAM Jounal on Control and Optimization*, v.34, N4, pp.1342-1364.

[5] Kurzhanski A.B. (1977) *Control and Observation Under Uncertainty* Nauka, Moscow, (in Russian).

[6] Kurzhanski A.B. (1999) Pontryagin's Alternated Integral in the Theory of Control Synthesis. *Proc.Steklov Math. Inst.*, v.224, (Engl.Transl.) pp.234-248.

[7] Kurzhanski A.B., Fillippova T.F., (1993) On the Theory of Trajectory Tubes: a Mathematical Formalism for Uncertain Dynamics, Viability and Control, *Advances in Nonlinear Dynamics and Control, ser. PSCT 17, Birkhäuser, Boston*,pp. 122 - 188.

[8] Kurzhanski A.B., Vályi I. (1997) *Ellipsoidal Calculus for Estimation and Control*, SCFA, Birkhäuser. Boston.

[9] Kurzhanski A.B., Varaiya P. (2001) Optimization Techniques for Reachability Analysis. In:*JOTA, N2*.

[10] Kurzhanski A.B., Varaiya P. (2002) Ellipsoidal Techniques for Reachability Analysis: Part I : External Approximations. Part II: Internal approximationa. Box-valued Constraints. In: *Optimization. Methods and Software*. Taylor and Francis, to appear in 2002.

[11] Kurzhanski A.B., Varaiya P. (2002) Reachability Analysis for Uncertain Systems - the Ellipsoidal Technique. *Dynamics of Continuous, Discrete and Impulsive Systems, ser.B*, v.9, N.3, pp.347-367.

[12] Kurzhanski A.B., Varaiya P. (2002) Reachability under State Constraints - the Ellipsoidal Technique. In: *Proc. of the IFAC-2002 World Congress, Barcelona, Spain*.

[13] Lindquist A., Byrnes C.I.,(2002) A Convex Optimization Approach to the Generalized Moment Problem. In: *Proc. of the MTNS-2002 Conf., Notre Dame, USA*

[14] Milanese M., et al.eds(1995) *Bounding Approach to System Identification*, Plenum Press.

[15] Rockafellar R.T., Wets R.J.B.(1998) *Variational Analysis* Springer-Verlag.

[16] Subbotin A.I.(1995) *Generalized Solutions of First-order PDE's*, SCFA, Birkhäuser. Boston.

15

Linear System Identification as Curve Fitting

Lennart Ljung

Abstract

The purpose of this contribution is to point out and exploit the kinship between identification of linear dynamic systems and classical curve fitting. For curve fitting we discuss both global and local parametric methods as well as non-parametric ones, such as local polynomial methods. We view system identification as the estimation of the frequency function curve. The empirical transfer function estimate is taken as the "observations" of this curve. In particular we discuss how this could be done for multi-variable systems. Local and non-parametric curve fitting methods lead to variants of classical spectral analysis, while the prediction error/maximum likelihood framework corresponds to global parametric methods. The role of the noise model in dynamic models is also illuminated from this perspective.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 203-215, 2003.
© Springer-Verlag Berlin Heidelberg 2003

15.1 Introduction

A linear dynamic system in discrete or continuous time can be described by

$$y(t) = G(\sigma)u(t) + v(t) \tag{15.1}$$

where σ is the differentiation operator p in continuous time and the shift operator q in discrete time. The identification problem is to find the transfer operator G and possibly also the spectrum of the additive noise v. There is an extensive literature on this problem, see among many books, e.g. [7] and [11].

One may distinguish three main approaches to this problem:

- *Spectral Analysis* which forms estimates of the spectra of the output, the input and the cross spectrum between input and output, $\Phi_y(\omega), \Phi_u(\omega)$, $\Phi_{yu}(\omega)$ and then the estimates

$$\hat{G}(i\omega) = \frac{\Phi_{yu}(\omega)}{\Phi_u(\omega)} \tag{15.2}$$

$$\hat{\Phi}_v(\omega) = \Phi_y(\omega) - \hat{G}(i\omega)\Phi_u(\omega) \tag{15.3}$$

  See, e.g. [5], [1] for thorough treatments of this approach.

- *Parametric Methods* that explicitly parameterize the sought transfer functions and estimate the parameters by maximum likelihood or related techniques.

- *Subspace Methods* that conceptually can be described as model reduction schemes of simple high order estimates. See, e.g. [14], [6] or [15].

It is the purpose of this contribution to put such identification methods into the perspective of simple-minded curve fitting techniques. To put it another way: We are not so far from Gauss' basic contributions on least squares.

15.2 Curve Fitting

To bring out the basic features of the estimation problem, let us study a simple example. Suppose the problem is to estimate an unknown function $g_0(x), -1 \leq x \leq 1$. The observations we have are noise measurements $y(k)$ at points x_k which we may or may not choose ourselves:

$$y(k) = g_0(x_k) + e(k) \tag{15.4}$$

How to approach this problem?

Parametric Methods

Global Parameterizations One way or another we must decide "where to look for" g. We could, for example, have the information that g is a third order polynomial. This would lead to the – in this case – grey box model structure

$$g(x, \theta) = \theta_1 + \theta_2 x + \theta_3 x^2 + \ldots + \theta_n x^{n-1} \tag{15.5}$$

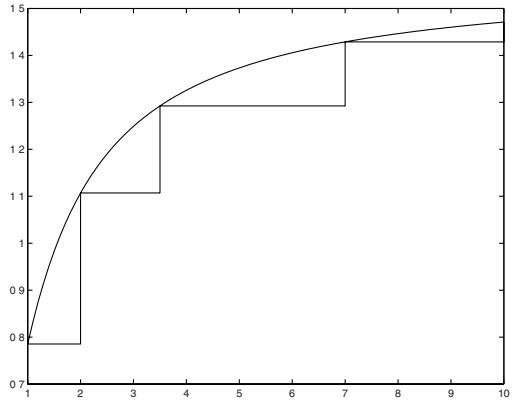


Figure 15.1 A piece-wise constant approximation.

with $n = 4$, and we would estimate the parameter vector θ from the observations y, using for example the classical least squares method.

Now suppose that we have no structural information at all about g. We would then still have to assume something about it, e.g. it is an analytical function, or that it is piecewise constant or something like that. In this situation, we could still use (15.5), but now as *black-box* model: if we assume g to be analytic we know that it can be approximated arbitrarily well by a polynomial. The necessary order n would not be known, and we would have to find a good value of it using some suitable scheme.

Note that there are several alternatives in this black-box situation: We could use rational approximations:

$$g(x,\theta) = \frac{\theta_1 + \theta_2 x + \theta_3 x^2 + \ldots + \theta_n x^{n-1}}{1 + \theta_{n+1} x + \theta_{n+2} x^2 + \ldots + \theta_{n+m-1} x^{m-1}} \tag{15.6}$$

or Fourier series expansions

$$g(x,\theta) = \theta_0 + \sum_{\ell=1}^{n} \theta_{2\ell-1} \cos(\ell \pi x) + \theta_{2\ell} \sin(\ell \pi x) \tag{15.7}$$

Local Parameterizations Alternatively, we could approximate the function by piece-wise constant functions, as illustrated in Figure 15.1. The mapping g can be parameterized as a function expansion

$$g(x,\theta) = \sum_{k=1}^{n} \alpha_k \kappa(\beta_k (x - \gamma_k)) \tag{15.8}$$

Here, κ is a "mother basis function", from which the actual functions in the function expansion are created by *dilation* (parameter β) and *translation* (parameter γ). For example, with $\kappa = \cos$ we would get Fourier series expansion with β as frequency and γ as phase. More common are cases where κ is a unit pulse. With

that choice, (15.8) can describe any piecewise constant function, where the granularity of the approximation is governed by the dilation parameter β. Compared to Figure 15.1 we would in that case have

- $n = 4$,
- $\gamma_1 = 1, \gamma_2 = 2, \gamma_3 = 3.5, \gamma_4 = 7$,
- $\beta_1 = 1, \beta_2 = 2/3, \beta_3 = 1/3.5, \beta_4 = 1/3$
- $\alpha_1 = 0.79, \alpha_2 = 1.1, \alpha_3 = 1.3, \alpha_4 = 1.43$

A related choice is a soft version of a unit pulse, such as the Gaussian bell. Alternatively, κ could be a unit step (which also gives piecewise constant functions), or a soft step, such as the sigmoid.

Estimation Techniques and Basic Properties

It suggests itself that the basic least-squares like approach is a natural approach for curve fitting:

$$\begin{aligned} \hat{\theta}_N &= \arg\min_{\theta} V_N(\theta) \\ V_N(\theta) &= \sum_{k=1}^{N} \mu_k (y(k) - g(x_k, \theta))^2 \end{aligned} \tag{15.9}$$

Here μ_k is a weight that in a suitable way reflects the

- "reliability" of the measurement k. This is typically evaluated as the variance of $e(k)$, so we would have $\mu_k \sim 1/\lambda_k$, where λ_k is the variance of $e(k)$.
- "relevance" of the measurement k. It could be that we do not fully believe that the underlying model $g(x, \theta)$ is capable of describing the data for all x. We could then downweigh a measurement at a point x_k outside a region of prime relevance for the model.

In case y and g_0 are vectorvalued (column vectors), the criterion takes the form

$$V_N(\theta) = \sum_{k=1}^{N} (y(k) - g(x_k, \theta))^T \Lambda_k^{-1} (y(k) - g(x_k, \theta) \tag{15.10}$$

where the matrix Λ_k takes care of the weightings. For the reliablity aspect, Λ_k would be the covariance matrix of $e(k)$.

Non-Parametric Methods

Basic Idea: Local Smoothing A simple idea to form an estimate of the function value $g(x)$ at a point x is to form some kind of average of the observations $y(k)$ corresponding to x_k in the neigbourhood of x:

$$\hat{g}(x) = \sum_{k=1}^{N} W(x, x_k) y(k) \tag{15.11}$$

where the weights W are chosen appropriately, and typically being zero when the distance between x and x_k is larger than a certain value ("the bandwidth"). The choice of such weights is the subject of an extensive literature in statistics. See, among many references, e.g. [3], [13], [2], and [9]. It is not the purpose of this paper to give an overview of such methods, but we can point to some basic choices.

Nearest Neighbor Smoothing Perhaps the simplest idea is to take as an estimate of $\hat{g}$ the observed value $y(k)$ at the nearest observation point. This corresponds to choosing $W(x, x_k)$ to be 1 for that x_k in the observation set that is closest to x and 0 otherwise.

Kernel Smoothers Take

$$W(x, z) = \kappa(|x - z|/h) \tag{15.12}$$

where $\kappa(\xi)$ is some bell-shaped function that is zero for $\xi > 1$. A common choice is the *Epanechnikov kernel*

$$\kappa(x) = \begin{cases} 1 - x^2 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases} \tag{15.13}$$

Local Polynomial Methods Polynomial approximations of a curve may be efficient, at least locally. Instead of the global model (15.5) one may seek a local approximation in the neigbourhood of x by

$$g(z, \theta) = \theta_1 + \theta_2(z - x) + \theta_3(z - x)^2 + \ldots \theta_n(z - x)^{n-1} \tag{15.14}$$

Then estimate these coefficients e.g. by weighted least squares

$$\min \sum_{k=1}^{N} \mu_k \cdot \tilde{W}(x - x_k) \cdot |y(k) - \theta_1 - \theta_2(x_k - x) - \ldots - \theta_n(x_k - x)^{n-1}|^2 \tag{15.15}$$

Here μ_k is a weight that reflects the reliability of the k:th measurement (the variance of $e(k)$) and $\tilde{W}$ concentrates the fit to a suitable neighbourhood of x. The resulting estimate $\hat{\theta}$ will be linear in $y(k)$ and choosing $\hat{\theta}_1$ as an estimate for $g(x)$ will give an expression of the type (15.11).

Direct Optimization of Weights In a recent approach, [10], the choice of weights W in (15.11) is formulated as a min-max optimization problem for the mean square error $\mathrm{E}|g(x) - \hat{g}(x)|^2$. This is either a quadratic programming problem or a convex one and can be solved quite efficiently to achieve optimal weights without relying upon asymptotic theory.

15.3 Linear Dynamic Models

The Curve: The Frequency Function

A linear dynamic system is uniquely defined, e.g., by its impulse response or by its *Frequency Function*

$$G(e^{i\omega T}) \quad \text{or} \quad G(i\omega) \tag{15.16}$$

in discrete time, or continuous time, respectivly. For simplicity in the sequel we will use the continuous time notation.

Therefore it is possible to interpret all methods for linear system identification as methods to estimate the frequency function curve.

The Observations: The ETFE

The primary observations from a dynamical system are always the sequences of sampled inputs and outputs, u and y,

$$Z^N = \{u(t_1), y(t_1), \ldots, u(t_N), y(t_N)\} \tag{15.17}$$

From these we may form the Fourier transforms

$$U_N(\omega) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} u(t_k) e^{i\omega(k-1)T} \tag{15.18}$$

$$Y_N(\omega) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} y(t_k) e^{i\omega(k-1)T} \tag{15.19}$$

(These expressions give the DFT for equidistantly sampled data: $t_{k+1} - t_k \equiv T$, but several variants can be envisioned.)

For a scalar input we may now form *the Empirical Transfer Function Estimate, ETFE* as

$$\hat{\hat{G}}_N(i\omega) = \frac{Y_N(\omega)}{U_N(\omega)} \tag{15.20}$$

In case the observations y and u have been obtained from a noise-corrupted linear system with frequency function $G_0(i\omega)$ it can be shown that the ETFE has the following statistical properties: (Lemma 6.1 in [7].)

$$\mathrm{E}\hat{\hat{G}}_N(i\omega) = G_0(i\omega) + \frac{\rho_1}{\sqrt{N} U_N(\omega)} \tag{15.21}$$

$$\mathrm{E}|\hat{\hat{G}}_N(i\omega) - G_0(i\omega)|^2 = \frac{\Phi_v(\omega)}{|U_N(\omega)|^2} + \frac{\rho_2}{N|U_N(\omega)|^2} \tag{15.22}$$

Here $\Phi_v(\omega)$ is the spectrum of the additive noise (at the output of the system)and ρ_i are constant bounds that depend on the impulse response of the system, the bound on the input, and the covariance function of the noise. Moreover, it can be shown that the EFTE's are asymptotically uncorrelated at frequencies on the DFT grid.

All this means that we can think of the ETFE as a "noisy measurement" of the frequency function:

$$\hat{\hat{G}}_N(i\omega_k) = G_0(i\omega_k) + v_k \tag{15.23}$$

with v_k being a zero mean random variable with variance $\Phi_v(\omega_k)/|U_N(\omega_k)|^2$. We have then ignored the terms with ρ in the expressions above.

Something must also be said about the frequency grid in (15.23): If the Fourier transforms are obtained by DFT of equidistantly sampled data, the natural frequencies to use in (15.23) are the DFT grid:

$$\omega_k = k\pi/N; \qquad k = 0, \ldots, N-1 \tag{15.24}$$

This gives two advantages:

- Frequencies in between these carry no extra information: they are merely (trigonometric) interpolations of the values on the DFT grid. This also determines the maximum frequency resolution of the frequency function.
- v_k are (asymptotically) uncorrelated on this grid.

In the case of p outputs, v_k is a column vector and Φ_v is a $p \times p$ matrix.

The Multi-input Case

When there are several inputs, so that u is an m-dimensional column vector and G is a $p \times m$ matrix, there is no unique way of forming the ETFE, and this has apparently not been discussed in the literature. One of the possibilities is to split the data record into m parts, like (assume $N = mM$ and $t_{k+1} - t_k \equiv 1$ for simplicity)

$$U_N^{(r)}(\omega) = \frac{1}{\sqrt{M}} \sum_{k=1}^{M} u((r-1)M + k) e^{i\omega(k-1)}; \qquad r = 1, \ldots, m \tag{15.25}$$

and similarly for $Y_N^{(r)}$. The corresponding DFT-grid for ω will be reduced by a factor m to

$$\omega_\ell = \ell\pi/M; \ell = 0, \ldots, M-1 \tag{15.26}$$

On this grid we can define

$$\begin{aligned} \hat{\hat{G}}_N(i\omega) &= \left[Y_N^{(1)}(\omega) \quad \ldots \quad Y_N^{(m)}(\omega) \right] \left[U_N^{(1)}(\omega) \quad \ldots \quad U_N^{(m)}(\omega) \right]^{-1} \\ &= \mathcal{Y}_N(\omega) \mathcal{U}_N(\omega)^{-1} \end{aligned} \tag{15.27}$$

provided the $m \times m$ inverse exists (which is the generic case). A related possibility is to take DFTs of the whole data record and form the estimate using m-tuples of neighboring frequencies.

It can be shown, analogously to the single input case that

$$\hat{\hat{G}}_N(i\omega_k) = G_0(i\omega_k) + v_k \tag{15.28}$$

where now v_k is a sequence of $p \times m$ matrices with (asymptotically) zero means and asymptotically uncorrelated on the DFT-grid (15.26) with covariance matrix

$$[\mathcal{U}_N(\omega) \mathcal{U}_N^*(\omega)]^{-1} \otimes \Phi_v(\omega) \tag{15.29}$$

for $\mathrm{vec}(v_k)$. Here $\otimes$ denotes the Kronecker product, and vec means stacking the columns of a matrix on top of each other.

15.4 Fitting the Frequency Function Curve by Local Methods

The basic relation between the Fourier data and the frequency function is

$$Y(\omega_k) = G(i\omega_k)U(\omega_k) + V(\omega_k), \qquad k = 1, \ldots, N \tag{15.30}$$

This holds in the general multi-variable case. The covariance matrix for $V(\omega)$ is the $m \times m$ spectrum matrix $\Phi(\omega)$. For compact notation we form the $p \times N$ matrix $\mathcal{Y}$ as

$$\mathcal{Y} = [\, Y(\omega_1), \ldots, Y(\omega_N)\,] \tag{15.31}$$

and similarly the $m \times N$ matrix $\mathcal{U}$ and the $p \times N$ matrix $\mathcal{V}$. Then, if G were a constant (complex valued) $p \times m$ matrix (15.30) could be rewritten

$$\mathcal{Y} = G\mathcal{U} + \mathcal{V} \tag{15.32}$$

This could be compared to the local polynomial approach (15.14) where we have only taken the constant term ($n = 1$). To estimate this constant G we would apply weighted least squares (15.15) where the weighting function $\tilde{W}(\omega_k - \omega)$ would measure the distance between the value ω, where we seek an estimate of G and the frequencies ω_k where we have the observations. Form the $N \times N$ matrix $\mathcal{W}$ as the diagonal, real-valued matrix of these weights. Then the weighted least squares estimate of G is

$$\hat{G} = \mathcal{Y}\mathcal{W}\mathcal{U}^*\,[\,\mathcal{U}\mathcal{W}\mathcal{U}^*\,]^{-1} \tag{15.33}$$

This is the estimate at frequency ω and the dependence on ω in this expression is hidden in $\mathcal{W}$.

To estimate the disturbance spectrum $\Phi_v(\omega)$ we estimate $\mathcal{V}$ by

$$\hat{\mathcal{V}} = \mathcal{Y} - \hat{G}\mathcal{U} = \mathcal{Y}\,(I - \mathcal{W}\mathcal{U}^*\,[\,\mathcal{U}\mathcal{W}\mathcal{U}^*\,]^{-1}\,\mathcal{U}) = \mathcal{Y}P_u \tag{15.34}$$

where the last step is a definition of P_u. Note that

$$P_u\mathcal{W}P_u^* = P_u\mathcal{W} \tag{15.35}$$

A natural estimate of the spectrum Φ_v is to form a weighted sum

$$\hat{\Phi}_v = \frac{1}{\rho}\hat{\mathcal{V}}\mathcal{W}\hat{\mathcal{V}}^* \tag{15.36}$$

The question is, what should the normalization factor ρ be? Consider the i,j element of the matrix above and note that $\mathcal{Y}P_u = \mathcal{V}P_u$. Thus

$$\mathrm{E}\hat{\Phi}_v^{i,j} = \frac{1}{\rho}\mathcal{V}_i P_u \mathcal{W} P_u^* \mathcal{V}_j^* = \frac{1}{\rho}\,\mathrm{tr}\mathrm{E}\mathcal{V}_j^*\mathcal{V}_i P_u \mathcal{W}$$

Assuming Φ_v to be constant (at least over the interval covered by the weights), and that the frequency grid is such that $v(\omega_k)$ are uncorrelated (recall the comment below (15.24)) we have

$$\mathrm{E}\mathcal{V}_j^*\mathcal{V}_i = \Phi_v^{i,j} \cdot I_{N\times N}$$

with the $N \times N$ identity matrix. Moreover

$$\eta = \mathrm{tr}P_u\mathcal{W} = \mathrm{tr}\mathcal{W} - \mathrm{tr}(\mathcal{U}\mathcal{W}\mathcal{U}^*)^{-1}\mathcal{U}\mathcal{W}^2\mathcal{U}^* \tag{15.37}$$

This shows that the correct normalization in (15.36) is $\rho = \eta$ (Note that in case equal weigths are used, that is, $\mathcal{W} = I$, then the normalization becomes the familiar $\rho = N - m$.)

This way (15.33) of estimating the frequency function is closely related to classical spectral analysis, since $\mathcal{U}\mathcal{W}\mathcal{U}^*$ can be seen as an estimate of the input spectral matrix and $\mathcal{Y}\mathcal{W}\mathcal{U}^*$ is an estimate of the cross spectrum. (See also Chapter 6 in [7].) The weights in $\mathcal{W}$ then correspond to the frequency (or tapering) windows, like the Bartlett, Parzen or Hamming windows typically applied in spectral analysis. Displaying the kinship to local polynomial modeling in curve fitting gives extra insight into the role of these windows. It also shows how to find the right normalization for unbiased estimation of the additive disturbance spectrum, which is important for narrow frequency windows. It may be easy to overlook the second term of (15.37).

Nothing in this treatment requires that the window $\mathcal{W}$ is the same for all frequencies. On the contrary, it is natural to let the bandwidth depend on the actual frequency where G is estimated. This is how the function `spafdr` (spectral analysis with frequency dependent resolution) in the SYSTEM IDENTIFICATION TOOLBOX, [8], is implemented. A related approach to smoothing the ETFE is described in [12].

15.5 Fitting the Frequency Function by Parametric Methods

The Model Structure

A model structure for a linear system is simply a parameterization of the frequency function

$$G(i\omega, \theta) \tag{15.38}$$

possibly together with a parameterization of the additive noise spectrum

$$\Phi_v(\omega, \theta) = H(i\omega, \theta)\Lambda(\theta)H^*(i\omega, \theta) \tag{15.39}$$

where the second step shows the spectrum factorized using a monic, stable and inversely stable transfer function H.

The actual parameterization can be done in many different ways. The underlying description could be a discrete time ARMAX model

$$A(q)y(t) = B(q)u(t) + C(q)e(t)$$

with the coefficients of the polynomials (in q^{-1}) A, B and C comprising θ. This gives

$$G(e^{i\omega}, \theta) = \frac{B(e^{i\omega})}{A(e^{i\omega})}$$
$$H(e^{i\omega}, \theta) = \frac{C(e^{i\omega})}{A(e^{i\omega})}$$

Note the similarity with the basic forms (15.5) and (15.6) for $x = e^{-i\omega}$.

A physically parameterized state space model

$$\dot{x}(t) = A(\theta)x(t) + B(\theta)u(t) + w(t); \qquad \mathrm{E}w(t)w^T(s) = Q(\theta)\delta(t-s)$$
$$y(t) = C(\theta)x(t) + D(\theta)u(t) + e(t); \qquad \mathrm{E}e(t)e^T(s) = R(\theta)\delta(t-s)$$

corresponds to

$$G(i\omega, \theta) = C(\theta)(i\omega I - A(\theta))^{-1}B(\theta) + D(\theta)$$
$$H(i\omega, \theta) = C(\theta)(i\omega I - A(\theta))^{-1}K(\theta) + I$$

where $K(\theta)$ and $\Lambda(\theta)$ are computed from A, C, Q and R as the steady state Kalman filter's gain and innovations variance.

Many other types of parameterizations are of course possible.

Frequency Domain Data

Assume that the data are given in the frequency domain. We can then form the ETFE as described in Sections 15.3 and 15.3. The weighted least squares fit between the ETFE and the parameterized curve is then obtained as in (15.10):

$$V(\theta) = \sum_{\ell=1}^{M} \mathrm{vec}(\hat{\hat{G}}(i\omega_\ell) - G(i\omega_\ell, \theta))^* \{[\mathcal{U}_N \mathcal{U}_N^*] \otimes \Phi_v^{-1}(\omega_\ell)\} \tag{15.40}$$
$$\times \mathrm{vec}(\hat{\hat{G}}(i\omega_\ell) - G(i\omega_\ell, \theta)) \tag{15.41}$$

where the frequencies ω_ℓ are from the grid (15.26). Here we have first formed a column vector from the matrix $\hat{\hat{G}} - G$ and then weighted with the inverse covariance matrix, (15.29), of the measurement noise at the frequency in question. (Note that $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.) Applying the formula

$$\mathrm{vec}(D^*)^*(C^* \otimes A)\mathrm{vec}(B) = \mathrm{tr}(ABCD)$$

now gives

$$V(\theta)=\sum_{\ell=1}^{M}\mathrm{tr}[\hat{\hat{G}}(i\omega_\ell)-G(i\omega_\ell,\theta)][\mathcal{U}_N\mathcal{U}_N^*][\hat{\hat{G}}(i\omega_\ell)-G(i\omega_\ell,\theta)]^*\Phi_v^{-1}(\omega_\ell) \quad (15.42)$$

which in view of (15.27) also can be written

$$V(\theta)=\sum_{\ell=1}^{M}\mathrm{tr}[\mathcal{Y}_N(\omega_\ell)-G(i\omega_\ell,\theta)\mathcal{U}_N(\omega_\ell)]^*\Phi_v^{-1}(\omega_\ell)[\mathcal{Y}_N(\omega_\ell)-G(i\omega_\ell,\theta)\mathcal{U}_N(\omega_\ell)] \quad (15.43)$$

The expression within bracket is a $p\times m$ matrix which means that we can go back to the original vectors $U^{(r)}(\omega_\ell)$ and $Y^{(r)}(\omega_\ell)$ in (15.25) to obtain

$$\begin{aligned}V(\theta)=&\sum_{\ell=1}^{M}\sum_{r=1}^{m}[Y_N^{(r)}(\omega_\ell)-G(i\omega_\ell,\theta)U_N^{(r)}(\omega_\ell)]^*\Phi_v^{-1}(\omega_\ell)\\&\times[Y_N^{(r)}(\omega_\ell)-G(i\omega_\ell,\theta)U_N^{(r)}(\omega_\ell)]\end{aligned} \quad (15.44)$$

Here we have an m-fold repetition of frequencies on the courser grid (15.26). By a small approximation we can move to the finer grid (15.24) and obtain

$$V(\theta)=\sum_{k=1}^{N}[Y_N(\omega_k-G(i\omega_k,\theta)U_N(\omega_k)]^*\Phi_v^{-1}(\omega_k)[Y_N(\omega_k)-G(i\omega_k,\theta)U_N(\omega_k)] \quad (15.45)$$

Prediction Errors From Time Domain Data

If we start with time domain data (15.17) we could of course directly transform to frequency domain data and go through steps of the previous subsection. It is instructive to follow the calculations directly in the time domain.

The discrete time domain version of the model (15.38), (15.39) is

$$y(t)=G(q,\theta)u(t)+H(q,\theta)e(t) \quad (15.46)$$

For equidistantly sampled data we can form the prediction errors for this model as

$$\varepsilon(t,\theta)=H^{-1}(q,\theta)(y(t)-G(q,\theta)u(t)) \quad (15.47)$$

Assuming that these have covariance matrix Λ, a natural criterion to minimize (which equals the maximum likelihood criterion if Λ is known and e is Gaussian) is

$$V(\theta)=\sum_{k=1}^{N}\varepsilon^T(t,\theta)\Lambda^{-1}\varepsilon(t,\theta) \quad (15.48)$$

Applying Parseval's relationship to (15.48), (15.47) and ignoring transient effects (or assuming periodic data) now gives for this criterion

$$\begin{aligned} V(\theta) &= \sum_{k=1}^{N} [Y_N(\omega_k) - G(i\omega_k,\theta)U_N(\omega_k)]^* \Phi_v^{-1}(\omega_k,\theta) \\ &\quad \times [Y_N(\omega_k) - G(i\omega_k,\theta)U_N(\omega_k)] \\ \Phi_v(\omega,\theta) &= H(i\omega,\theta)\Lambda H^*(i\omega,\theta) \end{aligned} \tag{15.49}$$

This is the same as (15.45) and we can now track back to the curve fitting expression between the ETFE and the parameterized curve in (15.40). Even in the time domain, multi-variable case the basic methods consequently still are curve fitting. We have also displayed the nature of the noise model in (15.46): It just provides the weighting in this fit.

15.6 Conclusions

Phrasing standard methods for linear system identification as curve fitting brings out several common features and gives some additional insight. It also shows that the bottom line in identification is quite simple and relies upon early work in statistics.

The subspace methods were not explicitly discussed in this contribution. The link to curve fitting is conceptually as follows: The first step in subspace methods is to form a big Hankel matrix from observed data. This can be seen as a high order ARX model for the system. The second step is to approximate this matrix with a low rank one using SVD. This is basically related to Hankel norm approximation of high order systems by lower order ones, which in turn can be interpreted as approximating the corresponding frequency functions, see [4].

15.7 References

[1] D.R. Brillinger. *Time Series: Data Analysis and Theory*. Holden-Day, San Francisco, 1981.

[2] W.S. Cleveland and E. Grosse. Computational methods for local regression. *Statistics and Computing*, 1:47–62, 1991.

[3] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Number 66 in Monographs on Statistics and Applied Probability. Chapman & Hall, 1996.

[4] K. Glover. All optimal Hankel norm approximations of linear multivariable systems and their L-infinity error bounds. *Int. Journal of Control*, 39:1115–1193, 1984.

[5] G.M. Jenkins and D.G. Watts. *Spectral Analysis*. Holden-Day, San Francisco, 1968.

[6] W. E. Larimore. Canonical variate analysis in identification, filtering and adaptive control. In *Proc. 29th IEEE Conference on Decision and Control*, pages 596–604, Honolulu, Hawaii, December 1990.

[7] L. Ljung. *System Identification - Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., 2nd edition, 1999.

[8] L. Ljung. *System Identification Toolbox for use with* MATLAB. *Version 6*. The MathWorks, Inc, Natick, MA, 6th edition, 2002.

[9] E. Nadaraya. On estimating regression. *Theory of Prob. and Applic.*, 9:141–142, 1964.

[10] J. Roll, A. Nazin, and L. Ljung. A non-asymptotic approach to local modelling. In *Proc. IEEE Conf. on Decision and Control*, Las Vegas, NV, Dec 2002. To appear.

[11] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall Int., London, 1989.

[12] A. Stenman, F. Gustafsson, D.E. Rivera, L. Ljung, and T. McKelvey. On adaptive smoothing of empirical transfer function estimates. *Control Engineering Practice*, 8:1309–1315, Nov 2000.

[13] C.J. Stone. Consistent non-parametric regression (with discussion). *Ann. Statist.*, 5:595–645, 1977.

[14] P. Van Overschee and B. DeMoor. *Subspace Identification of Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, 1996.

[15] M. Verhaegen. Identification of the deterministic part of MIMO state space models, given in innovations form from input-output data. *Automatica*, 30(1):61–74, January 1994.

16

Optimal Model Order Reduction for Maximal Real Part Norms

A. Megretski

Abstract

An observation is made that a polynomial time algorithm exists for the problem of optimal model order reduction of finite order systems in the case when the approximation error measure to be minimized is defined as maximum of the real part of the model mismatch transfer function over a certain set of frequencies. Applications to H-infinity model reduction and comparison to the classical Hankel model reduction are discussed.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 217-226, 2003.
© Springer-Verlag Berlin Heidelberg 2003

16.1 Introduction

This paper deals with problems of model order reduction for linear time-invariant (LTI) systems. Reduced order transfer functions are frequently used in modeling, design, and computer simulation of complex engineering systems. Despite significant research efforts, several fundamental questions concerning LTI model reduction remain unsolved.

A mathematical formulation of a model reduction problem can be given in terms of finding a stable transfer function $\hat{G}$ (the *reduced model*) of order less than k such that $\|G-\hat{G}\|$ (the *approximation error measure* quantifying the size of the *model mismatch* $\Delta = G - \hat{G}$) is minimal. Here G is a given stable LTI system (the *non-reduced model*), and $\|\cdot\|$ is a given norm (or sometimes a seminorm) on the vector space of stable transfer functions. While G is not rational in many applications, it is usually reasonable to assume that a high order high quality finite order approximation G_0 of G is available, and therefore the optimal model reduction problem is formulated as

$$\|G_0 - \hat{G}\| \to \min,\ (\mathrm{order}(\hat{G}) < k), \tag{16.1}$$

entirely in terms of finite order transfer functions.

Based on the wisdom of modern robust control, the most desirable norms $\|\cdot\|$ to be used in optimal model reduction are the so-called *weighted H-infinity norms*

$$\|\Delta\| = \|W\Delta\|_\infty, \tag{16.2}$$

where W is a given rational transfer function. However, to the author's knowledge, no polynomial time algorithms are known for solving (16.1) in the case of a non-zero weighted H-infinity norm $\|\cdot\|$ (even when $W \equiv 1$). (It is also not known whether the problem is NP-hard or not.)

The case when

$$\|\Delta\| = \|\Delta\|_h = \min_\delta \|\Delta + \bar{\delta}\|_\infty \tag{16.3}$$

(δ ranges over the set of stable transfer functions) is the so-called *Hankel norm* appears to be the only situation in which a polynomial time algorithm for solving (16.1) is commonly known. The theory of *Hankel model reduction* is the main concentration of rigorous results on model reduction. Since

$$\|\Delta\|_h \le \|\Delta\|_\infty \tag{16.4}$$

for every stable transfer function Δ, solving the Hankel model reduction problem provides a *lower bound* in the (unweighted) H-infinity model reduction problem. In addition, there is some evidence, both formal and experimental, that, for "reasonable" systems, the H-infinity model matching error delivered by the Hankel optimal reduced model is not much larger than the optimal Hankel model reduction error.

The positive statements concerning Hankel model reduction and its relation to H-infinity model reduction do not cover the case of weighted Hankel norms

$$\|\Delta\| = \|D\|_{h|W} = \min_\delta \|W(\Delta + \bar{\delta})\|_\infty.$$

The theory also does not extend to the case of G being defined by a finite number of frequency samples. The main point of this paper is that an alternative class of system norms $\|\cdot\|$, called *weighted maximal real part* norms, yields most of the good properties known of the Hankel model reduction, while providing the extra benefits of using sampled data and freqency weighted modeling error measures. For the weighted maximal real part model reduction, the paper provides a polynomial time optimization algorithm, and states a number of results concerning its relation to H-infinity and Hankel model reduction. Outcomes of some numerical experiments are also presented.

16.2 Maximal Real Part Model Reduction

For convenience, model reduction of *discrete-time systems* will be considered. Thus, a *stable transfer function* will be defined as a continuous function $f: \mathbf{T} \to \mathbf{C}$ for which the Fourier coefficients

$$\hat{f}[n] = \int_{-\pi}^{\pi} f(e^{jt}) e^{-jnt} dt$$

are all real and satisfy the condition

$$\hat{f}[n] = 0 \quad \forall\ n > 0.$$

Here $\mathbf{C}$ is the set of all complex numbers, and

$$\mathbf{T} = \{z \in \mathbf{C}:\ |z| = 1\}$$

is the unit circle centered at $z = 0$. The set of all stable transfer functions will be denoted by $\mathbf{A}$. The set of all rational stable transfer functions of order less than k will be denoted by $\mathbf{A}_k$.

The Unsampled Setup

The unsampled version of the *maximal real part model reduction problem* is defined as the task of finding $\hat{G} \in \mathbf{A}_k$ which minimizes $\|G_0 - \hat{G}\|_{r|W}$, where $W = |H|^2$,

$$\|\Delta\|_{r|W} = \|W\mathrm{Re}(\Delta)\|_\infty,$$

$G_0, H \in \mathbf{A}$ are given rational transfer functions, and

$$\|f\|_\infty = \max_{z \in \mathbf{T}} |f(z)|$$

for every continuous function $f: \mathbf{T} \to \mathbf{C}$.

It is easy to see that

$$0.5\|\Delta\|_{h|W} \le \|\Delta\|_{r|W} \le \|W\Delta\|_\infty \quad \forall\ \Delta \in \mathbf{A},$$

where the first inequality takes place because of

$$\|D\|_{h|W} = \min_{\delta} \|W(\Delta + \bar{\delta})\|_\infty \le \|W(\Delta + \bar{\Delta})\|_\infty = 2\|\Delta\|_{r|W}.$$

Therefore $\|\Delta\|_{r|W}$ is a norm which relaxes the corresponding weighted H-infinity norm and is stronger than the associated weighted Hankel norm.

It will be shown later in this section that the unsampled maximal real part optimal model reduction problem can be reduced to a semidefinite program of the size which grows linearly with k and the orders of G_0 and H. This extends significantly the set of model reduction setttings for which a solution can be found efficiently.

The Sampled Setup

Applications of model reduction often deal with the situation in which the order of G_0 (hundreds of thousands) is so large that it becomes not practical to handle the exact state-space or transfer funcion representations of G_0. In such cases it may be useful to work with sampled frequency domain values of G.

The *sampled* version of the maximal real part model reduction problem is defined as the task of finding $\hat{G} \in \mathbf{A}_k$ which minimizes $\|G_0 - \hat{G}\|_{s|V}$, where $V = \{(W_i, t_i)\}_{i=1}^N$ is a given sequence of pairs of real numbers $t_i \in [0, \pi]$, $W_i > 0$,

$$\|\Delta\|_{s|V} = \max_{1 \le i \le N} W_i |\mathrm{Re}(\Delta(e^{jt_i}))|,$$

and G_0 is a stable transfer function which is given (incompletely) by its samples $G_{0,i} = G(e^{jt_i})$.

Note that $\|G_0 - \hat{G}\|_{s|V}$ is completely determined by $\hat{G}$, $V = \{(W_i, t_i)\}_{i=1}^N$, and the samples $G_{0,i} = G(e^{jt_i})$. Practical use of the sampled setup usually relies on an assumption that $G(e^{jt})$ does not vary too much between the sample points t_i.

It is possible to propose various modifications of the sampled modeling error measure $\|\cdot\|_{s|W}$. For example, assuming that $0 = t_0 \le t_1 \le t_2 \le \cdots \le t_N \le t_{N+1}$, one can use the mixed cost

$$J(G_0, \hat{G}) = \max_{1 \le i \le N} W_i \max_{t \in [t_{i-1}, t_{i+1}]} |G(e^{jt_i}) - \hat{G}(e^{jt})|.$$

The main point, however, is the possibility to reduce the model reduction setup to an equivalent semidefinite program, to be shown in the next subsection.

The Convex Parameterization

By a *trigonometric polynomial* f of degree $m = \deg(f)$ we mean a function $f : \mathbf{T} \to \mathbf{R}$ of the form

$$g(e^{jt}) = \sum_{k=0}^{m} g_k \cos(kt),$$

where $g_k \in \mathbf{R}$ and $g_m \neq 0$.

The following simple observation is a key to the convexification of maximal real part optimal model reduction problems.

LEMMA 16.1
For every $f \in \mathbf{A}_m$ there exist trigonometric polynomials a, b such that

$$\deg(a) < m,\ \deg(b) < m,\quad a(z) > 0 \quad \forall\ z \in \mathbf{T}, \tag{16.5}$$

and

$$\mathrm{Re}(f(z)) = \frac{b(z)}{a(z)} \quad \forall\ z \in \mathbf{T}. \tag{16.6}$$

Conversely, for every pair (a, b) of trigonometric polynomials satisfying (16.5) there exists $f \in \mathbf{A}_m$ such that (16.6) holds. □

Proof. If $f \in \mathbf{A}_m$ then

$$f(z) = \frac{p(z)}{q(z)} \quad \forall\ z \in \mathbf{T},$$

where p, q are polynomials of degree less than m, with real coefficients, and $q(z) \neq 0$ for $|z| \geq 1$. Hence (16.6) holds for a, b defined by

$$a(z) = q(z)q(1/z), \quad b(z) = \frac{1}{2}(q(z)p(1/z) + q(1/z)p(z)) \quad (z \neq 0), \tag{16.7}$$

or, equivalently, by

$$a(z) = |q(z)|^2, \quad b(z) = \mathrm{Re}(p(z)q(\bar{z})) \quad (|z| = 1).$$

It is easy to see that a, b defined by (16.7) are trigonometric polinomials satisfying (16.5).

Conversely, let a, b be trigonometric polinomials satisfying (16.5). Let $r = \deg(a)$. Then $h(z) = z^r a(z)$ is an ordinary polynomial of degree $2r$. Since $a(z) > 0$ for all $z \in \mathbf{T}$, $h(z)$ has no zeros in $\mathbf{T} \cup \{0\}$. Since $a(z) = a(1/z)$, all zeros of $h(z)$ can be arranged in pairs $(z_i, 1/z_i)$, where $|z_i| < 1$, $i = 1, \ldots, r$, i.e. $h(z) = q_0 z^r q_r(z) q_r(1/z)$ where q_0 is a constant, and

$$q_r(z) = (z - z_1)(z - z_2) \cdot \cdots \cdot (z - z_r)$$

is a polynomial with no zeros in the region $|z| \geq 1$. Moreover, since h has real coefficients, the non-real zeros of h come in conjugated pairs, and hence q_r has real coefficients as well, and $q_0 \in \mathbf{R}$. Equivalently, we have

$$a(z) = q_0 |q_r(z)|^2 \quad \forall\ z \in \mathbf{T}.$$

Since $a(z) > 0$ for all $z \in \mathbf{T}$, we have $q_0 > 0$. Let

$$q(z) = q_0^{1/2} z^{m-r-1} q_r(z).$$

Then q is a polinomial of degree $m - 1$ with no zeros in the region $|z| \geq 1$, and $|q(z)|^2 = a(z)$ for all $z \in \mathbf{T}$.

It is left to show that a polynomial $p(z)$ of degree less than m with real coefficients can be found such that

$$2b(z) = p(z)q(1/z) + p(1/z)q(z).$$

Indeed, the set V of all real polynomials p of degree less than m forms an m-dimensional real vector space. The map

$$M_q : \quad p(z) \ \mapsto\ p(z)q(1/z) + p(1/z)q(z)$$

is a linear transformation from V into the m-dimensional real vector space of trigonometric polinomials of degree less than m. Moreover, $\ker M_q = \{0\}$, since

$$p(z)q(1/z) = -p(1/z)q(z)$$

would imply that p and q have same set of zeros (here we use the fact that all zeros of q are in the open unit disc $|z| < 1$), hence $p(z) = cq(z)$ and $c = 0$, i.e. $p = 0$. Therefore M_q is a bijection. □

Using Lemma 1 it is easy to convexify the maximal real part optimal model reduction problems. In particular, the unsampled version originally has the form

$$y \to \min \text{ subject to } |H(z)|^2|(Re)(G_0(z) - \hat{G}(z))| \le y \ \forall\ z \in \mathbf{T},\ \hat{G} \in \mathbf{A}_m.$$

Replacing $\mathrm{Re}(\hat{G}(z))$ by $b(z)/a(z)$ where a, b are the trigonometric polynomials from Lemma 1, we obtain an equivalent formulation

$$y \to \min \text{ subject to } |H(z)|^2|a(z)Re(G_0(z)) - b(z)| \le ya(z) \ \forall\ z \in \mathbf{T},$$

where the decision parameters a, b are constrained by (16.5). Since, for a given y, the constraints imposed on a, b are convex, the optimization problem is quasi-convex, and can be solved by combining a binary search over y with a convex feasibility optimization over a, b for a fixed y. For practical implementation, using an interior point cutting plane algorithm with a feasibility oracle utilizing the Kalman-Yakubovich-Popov lemma is advisable here. Another option would be to use the Kalman-Yakubovich-Popov lemma to transform the frequency domain inequalities

$$a(z) > 0,\ \pm|H(z)|^2(a(z)Re(G_0(z)) - b(z)) \le ya(z) \ \forall\ z \in \mathbf{T},$$

(which defines an infnite set of inequalities which are linear with respect to the coefficients of a, b but infinitely parameterized by $z \in \mathbf{T}$) into a set of *three* matrix inequalities, linear with respect to the coefficients of a, b, and three auxiliary symmetric matrix variables P_0, P_+, P_-. If n denotes the sum of m and the orders of H and G_0, the matrix inequalities will have sizes $m+1$,$n+1$ and $n+1$ respectively, and the sizes of $P_0, P_\pm$ will be m-by-m and n-by-n. Therefore, the model reduction problem will be reduced to semidefinite programming.

Similarly, the sampled version of the maximal real part optimal model reduction problem can be reduced to the convex optimization problem

$$y \to \min \text{ subject to } W_i|a(e^{jt_i})\mathrm{Re}(G_{0,i}) - b(e^{jt_i})| \le y \quad (1 \le i \le N),$$

where the decision parameters a, b are constrained by (16.5).

Note that the complexity of this quasi-convex optimization grows slowly with N, which can be very large. The complexity grows faster with m, but this does not appear to be a significant problem, since m, the desired reduced order, is small in most applications.

16.3 H-Infinity Modeling Error Bounds

In this section, H-infinity approximation quality of maximal real part optimal reduced models is examined. For a given stable transfer function $G \in \mathbf{A}$ let

$$d_m^\alpha(G) = \min_{\hat{G} \in \mathbf{A}_m} \|G - \hat{G}\|_\alpha,$$

$$\hat{G}_m^\alpha = \arg\min_{\hat{G} \in \mathbf{A}_m} \|G - \hat{G}\|_\alpha,$$

denote the minimum and the argument of minimum in the corresponding optimal model reduction problems, where $\alpha \in \{\infty, h, r\}$ indicates one of the unweighted norms: H-infinity, Hankel, or maximal real part. In the case when the optimal reduced model is not unique, the model delivered by a particular optimization algorithm can be considered. Let

$$\tilde{G}_m^\alpha = \hat{G}_m^\alpha + \arg\min_{c \in \mathbf{R}} \|G - \hat{G}_m^\alpha - c\|_\infty.$$

In other words, let $\tilde{G}_m^\alpha$ be the result of an adjustment of $\hat{G}_m^\alpha$ by an additive constant factor (which obviously does not change the order) to further optimize the H-infinity model reduction error. Practically, the modified $\tilde{G}_m^\alpha$ is easy to calculate. Obviously, $\hat{G}_m^\infty = \tilde{G}_m^\infty$.

Since

$$\|\Delta\|_\infty \geq \|\Delta\|_h, \quad \|\Delta\|_\infty \geq \|\Delta\|_r$$

for all $\Delta \in \mathbf{A}$, the quantities $d_m^h(G)$ and $d_m^r(G)$ are lower bounds of $d_m^\infty(G)$. One can argue that a maximal real part modified optimal reduced model $\tilde{G}_m^r$ (or, alternatively, a Hankel optimal modified reduced model $\tilde{G}_m^h$) is an acceptable surrogate of $\hat{G}_m^\infty$ when $\|G - \tilde{G}_m^r\|_\infty$ is not much larger than d_m^r (respectively, when $\|G - \tilde{G}_m^h\|_\infty$ is not much larger than d_m^h). For the Hankel optimal reduced models, a theoretical evidence that this will frequently be the case is provided by the inequality

$$\|G - \tilde{G}_m^h\|_\infty \leq \sum_{k \geq m} d_m^h(G). \tag{16.8}$$

Since for a "nice" smooth transfer function G the numbers $d_m^h(G)$ converge to zero quickly, (16.8) gives some assurance of good asymptotic behavior of H-infinity modeling errors for Hankel optimal reduced models.

Since $2\|\Delta\|_r \geq \|\Delta\|_h$ for all $\Delta \in \mathbf{A}$, one can argue that the maximal real part norm is "closer" to the H-infinity norm than the Hankel norm. However, the author was not able to prove a formal statement confirming the conjectured asymptotic superiority of maximal real part reduced models over the Hankel reduced models. Instead, a theorem demonstrating asymptotic behavior roughly comparable to that of the Hankel reduced models is given below.

THEOREM 16.1
For all $G \in \mathbf{A}$ and $m > 0$

$$\|G - \tilde{G}_m^r\|_\infty \leq 12 \sum_{k=0}^{\infty} 2^k m d_{2^k m}^r(G). \tag{16.9}$$

□

Proof. From (16.8), for every $f \in \mathbf{A}_n$

$$\|f - \tilde{f}_0^h\|_\infty \le \sum_{k=0}^{\infty} d_k^h(f) \le n d_0^h(f) = n\|f\|_h \le 2n\|f\|_r.$$

Note that $\tilde{f}_0^h$ is a constant transfer function. We have

$$\|\hat{G}_n^r - \hat{G}_{2n}^r\|_r \le \|\hat{G}_n^r - G\| + \|\hat{G}_{2n}^r - G\|_r \le 2d_n^r(G).$$

Hence

$$\|\hat{G}_n^r - \hat{G}_{2n}^r - c_n\|_\infty \le 12 n d_n^r(G)$$

for an appropriately chosen real constant c_n. Hence

$$\|\hat{G}_m^r - G - c\|_\infty \le 12 \sum_{k=0}^{\infty} 2^k m d_{2^k m}^r(G)$$

for an appropriately chosen constant c, which in turn implies (16.9). □

It may appear that (16.8) is a much better bound than (16.9), since $d_m(G)$ is not being multiplied by m in (16.8). However, the calculations for $d_m \approx 1/m^q$ as $m \to \infty$, where $q > 1$, result in the same rate of asymtotic convergence for the two upper bounds:

$$\sum_{k=m}^{\infty} \frac{1}{k^q} \approx \frac{c_1}{m^{q-1}},$$

$$\sum_{k=m}^{\infty} 2^k m \frac{1}{(2^k m)^q} \approx \frac{c_2}{m^{q-1}}.$$

16.4 Minor Improvements and Numerical Experiments

There is a number of ways in which the H-infinity quality of the maximal real part optimal reduced models can be improved. One simple trick is to re-optimize the numerator p of $\hat{G}_m^r = p/q$ with the optimal q being fixed (this is partially used in Hankel model reduction when $\hat{G}_m^h$ is being replaced by $\tilde{G}_m^h$). A further improvement of the *lower bound* $d_m^r(G)$ can be achieved when, in a weighted H-infinity model reduction setup $\|W(G - \hat{G}_m)\|_\infty \to \min$ $\hat{G}_m$ will be replaced by

$$\hat{G}_m^e = \frac{b(z) + (z - \bar{z})c(z)}{a(z)},$$

where a, b are constrained by (16.5), and c is an arbitrary trigonometric polynomial of degree less than $m-1$. Then optimization with respect to a, b, c is convex, and the optimal a can be used to get the *denominator* q of the reduced model,

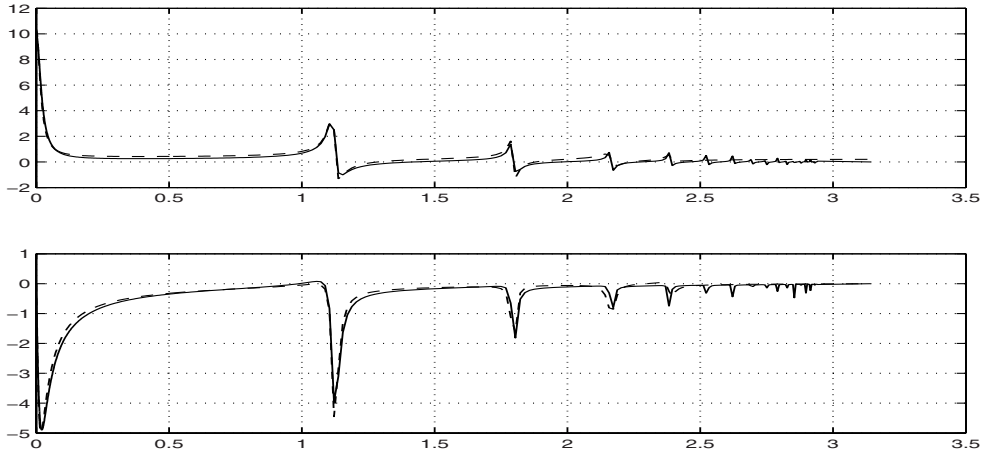


Figure 16.1 functions with delay

after which the numerator p is to be found via a separate optimization round. Note that, since

$$|G(z) - \hat{G}_m^e(z)| \geq |\mathrm{Re}(G(z)) - b(z)/a(z)|,$$

where the equality is achieved for $c \equiv 0$, the original maximal real part model reduction is a special case of the general scheme.

With these improvements implemented, the maximal real part model reduction performs reasonably well, as demonstrated by the following examples. The software used to produce the data (requires MATLAB and CPLEX) can be obtained by sending a request to ameg@mit.edu.

Functions With Delay

Here G is the infinite dimensional transfer function

$$G = \frac{1}{(1 - .9e^{-s})(1 + .3s)}.$$

For a 10th order approximation, a lower bound $d_{10}^r(G) \geq 0.35$ was found. The resulting reduced 10th order model $\tilde{G}_{10}^r$ satisfies

$$\|G - \hat{G}_{10}^r\|_\infty < 1.4.$$

Focus Servo of a DVD Player

59 frequency samples obtained as an experimental data were provided. For a 10th order fit, the lower bound of about 1.6 was obtained. The actual H-infinity error on the sampled data was approximately 6.2.

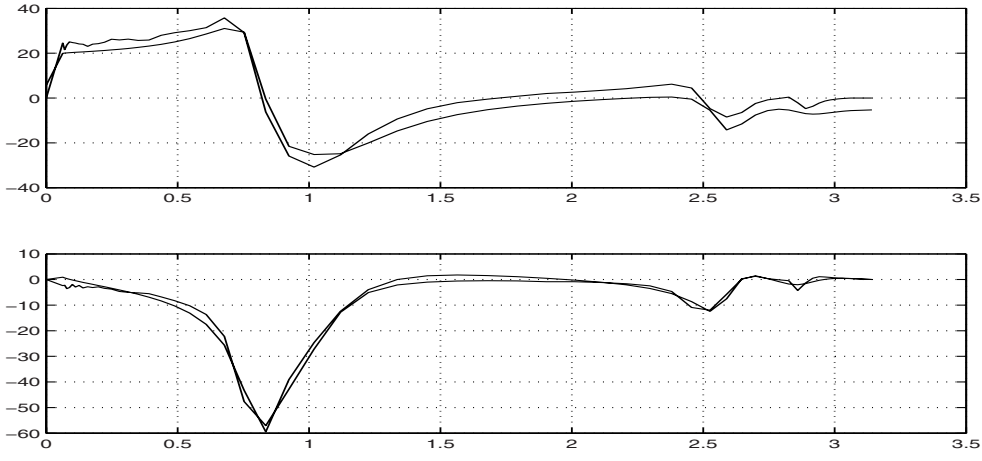


Figure 16.2 DVD focus servo

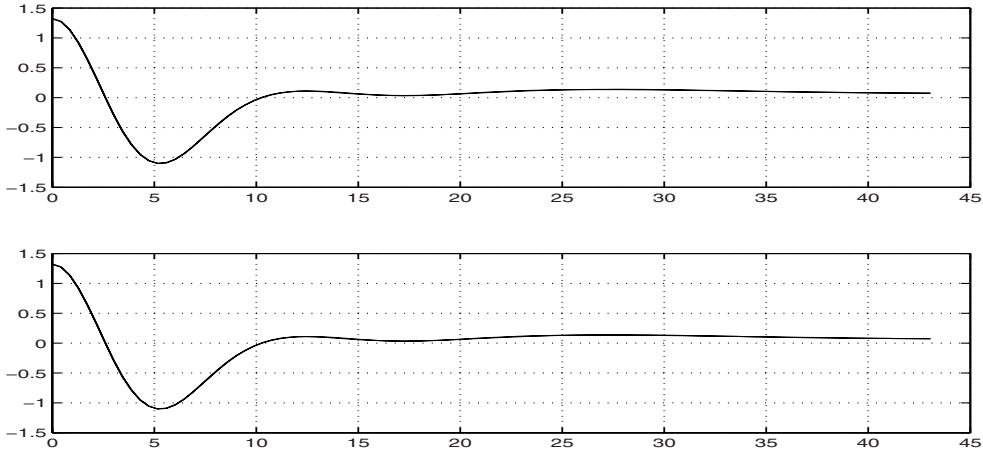


Figure 16.3 fluid dynamics control

Fluid Dynamics Control

A jet engine outlet pressure is to be controlled by regulating a discharge valve in midstream. The complete dynamical model of the actuator dynamics is given by partial differential equations. Computational fluid dynamics simulations provided a 101 sample of the transfer function. A 5th order reduced model was sought for the system. The lower bound of achievable H-infinity quality is 0.0069. A reduced model of quality 0.009 was found.

Acknowledgment

This work was supported by NSF, AFOSR, and DARPA

17

Quantum Schrödinger Bridges

Michele Pavon

Abstract

Elaborating on M. Pavon, *J. Math. Phys.* **40** (1999), 5565-5577, we develop a simplified version of a variational principle within Nelson stochastic mechanics that produces the von Neumann wave packet reduction after a position measurement. This stochastic control problem parallels, with a different kinematics, the problem of the Schrödinger bridge. This gives a profound meaning to what was observed by Schrödinger in 1931 concerning Schrödinger bridges: "*Merkwürdige Analogien zur Quantenmechanik, die mir sehr des Hindenkens wert erscheinen*".

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 227-238, 2003.
© Springer-Verlag Berlin Heidelberg 2003

17.1 Introduction: Schrödinger's Problem

In 1931/32 [1, 2], Schrödinger considered the following problem. A cloud of N Brownian particles in $\mathbb{R}^n$ has been observed having at time t_0 an empirical distribution approximately equal to $\rho_0(x)dx$. At some later time t_1, an empirical distribution approximately equal to $\rho_1(x)dx$ is observed. Suppose that $\rho_1(x)$ considerably differs from what it should be according to the law of large numbers (N is large), namely

$$\int_{t_0}^{t_1} p(t_0, y, t_1, x)\rho_0(y)dy,$$

where

$$p(s, y, t, x) = [2\pi(t-s)]^{-\frac{n}{2}} \exp\left[-\frac{|x-y|^2}{2(t-s)}\right], \quad s < t,$$

is the transition density of the Wiener process. It is apparent that the particles have been transported in an unlikely way. But of the many unlikely ways in which this could have happened, which one is the most likely?

In modern probabilistic language, this is a problem of large deviations of the empirical distribution [3]. By discretization and passage to the limit, Schrödinger computed the most likely intermediate empirical distribution as $N \to \infty$. It turned out that the optimal random evolution, the *Schrödinger bridge* from ρ_0 to ρ_1 over Brownian motion, had at each time a density $\rho(\cdot, t)$ that factored as $\rho(x,t) = \phi(x,t)\hat{\phi}(x,t)$, where ϕ and $\hat{\phi}$ are a p-harmonic and a p-coharmonic functions, respectively. That is

$$\phi(t,x) = \int p(t,x,t_1,y)\phi(t_1,y)dy, \tag{17.1}$$

$$\hat{\phi}(t,x) = \int p(t_0,y,t,x)\hat{\phi}(t_0,y)dy. \tag{17.2}$$

The existence and uniqueness of a pair $(\phi, \hat{\phi})$ satisfying (17.1)-(17.2) and the boundary conditions $\phi(x,t_0)\hat{\phi}(x,t_0) = \rho_0(x)$, $\phi(x,t_1)\hat{\phi}(x,t_1) = \rho_1(x)$ was guessed by Schrödinger on the basis of his intuition. He was later shown to be quite right in various degrees of generality by Fortet [4], Beurlin [5], Jamison [6], Föllmer [3]. Jamison showed, in particular, that the Schrödinger bridge is the unique *Markov* process $\{x(t)\}$ in the class of *reciprocal processes* (one-dimensional Markov fields) introduced by Bernstein [7] having as two-sided transition density

$$q(s,x;t,y;u,z) = \frac{p(s,x;t,y)p(t,y;u,z)}{p(s,x;u,z)}, \quad s < t < u,$$

namely $q(s,x;t,y;u,z)dy$ is the probability of finding the process x in the volume dy at time t given that $x(s) = x$ and $x(u) = z$. Schrödinger was struck by the following remarkable property of the solution: The Schrödinger bridge from ρ_1 to ρ_0 over Brownian motion is just the time reversal of the Schrödinger bridge from ρ_0 to ρ_1. In Schrödinger's words: "Abnormal states have arisen with high probability by an exact time reversal of a proper diffusion process". This led him to entitle [1]: "On the reversal of natural laws" A few years later, Kolmogorov wrote a paper on the subject with a very similar title [8]. Moreover, the fact that the

Schrödinger bridge has density $\rho(x,t) = \phi(x,t)\hat{\phi}(x,t)$ resembles the fact that in quantum mechanics the density may be expressed as $\rho(x,t) = \psi(x,t)\bar{\psi}(x,t)$. This analogy has inspired various attempts to construct a stochastic reformulation of quantum mechanics [9]-[12] starting from [1, 2, 7]. In order to discuss a more general Schrödinger bridge problem, we recall in the next session some essential facts on the kinematics of finite-energy diffusions as presented in [13, 14, 15, 16].

17.2 Elements of Nelson-Föllmer Kinematics of Finite Energy Diffusions

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a complete probability space. A stochastic process $\{\xi(t); t_0 \leq t \leq t_1\}$ mapping $[t_0, t_1]$ into $L_n^2(\Omega, \mathcal{F}, \mathbf{P})$ is called a *finite-energy diffusion* with constant diffusion coefficient $I_n\sigma^2$ if the path $\xi(\omega)$ belongs a.s. to $C([t_0, t_1]; \mathbb{R}^n)$ (n-dimensional continuous functions) and

$$\xi(t) - \xi(s) = \int_s^t \beta(\tau)d\tau + \sigma w_+(s,t), \quad t_0 \leq s < t \leq t_1, \tag{17.3}$$

where the *forward drift* $\beta(t)$ is at each time t a measurable function of the past $\{\xi(\tau); 0 \leq \tau \leq t\}$, and $w_+(\cdot,\cdot)$ is a standard, n-dimensional *Wiener difference process* with the property that $w_+(s,t)$ is independent of $\{\xi(\tau); 0 \leq \tau \leq s\}$. Moreover, β must satisfy the finite-energy condition

$$E\left\{\int_{t_0}^{t_1} \beta(\tau)\cdot\beta(\tau)d\tau\right\} < \infty. \tag{17.4}$$

We recall the characterizing properties of the n-dimensional *Wiener difference process* $w_+(s,t)$, see [13, Chapter 11] and [15, Section 1]. It is a process such that $w_+(t,s) = -w_+(s,t)$, $w_+(s,u) + w_+(u,t) = w_+(s,t)$, and that $w_+(s,t)$ is Gaussian distributed with mean zero and variance $I_n|s-t|$. Moreover, (the components of) $w_+(s,t)$ and $w_+(u,v)$ are independent whenever $[s,t]$ and $[u,v]$ don't overlap. Of course, $w_+(t) := w_+(t_0,t)$ is a standard Wiener process such that $w_+(s,t) = w_+(t) - w_+(s)$. In [14], Föllmer has shown that a finite-energy diffusion also admits a reverse-time differential. Namely, there exists a measurable function $\gamma(t)$ of the future $\{\xi(\tau); t \leq \tau \leq t_1\}$ called *backward drift*, and another Wiener difference process w_- such that

$$\xi(t) - \xi(s) = \int_s^t \gamma(\tau)d\tau + \sigma w_-(s,t), \quad t_0 \leq s < t \leq t_1. \tag{17.5}$$

Moreover, γ satisfies

$$E\left\{\int_{t_0}^{t_1} \gamma(\tau)\cdot\gamma(\tau)d\tau\right\} < \infty, \tag{17.6}$$

and $w_-(s,t)$ is independent of $\{\xi(\tau); t \leq \tau \leq t_1\}$. Let us agree that dt always indicates a strictly positive variable. For any function f defined on $[t_0, t_1]$, let

$$d_+f(t) = f(t+dt) - f(t)$$

be the *forward increment* at time t, and

$$d_- f(t) = f(t) - f(t - dt)$$

be the *backward increment* at time t. For a finite-energy diffusion, Föllmer has also shown in [14] that the forward and backward drifts may be obtained as Nelson's conditional derivatives, namely

$$\beta(t) = \lim_{dt \searrow 0} E\left\{ \frac{d_+ \xi(t)}{dt} | \xi(\tau), t_0 \le \tau \le t \right\},$$

and

$$\gamma(t) = \lim_{dt \searrow 0} E\left\{ \frac{d_- \xi(t)}{dt} | \xi(\tau), t \le \tau \le t_1 \right\},$$

the limits being taken in $L_n^2(\Omega, \mathcal{F}, P)$. It was finally shown in [14] that the one-time probability density $\rho(\cdot, t)$ of $\xi(t)$ (which exists for every $t > t_0$) is absolutely continuous on $\mathbb{R}^n$ and the following duality relation holds $\forall t > 0$

$$E\{\beta(t) - \gamma(t) | \xi(t)\} = \sigma^2 \nabla \log \rho(\xi(t), t), \quad \text{a.s.} \tag{17.7}$$

REMARK 17.1
It should be observed that in the study of reverse-time differentials of diffusion processes, initiated by Nelson in [17] and Nagasawa in [18], see [19, 20] and references therein, important results have obtained by A. Linquist and G. Picci in the Gaussian case [21, 22] without assumptions on the reverse-time differential. *In particular, their results on Gauss-Markov processes have been crucial in order to develop a strong form of* stochastic realization theory *[21]-[24] together with a variety of applications [24]-[30].* □

Corresponding to (17.3) and (17.5) are two change of variables formulas. Let $f : \mathbb{R}^n \times [0, T] \to \mathbb{R}$ be twice continuously differentiable with respect to the spatial variable and once with respect to time. Then, if ξ is a finite-energy diffusion satisfying (17.3) and (17.5), we have

$$f(\xi(t), t) - f(\xi(s), s) = \int_s^t \left(\frac{\partial}{\partial \tau} + \beta(\tau) \cdot \nabla + \frac{\sigma^2}{2} \Delta \right) f(\xi(\tau), \tau) d\tau + \int_s^t \sigma \nabla f(\xi(\tau), \tau) \cdot d_+ w_+(\tau), \tag{17.8}$$

$$f(\xi(t), t) - f(\xi(s), s) = \int_s^t \left(\frac{\partial}{\partial \tau} + \gamma(\tau) \cdot \nabla - \frac{\sigma^2}{2} \Delta \right) f(\xi(\tau), \tau) d\tau + \int_s^t \sigma \nabla f(\xi(\tau), \tau) \cdot d_- w_-(\tau). \tag{17.9}$$

The stochastic integrals appearing in (17.8) and (17.9) are a (forward) Ito integral and a backward Ito integral, respectively, see [15] for the details.

17.3 Schrödinger Bridges

The solution to the Schrödinger problem can be obtained by solving a stochastic control problem. The Kullback-Leibler pseudo-distance between two probability densities $p(\cdot)$ and $q(\cdot)$ is defined by

$$H(p,q) := \int_{\mathbb{R}^n} \log \frac{p(x)}{q(x)} p(x) dx.$$

This concept can be considerably generalized. Let $\Omega := C([t_0,t_1],\mathbb{R}^n)$ denote the family of n-dimensional continuous functions, let W_x denote Wiener measure on Ω starting at x, and let

$$W := \int W_x \, dx$$

be stationary Wiener measure. Let $\mathbb{D}$ be the family of distributions on Ω that are equivalent to W. For $Q, P \in \mathbb{D}$, we define the *relative entropy* $H(Q,P)$ of Q with respect to P as

$$H(Q,P) = E_Q[\log \frac{dQ}{dP}].$$

It then follows from Girsanov's theorem that [16, 14, 3]

$$\begin{aligned} H(Q,P) &= H(q(t_0),p(t_0)) + E_Q\left[\int_{t_0}^{t_1} \frac{1}{2}[\beta^Q(t)-\beta^P(t)]\cdot[\beta^Q(t)-\beta^P(t)]dt\right] \\ &= H(q(t_1),p(t_1)) + E_Q\left[\int_{t_0}^{t_1} \frac{1}{2}[\gamma^Q(t)-\gamma^P(t)]\cdot[\gamma^Q(t)-\gamma^P(t)]dt\right] \quad (17.10) \end{aligned}$$

Here $q(t_0)$ is the marginal density of Q at t_0, β^Q and γ^Q are the forward and the backward drifts of Q, respectively. Now let ρ_0 and ρ_1 be two everywhere positive probability densities. Let $\mathbb{D}(\rho_0,\rho_1)$ denote the set of distributions in $\mathbb{D}$ having the prescribed marginal densities at t_0 and t_1. Given $P \in \mathbb{D}$, we consider the following problem:

$$\text{Minimize} \quad H(Q,P) \quad \text{over} \quad \mathbb{D}(\rho_0,\rho_1).$$

In view of (17.10), this is a stochastic control problem. It is connected through Sanov's theorem [3, 32] to a problem of large deviations of the empirical distribution, according to Schrödinger original motivation. Namely, if $X^1, X^2, \ldots$ is an i.i.d. sequence of random elements on Ω with distribution P, then the sequence $P^n[\frac{1}{n}\sum_{i=1}^n \delta_{X^i} \in \cdot]$ satisfies a large deviation principle with *rate function* $H(\cdot,P)$. If there is at least one Q in $\mathbb{D}(\rho_0,\rho_1)$ such that $H(Q,P) < \infty$, it may be shown that there exists a unique minimizer Q^* in $\mathbb{D}(\rho_0,\rho_1)$ called *the Schrödinger bridge* from ρ_0 to ρ_1 over P. If (the coordinate process under) P is Markovian with forward drift field $b_+^P(x,t)$ and transition density $p(\sigma,x,\tau,y)$, then Q^* is also Markovian with forward drift field

$$b_+^{Q^*}(x,t) = b_+^P(x,t) + \nabla \log \phi(x,t),$$

where the (everywhere positive) function ϕ solves together with another function $\hat{\phi}$ the system (17.1)-(17.2) with boundary conditions

$$\phi(x,t_0)\hat{\phi}(x,t_0) = \rho_0(x), \quad \phi(x,t_1)\hat{\phi}(x,t_1) = \rho_1(x).$$

Moreover, $\rho(x,t) = \phi(x,t)\hat{\phi}(x,t), \forall t \in [t_0, t_1]$. This result has been suitably extended to the case where P is non-Markovian in [31]. For a survey of the theory of Schrödinger bridges with an extended bibliography see [32].

Consider now the following simpler problem: We have a *reference stochastic model* $P \in \mathbb{D}$. We think of P as modeling the macroscopic evolution of a thermodynamic system. Suppose we observe at time t_1 the (everywhere positive) density ρ_1 different from the marginal density of P. Thus we need to solve the following optimization problem

$$\text{Minimize} \quad H(Q,P) \quad \text{over} \quad Q \in \mathbb{D}(\rho_1).$$

where $\mathbb{D}(\rho_1)$ denotes the set of distributions in $\mathbb{D}$ having density ρ_1 at t_1. Let us assume that $H(\rho_1, p(t_1)) < \infty$. In view of (17.10), this stochastic control problem can be trivially solved. The unique solution is given by the distribution Q^* having backward drift $\gamma^P(t)$ and marginal density ρ_1 at time t_1. Thus, the result of measurement at time t_1 leads to the replacement of the stochastic model P with Q^*. Notice that the backward drift $\gamma^P(t)$ is perfectly preserved by this procedure. Symmetrically, if we were to change the initial distribution at time t_0, the procedure would preserve the forward drift $\beta^P(t)$.

17.4 Elements of Nelson's Stochastic Mechanics

Nelson's stochastic mechanics is a quantization procedure for classical dynamical systems based on diffusion processes. Following some early work by Feynes [33] and others, Nelson and Guerra elaborated a clean formulation starting from 1966 [34, 13, 35], showing that the Schrödinger equation could be derived from a continuity type equation plus a Newton type law, provided one accepted a certain definition for the stochastic acceleration. In analogy to classical mechanics, the Newton-Nelson law was later shown to follow from a Hamilton-like stochastic variational principle [36, 37]. Other versions of the variational principle have been proposed in [38, 39, 40, 41].
Consider the case of a nonrelativistic particle of mass m. Let $\{\psi(x,t); t_0 \le t \le t_1\}$ be the solution of the *Schrödinger equation*

$$\frac{\partial \psi}{\partial t} = \frac{i\hbar}{2m}\Delta\psi - \frac{i}{\hbar}V(x)\psi, \tag{17.11}$$

such that

$$||\nabla\psi||_2^2 \in L^1_{\text{loc}}[t_0, +\infty). \tag{17.12}$$

This is Carlen's *finite action condition*. Under these hypotheses, the Nelson measure $P \in \mathbb{D}$ may be constructed on path space, [42],[43], [39, Chapter IV], and references therein. Namely, letting $\Omega := C([t_0, t_1], \mathbb{R}^n)$ the n-dimensional continuous functions on $[t_0, t_1]$, under the probability measure P, the canonical coordinate process $x(t,\omega) = \omega(t)$ is an n-dimensional Markov diffusion process $\{x(t); t_0 \le t \le t_1\}$, called *Nelson's process*, having (forward) Ito differential

$$dx(t) = \left[\frac{\hbar}{m}\nabla\left(\Re \log\psi(x(t),t) + \Im \log\psi(x(t),t)\right)\right]dt + \sqrt{\frac{\hbar}{m}}dw(t), \tag{17.13}$$

where w is a standard, n-dimensional Wiener process. Moreover, the probability density $\rho(\cdot, t)$ of $x(t)$ satisfies

$$\rho(x,t) = |\psi(x,t)|^2, \quad \forall t \in [t_0, t_1]. \tag{17.14}$$

Following Nelson [13, 38], for a finite-energy diffusion with stochastic differentials (17.3)-(17.5), we define the *current* and *osmotic* drifts, respectively:

$$v(t) = \frac{\beta(t) + \gamma(t)}{2}, \quad u(t) = \frac{\beta(t) - \gamma(t)}{2}.$$

Clearly v is similar to the classical velocity, whereas u is the velocity due to the "noise" which tends to zero when the diffusion coefficient σ^2 tends to zero. In order to obtain a unique time-reversal invariant differential [40], we take a complex linear combination of (17.3)-(17.5), obtaining

$$\begin{aligned} x(t) - x(s) &= \int_s^t \left[\frac{1-i}{2}\beta(\tau) + \frac{1+i}{2}\gamma(\tau)\right] d\tau \\ &+ \frac{\sigma}{2}\left[(1-i)(w_+(t) - w_+(s)) + (1+i)(w_-(t) - w_-(s))\right]. \end{aligned}$$

Let us define the *quantum drift*

$$v_q(t) := \frac{1-i}{2}\beta(t) + \frac{1+i}{2}\gamma(t) = v(t) - iu(t),$$

and the *quantum noise*

$$w_q(t) := \frac{1-i}{2}w_+(t) + \frac{1+i}{2}w_-(t).$$

Hence,

$$x(t) - x(s) = \int_s^t v_q(\tau)d\tau + \sigma[w_q(t) - w_q(s)]. \tag{17.15}$$

This representation enjoys the time reversal invariance property. It has been crucial in order to develop a Lagrangian and a Hamiltonian dynamics formalism in the context of Nelson's stochastic mechanics in [40, 44, 45]. Notice that replacing (17.3)-(17.5) with (17.15), we replace the pair of real drifts (v, u) by the unique *complex-valued* drift $v - iu$ that tends correctly to v when the diffusion coefficient tends to zero.

17.5 Quantum Schrödinger Bridges

We now consider the same problem as at the end of Section 17.3. We have a *reference stochastic model* $P \in \mathbb{D}$ given by the Nelson measure on path space that has been constructed through a variational principle [37, 38, 40]. This Nelson process $x = \{x(t); t_0 \le t \le t_1\}$ has an associated solution $\{\psi(x,t) : t_0 \le t \le t_1\}$ of the Schrödinger equation in the sense that the quantum drift of x is $v_q(t) = \frac{\hbar}{im}\nabla \log \psi(x(t), t)$ and the one-time density of x satisfies $\rho(x,t) = |\psi(x,t|^2$. Suppose

a position measurement at time t_1 yields the probability density $\rho_1(x) \neq |\psi(x,t_1)|^2$. We need a suitable variational mechanism that, starting from (P,ρ_1), produces the new stochastic model in $\mathbb{D}(\rho_1)$. It is apparent that the variational problem of Section 17.3 is not suitable as it preserves the backward drift. Since in stochastic mechanics both differentials must be granted the same status, we need to change both drifts as little as possible given the new density ρ_1 at time t_1. Thus, we employ the differential (17.15), and consider the variational problem:
Extremize on $(\tilde{x},\tilde{v}_q) \in (\mathbb{D}(\rho_1) \times \mathcal{V})$ the functional

$$J(\tilde{x},\tilde{v_q}) := E\left\{\frac{1}{2}\log\frac{\tilde{\rho}_1(\tilde{x}(t_1))}{\rho(\tilde{x}(t_1),t_1)} + \int_{t_1}^{t_2}\frac{mi}{2\hbar}(v_q(\tilde{x}(t),t)-\tilde{v}_q(t))\cdot(v_q(\tilde{x}(t),t)-\tilde{v}_q(t))\,dt\right\}$$

subject to: $\tilde{x}$ has quantum drift (velocity) $\tilde{v}_q$.
Here $v_q(x,t) = \frac{\hbar}{im}\nabla\log\psi(x,t)$ is quantum drift of Nelson reference process, and $\mathbb{D}(\rho_1)$ is family of finite-energy, $\mathbb{R}^n$-valued diffusions on $[t_0,t_1]$ with diffusion coefficient $\frac{\hbar}{m}$, and having marginal ρ_1 at time t_1. Moreover, $\mathcal{V}$ denotes the family of finite-energy, C^n - valued stochastic processes on $[t_0,t_1]$. Following the same variational analysis as in [45], we get a Hamilton-Jacobi-Bellman type equation

$$\frac{\partial\varphi}{\partial t} + v_q(x,t)\cdot\nabla\varphi(x,t) - \frac{i\hbar}{2m}\Delta\varphi(x,t) = \frac{i\hbar}{2m}\nabla\varphi(x,t)\cdot\nabla\varphi(x,t), \tag{17.16}$$

with terminal condition $\varphi(x,t_1) = \frac{1}{2}\log\frac{\rho_1(x)}{\rho(x,t_1)}$. Then $\tilde{x}\in\mathbb{D}(\rho_1)$ with quantum drift

$$v_q(\tilde{x}(t),t) + \frac{\hbar}{mi}\nabla\varphi(\tilde{x}(t),t)$$

solves the extremization problem. Write $\psi(x,t_1) = \rho(x,t_1)^{\frac{1}{2}}\exp[\frac{i}{\hbar}S(x,t_1)]$, and let $\{\tilde{\psi}(x,t)\}$ be solution of Schrödinger equation (17.11) on $[t_0,t_1]$ with terminal condition

$$\tilde{\psi}(x,t_1) = \rho_1(x)^{\frac{1}{2}}\exp[\frac{i}{\hbar}S(x,t_1)].$$

Next, notice that for $t\in[t_0,t_1]$

$$\left[\frac{\partial}{\partial t} + v_q(x,t)\cdot\nabla - \frac{i\hbar}{2m}\Delta\right]\left(\frac{\tilde{\psi}}{\psi}\right) = 0,\ \frac{\tilde{\psi}}{\psi}(x,t_1) = \left(\frac{\rho_1(x)}{\rho(x,t_1)}\right)^{\frac{1}{2}},$$

where $v_q(x,t) = \frac{\hbar}{im}\nabla\log\psi(x(t))$. It follows that $\varphi(x,t) := \log\frac{\tilde{\psi}}{\psi}(x,t)$ solves (17.16), and the corresponding quantum drift is

$$v_q(\tilde{x}(t),t) + \frac{\hbar}{mi}\nabla\varphi(\tilde{x}(t),t) = \frac{\hbar}{mi}\nabla\log\tilde{\psi}(\tilde{x}(t),t).$$

Thus, new process after measurement at time t_1 (*quantum Schrödinger bridge*) is just the Nelson process associated to another solution $\tilde{\psi}$ of the same Schrödinger equation. Invariance of phase at t_1 follows from the variational principle.

17.6 Collapse of the Wavefunction

Consider the case where measurement at time t_1 only gives the information that x lies in subset D of configuration space $\mathbb{R}^n$ of the system. The density $\rho_1(x)$ just after measurement is

$$\rho_1(x) = \frac{\chi_D(x)\rho(x,t_1)}{\int_D \rho(x',t_1)dx'},$$

where $\rho(x,t_1)$ is density of Nelson reference process x at time t_1.. Let A be subspace of $L^2(\mathbb{R}^n)$ of functions with support in D. Then $A^\perp$ is subspace of $L^2(\mathbb{R}^n)$ functions with support in D^c. Decompose $\psi(x,t_1)$ as

$$\psi(x,t_1) = \chi_D(x)\psi(x,t_1) + \chi_{D^c}(x)\psi(x,t_1) = \psi_1(x) + \psi_2(x),$$

with $\psi_1 \in A$ and $\psi_2 \in A^\perp$. The probability p_1 of finding particle in D is

$$p_1 = \int_D |\psi(x,t_1)|^2 dx = \int_{\mathbb{R}^n} |\psi_1(x)|^2 dx.$$

If the result of the measurement at time t_1 is that the particle lies in D, the variational principle replaces $\{x(t)\}$ with $\{\tilde{x}(t)\}$ and, consequently, replaces $\psi(x,t_1) = \psi_1(x) + \psi_2(x)$ with $\tilde{\psi}(x,t_1)$ where

$$\tilde{\psi}(x,t_1) = \rho(x)_1^{\frac{1}{2}} \exp[\frac{i}{\hbar}S(x,t_1)] = \frac{\psi_1(x)}{||\psi_1||_2}.$$

Postulating the variational principle of the previous section (rather than the invariance of the phase at t_1), we have therefore obtained the so-called “collapse of the wavefunction", see e.g. [46] and references therein. The collapse is instantaneous, precisely as in the orthodox theory. It occurs “when the result of the measurement enters the consciousness of the observer" [47]. We mention here that, outside of stochastic mechanics, there exist alternative stochastic descriptions of (non instantaneous) quantum state reduction such as those starting from a stochastic Schrödinger equation, see e.g. [48] and references therein.

17.7 Conclusion and Outlook

We shall show elsewhere [49] that the variational principle of section 17.5 may be replaced by two stochastic differential games with real velocities with an appealing classical interpretation. We shall also show that, using Nelson’s observation in [38, 15] and this variational principle, it is possible to obtain a completely satisfactory classical probabilistic description of the two-slit experiment.

If the variational mechanism described here can be extended to the case where both the initial and final quantum states are varied, it would provide a general approach to the steering problem for quantum systems (extending [50]) that has important applications in quantum computation [51], control of molecular dynamics [52] and many other fields.

17.8 References

[1] E. Schrödinger, Über die Umkehrung der Naturgesetze, *Sitzungsberichte der Preuss Akad. Wissen. Berlin, Phys. Math. Klasse* (1931), 144-153.

[2] E. Schrödinger, Sur la théorie relativiste de l'électron et l'interpretation de la mécanique quantique, *Ann. Inst. H. Poincaré* **2**, 269 (1932).

[3] H. Föllmer, Random fields and diffusion processes, in: *Ècole d'Ètè de Probabilitès de Saint-Flour XV-XVII*, edited by P. L. Hennequin, Lecture Notes in Mathematics, Springer-Verlag, New York, 1988, vol.1362,102-203.

[4] R. Fortet, Résolution d'un système d'equations de M. Schrödinger, *J. Math. Pure Appl.* IX (1940), 83-105.

[5] A. Beurling, An automorphism of product measures, *Ann. Math.* **72** (1960), 189-200.

[6] B. Jamison, Reciprocal processes, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **30** (1974), 65-86.

[7] S. Bernstein, Sur les liaisons entre les grandeurs aléatoires, *Verh. Int. Math. Kongress*, Zürich, Vol. I (1932), 288-309.

[8] A. Kolmogorov, Zur Umkehrbarkeit der statistischen Naturgesetze, *Math. Ann.* **113** (1936), 766-772.

[9] J.C.Zambrini, Stochastic mechanics according to E. Schrödinger, *Phys. Rev.* A **33** (1986) 1532.

[10] M. Nagasawa, Transformations of diffusions and Schrödinger processes, *Prob. Th. Rel. Fields* **82** (1989), 109-136.

[11] B.C.Levy and A.J.Krener, Kinematics and dynamics of reciprocal diffusions, *J.Math.Phys.* **34** (1993) 1846.

[12] B.C.Levy and A.J.Krener,Stochastic mechanics of reciprocal diffusions, *J.Math.Phys.* **37** (1996), 769.

[13] E. Nelson. *Dynamical Theories of Brownian Motion*. Princeton University Press, Princeton, 1967.

[14] H. Föllmer, in *Stochastic Processes - Mathematics and Physics* , Lecture Notes in Mathematics (Springer-Verlag, New York,1986), Vol. 1158, pp. 119-129.

[15] E. Nelson, in*Ècole d'Ètè de Probabilitès de Saint-Flour XV-XVII*, edited by P. L. Hennequin, Lecture Notes in Mathematics (Springer-Verlag, New York, 1988), Vol.1362, pp. 428-450.

[16] I. Karatzas and S. E. Shreve, Brownian Motion and Stochastic Calculus, Springer-Verlag, New York, 1988.

[17] E. Nelson, The adjoint Markov process, Duke Math. J. **25** (1958), 671.

[18] M.Nagasawa, The adjoint process of a diffusion with reflecting barrier, Kodai Math.Sem.Rep. 13, 1961, 235.

[19] L. M. Morato, On the dynamics of diffusions and the related general electromagnetic potentials, J. Math. Phys. **23** (1982), 1020.

[20] U.G.Haussmann and E.Pardoux, Time reversal of diffusions, *The Annals of Probability* **14** (1986), 1188.

[21] A. Lindquist and G.Picci, On the stochastic realization problem, *SIAM J. Control and Optimization* **17** (1979), 365-389.

[22] A. Lindquist and G. Picci, Forward and backward semimartingale models for Gaussian processes with stationary increments, *Stochastics* **15** (1985), 1-50.

[23] A. Lindquist and G. Picci, Realization theory for multivariate stationary Gaussian processes, *SIAM J. Control and Optimization* **23** (1985), 809-857.

[24] A. Lindquist and G. Picci, A geometric approach to modeling and estimation of linear stochastic systems, *J. Mathematical Systems, Estimation, and Control* **1** (1991), 241-333.

[25] F. Badawi, A. Lindquist and M. Pavon, A stochastic realization approach to the smoothing problem, *IEEE Trans. Autom. Control*, **AC-24** (1979), 878-888.

[26] M.Pavon, New results on the interpolation problem for continuous time stationary increments processes, *SIAM J. Control and Optimiz.* **22** (1984), 133-142.

[27] M.Pavon, Canonical correlations of past inputs and future outputs for linear stochastic systems, *Systems and Control Letters* **4** (1984), 209-215.

[28] A. Lindquist and G. Picci, Geometric methods for state space identification, in *Identification, Adaptation, Learning: The Science of Learning Models from Data*,S. Bittanti and G. Picci (editors), Nato ASI Series (Series F, Vol 153), Springer, 1996, 1–69.

[29] A. Lindquist and G. Picci, Canonical correlation analysis, approximate covariance extension, and identification of stationary time series, *Automatica* **32** (1996), 709–733.

[30] A. Lindquist and Gy. Michaletzky, Output-induced subspaces, invariant directions and interpolation in linear discrete-time stochastic systems, *SIAM J. Control and Optimization* **35** (1997), 810–859.

[31] M.Pavon, Stochastic control and non-Markovian Schrödinger processes,in *Systems and Networks: Mathematical Theory and Applications*„ vol. II, U. Helmke, R.Mennichen and J.Saurer Eds., Mathematical Research vol.79, Akademie Verlag, Berlin, 1994,409-412.

[32] A. Wakolbinger, Schrödinger Bridges from 1931 to 1991, in: E. Cabaña et al. (eds) , *Proc. of the 4th Latin American Congress in Probability and Mathematical Statistics*, Mexico City 1990, Contribuciones en probabilidad y estadistica matematica 3 (1992) , pp. 61-79.

[33] I. Fenyes, Z. Physik **132**, 81 (1952).

[34] E. Nelson, Derivation of the Schrödinger equation from Newtonian mechanics, *Phys. Rev.* **150** 1079 (1966).

[35] F. Guerra, Structural aspects of stochastic mechanics and stochastic field theory, *Phys.Rep.* **77** (1981) 263.

[36] K. Yasue, Stochastic calculus of variations, *J. Functional Analysis* 41 (1981), 327-340.

[37] F. Guerra and L. Morato, *Phys.Rev.D* **27**, 1774 (1983).

[38] E. Nelson. *Quantum Fluctuations*. Princeton University Press, Princeton, 1985.

[39] Ph. Blanchard, Ph. Combe and W. Zheng. *Math. and Physical Aspects of Stochastic Mechanics*. Lect. Notes in Physics vol. 281, Springer-Verlag, New York, 1987.

[40] M. Pavon, Hamilton's principle in stochastic mechanics, *J. Math. Phys.* **36** (1995), 6774.

[41] H. H. Rosenbrock, Doing quantum mechanics with control theory, *IEEE Trans. Aut. Contr.* **54** (2000), 73-77.

[42] E. Carlen, *Comm. Math. Phys.*, **94**, 293 (1984).

[43] R. Carmona, Probabilistic construction of the Nelson process, Taniguchi Symp. PMMP Katata (1985), 55-81.

[44] M. Pavon, A new formulation of stochastic mechanics, *Physics Letters A* **209** (1995) 143-149.

[45] M. Pavon, Derivation of the wavefunction collapse in the context of Nelson's stochastic mechanics, *J. Math. Physics* **40** (1999), 5565-5577.

[46] J.B. Keller, Am. J. Phys. **58**, 768 (1990).

[47] E. Wigner, Two kinds of reality, *The Monist* **48** (1964), 248-264.

[48] S. Adler, D. Brody, T. Brun and L. Hughston, Martingale models for quantum state reduction, arXiv quant-ph/0107153, July 2001.

[49] M. Pavon, under preparation.

[50] A. Beghi, A. Ferrante and M. Pavon, How to steer a quantum system over a Schrödinger bridge, *Quantum Information Processing*, **1**, n.3, June 2002.

[51] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*, Cambridge Univ. Press, Cambridge, UK, 2000.

[52] H. Rabitz, R. de Vivie-Riedle, M. Motzkus, and K. Kompa, *Science* **288** (2000), 824.

18

Segmentation of Diffusion Tensor Imagery

Eric Pichon *Guillermo Sapiro* *Allen Tannenbaum*

Abstract

Segmentation paradigms in diffusion tensor imagery (DTI) are discussed in this paper. We present a technique for determining paths of anatomical connectivity from the tensor information obtained in magnetic resonance diffusion tensor measurements. The basic idea is to construct optimal curves in 3D space, where the optimality criteria is based on the eigenvalues and eigenvectors of the tensor. These curves are constructed via partial differential equations computed using multiple level-set functions. We also discuss our current efforts in clustering in DTI.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 239-247, 2003.
© Springer-Verlag Berlin Heidelberg 2003

18.1 Introduction

Fundamental advances in understanding complex biological systems require detailed knowledge of structural and functional organization in the living system. In the case of the human brain for instance, anatomical connections are related to the information pathways and how this information is processed.

During the past three decades, the neuroscience and medical communities have witnessed a tremendous growth in the field of in-vivo, non-invasive imaging of brain function. Magnetic Resonance Imaging (MRI) evolved into the modality of choice for neuroradiological examination, due to its ability to visualize soft tissue with exquisite anatomical detail and contrast.

However the resolution of MRI is well above the dimension of neural fibers and the current understanding of the nervous system is still incomplete because of the lack of fundamental connectivity information.

Diffusion Tensor MRI (DT-MRI) adds to conventional MRI the capability of measuring the random motion (diffusion) of water molecules due to intrinsic thermal agitation [2, 3]. In highly structured tissues containing a large number of fibers, like skeletal muscle, cardiac muscle, and brain white matter, water diffuse fastest along the direction that the fibers are pointing in, and slowest at right angles to it. By taking advantage of the very structure of such tissues, DT-MRI can be used to track fibers well below the resolution of conventional MRI.

Information obtained by DT-MRI consists of the diffusivities and orientations of the local principal axes of diffusion for each voxel. A wide range of techniques have been explored to provide explicit connection information from this tensor field. Early work [13] attempted to use a similarity measure to group together neighboring voxels. Other methods [4, 21] follow locally the direction of highest diffusion (this is closely related to the so called "hyperstreamline" method of [10] for tensor field visualization). In [18] a Markovian approach is used to regularize the candidate curves.

In this paper we discuss a technique to compute the anatomical paths which is based on 3D optimal curves computed via multiple level-set functions. The ideas are based on prior work on geodesic active contours [7, 9, 14], combined with numerical techniques developed in [5, 6]. The basic idea is that given two user marked points, we construct a 3D optimal-effort curve connecting these points. The effort is based on weights obtained from the diffusion tensor. The computational construction of these curves is based on representing it as the intersection of two 3D surfaces, and evolving these surfaces via the techniques developed in [5, 6]. Alternatively, one could use the work introduced in [15] for this computation. Note that the fast techniques in [12, 19, 20, 23] can not be used in the general case we discuss below due to the type of energy we use. This is in contrast with the work in [17], where the energy is artificially modified to fit the requirements for using these fast numerical approaches. It is interesting to extend the work in [9] to be able to incorporate directionality, as done below, into the penalty function, thereby permitting the use of fast numerical techniques. If the images need to be regularized prior to the geodesic computation, the approaches in [8, 18] could be used for example (see also [22] for a general theory for denoising non-flat features).

18.2 Active Contours and Diffusion Tensor Imaging

Once we have enhanced the DTI, e.g., via the techniques in [8, 18], we can use this to construct the flow paths (fiber tracking), e.g., for visualization. The basic idea for this is to use our prior work on geometric active contours [7, 14], as well as [9].

Brief Background on Geodesic Snakes

We briefly review some of the relevant results from [7, 14] now. We work in the plane for simplicity. All of the results extend to $\mathbf{R}^3$ as well. We first define a positive stopping term $\phi : \mathbf{R}^2 \to \mathbf{R}$ which will act as a conformal factor in the new metric we consider for our snake model. For example, the term $\phi(x, y)$ may be chosen to be small near an edge, and so acts to stop the evolution when the contour gets close to an edge. Hence, one may take

$$\phi := \frac{1}{1 + \|\nabla G_\sigma * I\|^2}, \tag{18.1}$$

where I is the (grey-scale) image and G_σ is a Gaussian smoothing filter.

We use ϕ to modify the ordinary Euclidean arc-length function along a curve $C = (x(p), y(p))^T$ with parameter p given by

$$ds = (x_p^2 + y_p^2)^{1/2}\, dp,$$

to

$$ds_\phi = (x_p^2 + y_p^2)^{1/2} \phi\, dp.$$

Then we want to compute the corresponding gradient flow for shortening length relative to the new metric ds_ϕ.

Accordingly set

$$L_\phi(t) := \int_0^1 \|\frac{\partial C}{\partial p}\|\, \phi\, dp.$$

Let

$$\vec{T} := \frac{\partial C}{\partial p} / \|\frac{\partial C}{\partial p}\|\,,$$

denote the unit tangent. Then taking the first variation of the modified length function L_ϕ, and using integration by parts, we get that

$$L'_\phi(t) \quad = \quad -\int_0^{L_\phi(t)} \langle \frac{\partial C}{\partial t}, \phi\kappa\vec{N} - \nabla\phi \rangle\, ds$$

which means that the direction in which the L_ϕ perimeter is shrinking as fast as possible is given by

$$\frac{\partial C}{\partial t} = \phi\kappa\vec{N} - \nabla\phi$$

Since we can ignore the tangential component of the speed $\frac{\partial C}{\partial t}$ when evolving the curve C, this flow is geometrically equivalent to :

$$\frac{\partial C}{\partial t} = (\phi\kappa - \nabla\phi \cdot \vec{N}))\vec{N}. \tag{18.2}$$

This is precisely the weighted L^2-gradient flow corresponding to the minimization of the length functional L_ϕ. The level set, [16], version of this is

$$\frac{\partial \Psi}{\partial t} = \phi\|\nabla\Psi\|\mathrm{div}(\frac{\nabla\Psi}{\|\nabla\Psi\|}) + \nabla\phi \cdot \nabla\Psi. \tag{18.3}$$

One expects that this evolution should attract the contour very quickly to the feature which lies at the bottom of the potential well described by the gradient flow (18.3). Notice that for ϕ as in (18.1), $\nabla\phi$ will look like a doublet near an edge. Of course, one may choose other candidates for ϕ in order to pick out other features. This will be done for diffusion tensor images next.

Geodesic Snakes and Diffusion Tensor Imagery

For the case of DTI, we use a combination of the principal direction of the tensor with a measurement of anisotropy to define ϕ and construct curves that will indicate the principal direction of flow (fiber tracking). Note that this can be combined with our prior work [1], where we have shown how to smoothly construct and complete curves from partial tangent data. The explicit use of a directionality constraint limits the computational techniques that can be used to find the optimal curve.

Note that in contrast with what is primarily done in the literature, when only single slices are used, we use multiple-slices (and then 3D) for these works. For this we use the computational technique developed in [5, 6], where the 3D active contour that is deforming toward the minima of the energy is represented as the intersection of two 3D surfaces. We will then need to move 3D curves, with fix end points, having the curve represented as the intersection of two deforming surfaces.

To each point in the domain $\Omega \subset \mathbf{R}^3$ we associate a 3×3 positive semidefinite symmetric matrix $A(x, y, z)$ with (real eigenvalues) $\lambda_i(x, y, z) = \lambda_i,\ i = 1, 2, 3$ and associated unit eigenvectors $\varepsilon_i(x, y, z) = \varepsilon_i,\ i = 1, 2, 3$. We always assume that $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$.

We define the fractional anisotropy to be [17, 18]:

$$\phi(x, y, z) := \frac{\sqrt{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_1 - \lambda_3)^2}}{\sqrt{2}\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}.$$

(See our discussion below for some properties of this function.) We will also consider

$$v(x, y, z) := \phi(x, y, z)\, \varepsilon_1(x, y, z).$$

Diffusion Flow

In this section, we will formulate the flow which will move an arbitrary initial curve with given endpoints to a curve with respect to the weighted conformal metric defined by v. We state the results in both the plane and in space.

In what follows, we assume that we are given an embedded family of differentiable curves $C(p,t) : [0,1] \to \mathbf{R}^3$ where p is the curve parameter (independent of t), and t denotes the parameter of the family. The arc-length parametrization will be denoted by ds so that

$$ds = \sqrt{x_p^2 + y_p^2 + z_p^2}\, dp.$$

We can now state our result:

THEOREM 18.1
Let ϕ, ε_1 and v be as above. Consider the energy functional

$$L_A(t) := \frac{1}{2} \int_0^L \|\varepsilon_1 - C_s\|^2 \phi \, ds.$$

By minimizing $L_A(t)$, the following flow is obtained :

$$C_t = \phi\, C_{ss} - \mathrm{curl}(v) \times C_s - \nabla\phi. \tag{18.4}$$

□

Proof.

We note that

$$\begin{aligned} L_A(t) &= \frac{1}{2} \int_0^L \langle \varepsilon_1 - C_s, \varepsilon_1 - C_s \rangle \, \phi \, ds \\ &= \frac{1}{2} \int_0^L [\langle \varepsilon_1, \varepsilon_1 \rangle + \langle C_s, C_s \rangle - 2\langle \varepsilon_1, C_s \rangle] \, \phi \, ds \\ &= \underbrace{\int_0^L \phi \, ds}_{L_A^1(t)} - \underbrace{\int \langle \varepsilon_1, C_s \rangle \, \phi \, ds}_{L_A^2(t)} \end{aligned}$$

As above, we can compute that the first variation of $L_A^1(t)$ is :

$$L_A^{1'}(t) = -\int_0^L \langle \phi \; C_{ss} - \nabla\phi, C_t \rangle \, ds$$

In order to compute $L_A^{2'}(t)$, set

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = v.$$

Then since

$$L_A^2(t) = \int_0^1 (ax_p + by_p + cz_p)\, dp,$$

we have that (integrating by parts)

$$L_A^{2'}(t) = \int_0^1 [\nabla a \cdot \begin{pmatrix} x_t \\ y_t \\ z_t \end{pmatrix} x_p + \nabla b \cdot \begin{pmatrix} x_t \\ y_t \\ z_t \end{pmatrix} y_p + \nabla c \cdot \begin{pmatrix} x_t \\ y_t \\ z_t \end{pmatrix} z_p - a_p x_t - b_p y_t - c_p z_t]\, dp.$$

The expression in the integral is

$$(b_x y_p + c_x z_p - a_y y_s - a_z z_p)x_t + (a_y x_p + c_y z_p - b_x x_p - b_z z_p)y_t + (a_z x_p + b_z y_p - c_x x_p - c_y y_p)z_t. \tag{18.5}$$

Noting that

$$\text{curl}(v) = \begin{pmatrix} c_y - b_z \\ a_z - c_x \\ b_x - a_y \end{pmatrix},$$

we see that we may write (18.5) as

$$-\langle \text{curl}(v) \times \begin{pmatrix} x_p \\ y_p \\ z_p \end{pmatrix}, C_t \rangle,$$

and so

$$L_A^{2'}(t) = -\int \langle \text{curl}(v) \times \begin{pmatrix} x_p \\ y_p \\ z_p \end{pmatrix}, C_t \rangle\, ds.$$

Since

$$L_A'(t) = L_A^{1'}(t) - L_A^{2'}(t),$$

the theorem follows. QED

Remarks:

1. The above energy is minimum when the tangent of the curve C is most closely aligned with the direction of the principal eigenvector ε_1. Moreover this constraint is weighted by the anisotropy ϕ defined previously. When A is almost isotropic, we have $\lambda_1 \approx \lambda_2 \approx \lambda_3$. In this region, $\phi \approx 0$ ensures that we will not penalize a curve that would not be perfectly aligned with ε_1 (which here cannot be considered the unique direction of diffusion). On the other hand, if $\lambda_1 \gg \lambda_2 \geq \lambda_3$ there is no ambiguity : ε_1 is the preferred direction for diffusion and $\phi \approx 1$ ensures that the curve will follow closely.

2. In two dimensions, we can consider that $c = 0$ and $a_z = b_z = 0$. As is standard we define curl_{2D} to be the scalar $b_x - a_y$. We set $\text{curl}_{3D}(v) := \text{curl}(v)$. Therefore the following relation holds :

$$\begin{aligned} \text{curl}_{3D}(v) &= \begin{pmatrix} 0 \\ 0 \\ b_x - a_y \end{pmatrix} \\ &= \text{curl}_{2D} \begin{pmatrix} a \\ b \end{pmatrix} e_z \end{aligned}$$

Since $e_z \times \vec{T} = \vec{N}$ ($\vec{T}$ is the unit tangent and $\vec{N}$ the unit normal), we get the flow

$$C_t = (\phi\kappa - \langle \nabla\phi, \vec{N} \rangle - \text{curl}_{2D}(v))\vec{N} \tag{18.6}$$

Note that by the standard Frenet formulas in the plane

$$C_{ss} = \kappa\vec{N},$$

so that equations (18.4) and (18.6) are consistent.

3. The above curve deformation is implemented in 3D deforming the intersecting of two 3D surfaces. In addition, the end points of the deforming curve are fixed.

4. In case an advanced initialization is needed, we can use for example the technique in [17]. We are also investigating the use of the geodesics obtained from just $\int \phi ds$, which can be computed using fast numerical techniques, to initialize the flow.

5. The above described technique can be used for finding discrete connectivity lines. We are currently also working on the use of techniques such as those in [11] to cluster the diffusion direction information.

18.3 Conclusions

In this paper, we discussed a geodesic snake technique for the segmentation of diffusion tensor imagery. DTI is an increasing important non-invasive methodology which can reveal white matter bundle connectivity in the brain. As such it is useful in neuroscience as well as image guided surgery and therapy.

Using the conformal metric ideas we derived an explicit flow for the segmentation of such imagery in three dimensions. In future work, we plan to test our flow on some explicit examples. Level set ideas will of course be very important in the computer implementation and the development of fast reliable algorithms based on the equation (18.4).

Acknowledgment

This work was supported by grants from AFOSR, NSF, ONR, and MRI.

18.4 Bibliography

[1] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and grey levels," *IEEE Trans. Image Processing* **10**, pp. 1200-1211, August 2001.

[2] P. J. Basser, "Inferring microstructural features and the physiological state of tissues from diffusion-weighted images," *NMR Biomed* **8**, pp. 333-344, 1995.

[3] P.J. Basser, J. Mattiello, and D. LeBihan "MR diffusion tensor spectroscopy and imaging" *Biophys J* **66** (1994), pp 259-267

[4] P.J. Basser, S. Pajevic, C. Pierpaoli, J. Duda and A. Aldroubi "In vivo fiber tractography using DT-MRI data" *Magn. Reson. Med.* **44** (2000) pp. 625-632

[5] M. Bertalmio, G. Sapiro, and G. Randall, "Region tracking on level-sets methods," *IEEE Trans. Medical Imaging* **18**, pp. 448-451, 1999.

[6] P. Burchard, L-T. Cheng, B. Merriman, and S. Osher, "Motion of curves in three spatial dimensions using a level set approach," *UCLA CAM Report* **00-29**, July 2000

[7] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision* **22:1**, pp. 61-79, 1997.

[8] C. Chef d'hotel, D. Tschumperle, R. Deriche, and O. Faugeras, "Constrained flows of matrix-valued functions: Application to diffusion tensor regularization," *Proc. ECCV-LNCS* **2350**, p. 251, Springer, 2002.

[9] L. Cohen and R. Kimmel, "Global minimum for active contours models: A minimal path approach," *International Journal of Computer Vision* **24**, pp. 57-78, 1997.

[10] T. Delmarcelle, L. Hesselink "Visualizing second-order tensor fields with hyperstreamlines" *IEEE Computer Graphics and Applications* **13**, issue 4, pp. 25-33, July 1993

[11] H. Garcke, T. Preusser, M. Rumpf, A. Telea, U. Weikard, and J. van Wijk, "A continuous clustering method for vector fields," *Proceedings IEEE Visualization 2000*, pp. 351-358, 2001.

[12] J. Helmsen, E. G. Puckett, P. Collela, and M. Dorr, "Two new methods for simulating photolithography development in 3D," *Proc. SPIE Microlithography* **IX**, pp. 253, 1996.

[13] D.K. Jones, A. Simmons, S.C.R. Williams and M.A. Horsfield "Non-invasive assessment of axonal fiber connectivity in the human brain via diffusion tensor MRI" *Magn. Reson. Med.* **42** (1999) pp. 37-41

[14] S. Kichenassamy S, A. Kumar, P. Olver, A. Tannenbaum, A. Yezzi, "Conformal curvature flows: from phase transitions to active vision," *Archives on Rational Mechanics and Analysis,* **134** (1996), pp. 275-301.

[15] L. Lorigo, O. Faugeras, E. Grimson, R. Keriven, R. Kikinis, A. Nabavi, C. Westin, "Co-dimension 2 geodesic active contours for the segmentation of tubular strucures," *Proc. of CVPR'00*, 2000.

[16] S. Osher and J. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations," *Journal of Computational Physics* **79**, pp. 12-49, 1988.

[17] G. J. Parker, C. A. M. Wheeler-Kingshott, and G. J. Barker, "Estimating distributed anatomical connectivity using fast marching methods and diffusion tensor imaging," *IEEE Trans. Medical Imaging* **21**, pp. 505-512, 2002.

[18] C. Poupon, C. A. Clark, V. Froulin, J. Regis, D. Le Bihan, and J. F. Mangin, "Regularization of diffusion-based direction maps for the tracking of brain white matter fascicles," *Neuroimage* **12**, pp. 184-195, 2000.

[19] J. Sethian, "Fast marching level set methods for three-dimensional photolithography development," *Proc. SPIE International Symposium on Microlithography*, Santa Clara, California, March, 1996.

[20] J. A. Sethian, "A fast marching level-set method for monotonically advancing fronts," *Proc. Nat. Acad. Sci.* **93:4**, pp. 1591-1595, 1996.

[21] B. Stieljes, W.E. Kaufman, P.C.M. van Zijl, K. Fredricksen, G.D. Pearlson and M. Solaiyappan *et al.* "Diffusion tensor imaging and axonal tracking in the human brainstem" *NeuroImage* **14** (2001) pp. 723-735

[22] B. Tang, G. Sapiro, and V. Caselles, "Diffusion of general data on non-flat manifolds via harmonic maps theory: The direction diffusion case," *Int. Journal Computer Vision* **36:2**, pp. 149-161, February 2000.

[23] J. N. Tsitsiklis, "Efficient algorithms for globally optimal trajectories," *IEEE Transactions on Automatic Control* **40** pp. 1528-1538, 1995.

19

Robust Linear Algebra and Robust Aperiodicity

Boris T. Polyak

Abstract

We consider some simple robust linear algebra problems which provide new insight for the robust stability and aperiodicity. Uncertainties are defined via various matrix norms, which include vector-induced and component-wise norms. First, the solution set of uncertain systems of linear algebraic equations is described. Second, the radius of nonsingularity of a matrix family is calculated. Third, these results are applied for estimation of aperiodicity radius.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 249-260, 2003.
© Springer-Verlag Berlin Heidelberg 2003

19.1 Introduction

Uncertainty plays a key role in control [1, 2, 3] as well as in numerical analysis [4, 5]. We try to present an unified framework to treat uncertainties in linear algebra. For a nominal real matrix A a family of perturbed matrices is considered in a structured form $A + B\Delta C$ where Δ is a (rectangular) real matrix, bounded in some norm. We analyze various norms, which include such widely used ones as spectral, Frobenius, interval. The main tool is given by Theorem 19.1, which validates that the set $\{\Delta a, ||\Delta|| \leq \varepsilon\}$ is a ball in some specified norm (Section 2). Based on this result, we are in position to describe a set of all solutions of perturbed systems of linear equations $(A+B\Delta C)x = b$ (Section 3). The next issue to be addressed in Section 4 is the distance to singular matrices (nonsingularity radius). Another basic problem of robust linear algebra is pseudospectrum — set of all eigenvalues of perturbed matrices. We study a real version (i.e. the set of all real eigenvalues) of this notion in Section 5. Finally we apply the results to control. Namely, we address aperiodicity property of matrices and aperiodicity robustness (Section 6).

19.2 Matrix Norms and Preliminaries

In this Section we introduce various norms for matrix perturbations and establish their properties. All vectors and matrices in the paper are assumed to be *real*.

As usual $||x||_p$ denotes l_p norm of a vector $x \in R^n, 1 \leq p \leq \infty$, i.e. $||x||_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. For matrices $A \in R^{m\times n}$ with entries $a_{ij}, i = 1,\dots,m, j = 1,\dots,n$ two types of norms are considered. The first is *induced norm*:

$$||A||_{p,q} = \max_{x\neq 0} \frac{||Ax||_q}{||x||_p} = \max_{||x||_p\leq 1} ||Ax||_q.$$

The second is *component-wise norm*:

$$||A||_p = \left(\sum_{i,j} |a_{ij}|^p\right)^{1/p}.$$

The most important examples are listed below; explicit expressions for norms can be found in the literature or easily validated. Vector $a_i \in R^n$ denotes i-th row of A.

$$||A||_{\infty,\infty} = \max_i \sum_j |a_{ij}|;$$

$$||A||_{1,1} = \max_j \sum_i |a_{ij}|;$$

$$||A||_{2,2} = \bar{\sigma}(A) = \max_i(\lambda_i(A^TA)^{1/2}),$$

here $\bar{\sigma}(A)$ is the largest singular value of A while $\lambda_i(B)$ are eigenvalues of B. This is widely used *spectral* or *operator* norm of a matrix;

$$||A||_{1,\infty} = ||A||_\infty = \max_{i,j} |a_{ij}|,$$

this norm is sometimes called *interval* one (the family of matrices $A+\Delta, ||\Delta||_{1,\infty} \le \varepsilon$ is the interval matrix);

$$||A||_{2,\infty} = \max_i (\sum_j a_{ij}^2)^{1/2};$$

$$||A||_{1,2} = \max_j (\sum_i a_{ij}^2)^{1/2};$$

$$||A||_2 = ||A||_F = (\sum_{i,j} a_{ij}^2)^{1/2},$$

this is another widely used norm — *Frobenius* one;

$$||A||_1 = \sum_{i,j} |a_{ij}|;$$

$$||A||_{\infty,1} = \max_{||x||_\infty \le 1} \sum_i |(a_i, x)|;$$

$$||A||_{\infty,2} = \max_{||x||_\infty \le 1} (\sum_i (a_i, x)^2)^{1/2};$$

$$||A||_{2,1} = \max_{||x||_2 \le 1} \sum_i |(a_i, x)|;$$

Note that the first eight formulas provide explicit (or easily calculated, as for spectral norm) expressions, while three last ones reduce calculation of the norm to the optimization problems. These problems are maximization of a quadratic or piece-wise linear convex function on a convex set. They are known to be NP-hard; moreover it is proved in [6] that calculation of $||A||_{p,q}$ is NP-hard for any $p, q \ge 2, p + q > 4$. There are computationally tractable bounds for these norms with precise estimation of their tightness [6, 7, 8, 9], but we are unable to discuss this important problem here. Nevertheless, calculation of $||A||_{\infty,1}$ or $||A||_{\infty,2}$ is not a hard task for matrices of moderate size. Indeed, the solution of above optimization problems is achieved at a vertex of the unit cube, thus it suffices to check 2^n points. For $n \le 15$ this can be performed with no difficulties.

The main property of the above introduced norms is given by the following result, which will be intensively exploited.

THEOREM 19.1
For every $x \in R^n, \varepsilon > 0$ the set $\{y = \Delta x, ||\Delta|| \le \varepsilon\}$ is a ball, specifically

$$\{y = \Delta x, \quad \Delta \in R^{m \times n}, ||\Delta||_{p,q} \le \varepsilon\} = \{y \in R^m : ||y||_q \le \varepsilon ||x||_p\}, \tag{19.1}$$

$$\{y = \Delta x, \quad \Delta \in R^{m \times n}, ||\Delta||_p \le \varepsilon\} = \{y \in R^m : ||y||_p \le \varepsilon ||x||_{p_*}\}, \tag{19.2}$$

where $1 \le p_* \le \infty$ is the index conjugate to p : $\quad 1/p + 1/p_* = 1$. □

Proof. For induced norms the inequality $||\Delta x||_q \le ||\Delta||_{p,q}||||x||_p \le \varepsilon||x||_p$ follows from the definition. For component-wise norm it can be validated via Holder inequality:

$$||\Delta x||_p^p = \sum_i |\sum_j \Delta_{ij}x_j|^p \le \sum_i (\sum_j |\Delta_{ij}|^p)||x||_{p_*}^p = ||\Delta||_p^p||x||_{p_*}^p$$

hence $||y||_p \le \varepsilon||x||_{p_*}$. On the other hand, if $y \in R^m, ||y||_q \le \varepsilon||x||_p$, then take $\Delta = yv^T$, where $v \in R^n$ is the vector such that $v^Tx = 1, ||v||_{p_*}||x||_p = 1$. Then $\Delta x = yv^Tx = y$ and for induced norms

$$||\Delta||_{p,q} = \max_{||z||_p \le 1} ||y||_q|v^Tz| \le ||y||_q||v||_{p_*} \le \varepsilon||x||_p||v||_{p_*} = \varepsilon.$$

Similarly for component-wise norms if $y \in R^m, ||y||_p \le \varepsilon||x||_{p_*}$, take $\Delta = yv^T, v \in R^n, v^Tx = 1, ||v||_p||x||_{p_*} = 1$ and for this Δ

$$||\Delta||_p = (\sum_{i,j} |y_i|^p|v_j|^p)^{1/p} = ||y||_p||v||_p \le \varepsilon||x||_{p_*}||v||_p = \varepsilon.$$

Thus for both kinds of norms the equivalence of two sets in (19.1),(19.2) is validated.

19.3 Solution Set of Perturbed Linear Algebraic Equations

Let $A \in R^{n\times n}, B \in R^{n\times m}, C \in R^{r\times n}$ be given matrices, A is nonsingular, $b \in R^n$ is a given vector, $\varepsilon > 0$. The set

$$S_\varepsilon = \{x \in R^n : \exists \Delta \in R^{m\times r}, ||\Delta|| \le \varepsilon, (A + B\Delta C)x = b\} \tag{19.3}$$

is called *the solution set* for the nominal equation $Ax = b$ under structured perturbations. The level of perturbations is given by ε and the norm $||\cdot||$ should be specified. The result below provides the closed-form expression for S_ε.

THEOREM 19.2
If $||\Delta|| = ||\Delta||_{p,q}$ then

$$S_\varepsilon = \{x = A^{-1}(b - By) : ||y||_q \le \varepsilon||CA^{-1}(b - By)||_p\} \tag{19.4}$$

and if $||\Delta|| = ||\Delta||_p$ then

$$S_\varepsilon = \{x = A^{-1}(b - By) : ||y||_p \le \varepsilon||CA^{-1}(b - By)||_{p_*}\}. \tag{19.5}$$

□

Proof. If $(A + B\Delta C)x = b$ then denoting $y = \Delta Cx$ we get $x = A^{-1}(b - By)$ and $y = \Delta CA^{-1}(b - By)$. Due to Theorem 19.1 all y satisfying the last equation with $||\Delta|| \le \varepsilon$ coincide with the set $||y|| \le \varepsilon||CA^{-1}(b - By)||$ with corresponding norms

in the left- and right-hand sides. Substituting these norms and returning to x variables we arrive to the Theorem assertions.

Let us consider some particular cases of the above result.

1. *Unstructured perturbations.* $B = C = I$. Then we can express y via $x : y = b - Ax$ and the solution set becomes

$$S_\varepsilon = \{x : ||Ax - b||_q \le \varepsilon ||x||_p\} \tag{19.6}$$

for induced norms and

$$S_\varepsilon = \{x : ||Ax - b||_p \le \varepsilon ||x||_{p_*}\} \tag{19.7}$$

for component-wise norms.

2. *Spectral or Frobenius norm.* For $p = q = 2$ (induced norms) or $p = 2$ (component-wise norm) we get the same expression

$$S_\varepsilon = \{x = A^{-1}(b - By) : ||y||_2 \le \varepsilon ||CA^{-1}(b - By)||_2\}. \tag{19.8}$$

This leads to the following conclusion.

PROPOSITION 19.1
If $\varepsilon < 1/\bar{\sigma}(CA^{-1}B)$, then the solution set for perturbations bounded in spectral or Frobenius norm is an ellipsoid. □

Proof. Under above assumption on ε the quadratic form $||y||_2^2 - \varepsilon^2 ||CA^{-1}By||_2^2$ is positive definite, thus the set of y defined by (19.8) is an ellipsoid. Hence the set S_ε is an ellipsoid as well, being a linear image of the ellipsoid.

3 *Solution set for interval equations.* Suppose $B = C = I, p = 1, q = \infty$, that is S_ε is solution set for interval equations:

$$S_\varepsilon = \{x : \exists |\Delta_{ij}| \le \varepsilon, i, j = 1, \ldots, n, \quad (A + \Delta)x = b.\} \tag{19.9}$$

Then according to Theorem 19.2

$$S_\varepsilon = \{x : ||Ax - b||_\infty \le \varepsilon ||x||_1\}. \tag{19.10}$$

This is a polytopic set; however it can be non convex for arbitrary small $\varepsilon > 0$. For instance for $n = 2, A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ the set $S_\varepsilon = \{x \in R^2 : \max\{|x_1 - 1|, |x_2|\} \le \varepsilon(|x_1| + |x_2|)\}$ is nonconvex for any $\varepsilon > 0$.

The structure of all solutions for interval equations has been described first in [10]. More details can be found in the monograph [11].

In the above analysis we supposed that the right hand side of the equation — vector b — remains unperturbed. It is not hard to incorporate more general case, but we do not address the issue here.

19.4 Nonsingularity Radius

We consider the same framework as above: let $A \in R^{n\times n}, B \in R^{n\times m}, C \in R^{r\times n}$ be given matrices, A is nonsingular. The problem is to find the margin of perturbations Δ which preserve nonsingularity of the matrix $A + B\Delta C$. More rigorously, we define the *nonsingularity radius* as

$$\rho(A) = \min\{||\Delta|| : A + B\Delta C \text{ is singular. }\}$$

The norm in the above definition should be specified; we denote the radius as $\rho(A)_{p,q}$ or $\rho(A)_p$ for induced and component-wise norms respectively. Notice that standard definition deals with complex unstructured ($B = C = I$) perturbations and spectral norm while we address real structured perturbations and arbitrary norms.

THEOREM 19.3
The nonsingularity radius is given by

$$\rho(A)_{p,q} = 1/||CA^{-1}B||_{q,p} \tag{19.11}$$

$$\rho(A)_p = 1/||CA^{-1}B||_{p,p_*} \tag{19.12}$$

□

Proof. Matrix $A + B\Delta C$ is singular if and only if the equation $(A + B\Delta C)x = 0$ has nonvanishing solution, that is if the solution set for this equation contains a nonzero point. Consider induced norms case first. From Theorem 19.2 with $b = 0$ it means that the inequality $||y||_q \le \varepsilon||CA^{-1}By||_p$ holds for $y \ne 0, \varepsilon = ||\Delta||_{p,q}$. This is equivalent to $||\Delta||_{p,q} \ge 1/||CA^{-1}B||_{q,p}$. Thus nonsingularity arises if the last inequality holds; this leads to (19.11). The case of component-wise norms is treated similarly.

Some particular cases are of interest.

1. *Unstructured perturbations, induced norms with* $p = q$. For $B = C = I, p = q$ we obtain from (19.11):

$$\rho(A)_{p,p} = 1/||A^{-1}||_{p,p}.$$

This is the classical result due to Kahan [12].

2. *Interval norm.* Taking $p = 1, q = \infty, B = C = I$ we get

$$\rho(A)_{1,\infty} = \rho(A)_\infty = 1/||A^{-1}||_{\infty,1}.$$

Say it another way, nonsingularity radius for interval perturbations is reciprocal to the $(\infty, 1)$-norm of the inverse matrix. As we have mentioned, calculation of such norm is NP-hard problem, however it requires to compute 2^n numbers. Thus the problem is tractable for moderate n, say $n \le 15$.

3. *Scalar perturbation.* If $m = r = 1, B = e_i, C = e_j^T, e_i$ is i-th ort, then $B\Delta C$ is the matrix with the only ij-th entry nonvanishing (equal to $\Delta \in R^1$) and all other entries equal to 0. Thus

$$\min\{|\Delta| : A + \Delta E \text{ is singular } \} = 1/|m_{ji}|,$$

where $E = ((e_{kl})), k, l = 1, \ldots, n, e_{ij} = 1, e_{kl} = 0, (k, l) \ne (i, j), A^{-1} = ((m_{kl})), k, l = 1, \ldots, n$.

19.5 Real Pseudospectrum

We proceed to investigation of spectrum of perturbed matrices, which is often called *pseudospectrum*. In contrast with numerous works on this subject [13, 14, 15] we deal with *real* perturbations and eigenvalues. We call *(real) pseudospectrum* of a matrix $A \in R^{n\times n}$ the set

$$\Lambda_\varepsilon(A) = \{\lambda \in R^1 : \exists \Delta \in R^{m\times r}, ||\Delta|| \le \varepsilon, \lambda \text{ is an eigenvalue of } A + B\Delta C.\} \quad (19.13)$$

The level of perturbation $\varepsilon > 0$ and the norm in the above definition should be specified; we denote the pseudospectra as $\Lambda_\varepsilon(A)_{p,q}$ or $\Lambda_\varepsilon(A)_p$ for induced and component-wise norms respectively. Below we use notation

$$G(\lambda) = C(A - \lambda I)^{-1}B$$

THEOREM 19.4
The real pseudospectra is given by

$$\Lambda_\varepsilon(A)_{p,q} = \{\lambda \in R^1 : 1/||G(\lambda)||_{q,p} \le \varepsilon\} \quad (19.14)$$

$$\Lambda_\varepsilon(A)_p = \{\lambda \in R^1 : 1/||G(\lambda)||_{p,p_*} \le \varepsilon\}. \quad (19.15)$$

□

Proof. λ is an eigenvalue of a matrix $A + B\Delta C$ if and only if $A - \lambda I + B\Delta C$ is singular. Thus we can apply Theorem 19.3 to the matrix $A - \lambda I$; λ belongs to the pseudospectrum if and only if $\varepsilon \le \rho(A - \lambda I)$. This coincides with the assertion of Theorem 19.4.

COROLLARY 19.1
Suppose matrix $A \in R^{n\times n}$ has no real eigenvalues. Then

$$\min\{||\Delta||_{p,q} : A + B\Delta C \quad \text{has a real eigenvalue}\} = \min_{\lambda\in R^1}\{1/||G(\lambda)||_{q,p}\} \quad (19.16)$$

$$\min\{||\Delta||_p : A + B\Delta C \quad \text{has a real eigenvalue}\} = \min_{\lambda\in R^1}\{1/||G(\lambda)||_{p,p_*}\}. \quad (19.17)$$

□

For Frobenius norm we can provide another characterization of the distance to matrices with real eigenvalues, which avoids λ gridding. Denote

$$D_\varepsilon = \begin{pmatrix} A & -\varepsilon BB^T \\ -\varepsilon CC^T & A^T \end{pmatrix}.$$

PROPOSITION 19.1

$$\varepsilon^* = \min\{||\Delta||_2 : A + B\Delta C \quad \text{has a real eigenvalue}\} \quad (19.18)$$

$$= \min\{\varepsilon \in R^1 : D_\varepsilon \quad \text{has a real eigenvalue}\}. \quad (19.19)$$

Moreover, matrix D_ε has a real eigenvalues for all $\varepsilon \ge \varepsilon^*$. □

Proof. Optimal Δ in (19.18) is the solution of the optimization problem

$$\min ||\Delta||_2^2, \quad (A + B\Delta C)x = \lambda x, \quad x \neq 0, \lambda \in R^1, x \in R^n. \tag{19.20}$$

The Lagrange function for the problem is

$$L(x, \Delta, y) = ||\Delta||_2^2 + (y, (A + B\Delta C)x - \lambda x).$$

Writing the derivatives with respect to the vector variable x and the matrix variable Δ we have

$$L_\Delta = \Delta + B^T y x^T C^T = 0,$$

$$L_x = (A^T + C^T \Delta^T B^T)y - \lambda y = 0.$$

Excluding Δ we obtain equations

$$Ax - ||Cx||_2^2 B B^T y = \lambda x \tag{19.21}$$

$$A^T y - ||B^T y||_2^2 C^T C x = \lambda y. \tag{19.22}$$

These equations remain invariant if one replaces x with $\gamma x, y$ with y/γ for arbitrary $\gamma \neq 0$. We can choose γ so that $||Cx||_2 = ||B^T y||_2$. Recall that $\Delta = -B^T y x^T C^T$, thus for optimal solution $||\Delta||_2 = ||Cx||_2 ||B^T y||_2 = \varepsilon^*$ and (19.21), (19.22) becomes

$$Ax - \varepsilon^* B B^T y = \lambda x \tag{19.23}$$

$$A^T y - \varepsilon^* C^T C x = \lambda y. \tag{19.24}$$

Now consider the real eigenvalue problem (19.19): $D_\varepsilon w = \lambda w$; for $w = (u, v)$ it can be written as

$$Au - \varepsilon B B^T v = \lambda u \tag{19.25}$$

$$A^T v - \varepsilon C^T C u = \lambda v \tag{19.26}$$

and completely coincides with (19.21), (19.22) for $\varepsilon = \varepsilon^*$. Thus if ε^* is the solution of (19.18) then D_{ε^*} has a real eigenvalue. Multiplying (19.25) by v and (19.26) by u we conclude that $||B^T v||_2 = ||Cu||_2$. If we take

$$x = \sqrt{\varepsilon} u / ||Cu||_2, \quad y = \sqrt{\varepsilon} v / ||B^T v||_2,$$

then such x, y satisfy (19.21), (19.22) and $\Delta = -B^T y x^T C^T$ has the norm equal to ε. Thus existence of a real eigenvalue of the matrix D_ε implies that optimality conditions (19.21), (19.22) hold.

To prove the last assertion of Proposition 19.1 we can rewrite (19.25), (19.26) as

$$(A - \lambda I)u = \varepsilon B B^T v \tag{19.27}$$

$$(A - \lambda I)v = \varepsilon C^T C u \tag{19.28}$$

or denoting $t = Cu$

$$\varepsilon^{-2}t = G(\lambda)G(\lambda)^T t, \quad G(\lambda) = C(A - \lambda I)^{-1}B.$$

Thus ε^{-2} is an eigenvalue of the nonnegative definite matrix $G(\lambda)G(\lambda)^T$ if and only if (19.25), (19.26) hold. We conclude that if $\varepsilon \geq 1/\max_\lambda \overline{\sigma}(G(\lambda)) = \varepsilon^*$ then D_ε has a real eigenvalue.

Let us apply the above expressions of matrix pseudospectrum for finding zero sets of perturbed polynomials. If a family of polynomials has the form

$$\mathcal{P}_p = \{P(s,a) = a_1 + a_2 s + \cdots + a_n s^{n-1} + s^n, \quad ||a - a^*||_p \leq \varepsilon\} \tag{19.29}$$

where $a^* \in R^n$ are the coefficients of the nominal polynomial $P(s,a^*)$, then its *(real) zero set* is

$$Z_\varepsilon = \{\lambda \in R^1 : \exists P(s,a) \in \mathcal{P}, \quad P(\lambda,a) = 0\}. \tag{19.30}$$

Such sets are sometimes called *spectral sets*, see [3]; usually their complex counterparts for $p = 2$ are studied. To apply Theorem 19.4 for calculation of Z_ε let us take $m = 1, r = n, C = I$ and

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \dots & -a_1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \quad \Delta = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_{n-1} \\ \delta_n \end{pmatrix}^T.$$

Then $||\Delta||_p$ is the same as $||\delta||_p, \delta \in R^n$ and the matrix $A + B\Delta C$ has the same form as A with the last row a replaced with $a + \delta$. Hence the eigenvalues of $A + B\Delta C$ are equal to the zeros of $P(s, a + \delta)$ and the zero set coincides with pseudospectrum. It is easy to show that

$$G(\lambda) = (1, \lambda, \lambda^2, \dots, \lambda^n)^T / P(\lambda, a)$$

Then exploiting Theorem 19.4 we obtain:

PROPOSITION 19.2
The zero set of a polynomial family $\mathcal{P}_p$ is equal to

$$Z_\varepsilon = \{\lambda : |P(\lambda,a)| \leq \varepsilon ||(1, \lambda, \lambda^2, \dots, \lambda^n)^T||_{p_*}\} \tag{19.31}$$

□

For $p = 2$ this result (and its complex extension) was known [3], Theorem 16.3.4.

19.6 Aperiodicity Radius

A matrix $A \in R^{n\times n}$ is called *aperiodic*, if its eigenvalues are all real, negative and distinct. Such matrices play role in control, because solutions of a system $\dot{x} = Ax$ with aperiodic A are stable and do not oscillate (each component of $x(t)$ change sign not more than n times). *Robust aperiodicity* problem is to check aperiodicity of a family of perturbed matrices $A + B\Delta C, ||\Delta|| \leq \varepsilon$. The similar problem for polynomials is well studied, see e.g. [16, 17, 18]. However the matrix version of the problem remained open; just particular results have been obtained [19].

Define *aperiodicity radius* for an aperiodic matrix A as

$$\nu(A) = \min\{||\Delta|| : A + B\Delta C \quad \text{is not aperiodic}\}. \tag{19.32}$$

For specific norm $||\cdot||_{p,q}$ or $||\cdot||_p$ we obtain $\nu(A)_{p,q}$ and $\nu(A)_p$ respectively. If $\lambda_i \in R^1, i = 1,\dots,n$ are eigenvalues of A, then the functions

$$\phi(\lambda)_{p,q} = 1/||G(\lambda)||_{p,q}, \quad \phi(\lambda)_p = 1/||G(\lambda)||_{p,p_*}$$

are vanishing at points λ_i and positive for all other λ. Computationally it is not hard to find

$$\phi^i_{p,q} = \max_{\lambda_i\leq\lambda\leq\lambda_{i+1}} \phi(\lambda)_{p,q}, \phi^i_p = \max_{\lambda_i\leq\lambda\leq\lambda_{i+1}} \phi(\lambda)_p$$

for $i = 1,\dots,n-1$.

THEOREM 19.5
Aperiodicity radius is estimated by formulas

$$\nu(A)_{p,q} \geq \min\{\phi^1_{p,q},\dots,\phi^{n-1}_{p,q},\phi(0)_{p,q}\} \tag{19.33}$$

$$\nu(A)_p \geq \min\{\phi^1_p,\dots,\phi^{n-1}_p,\phi(0)_p\} \tag{19.34}$$

□

Proof. It follows from Theorem 19.4 that for ε less than right hand side of (19.33), (19.34) the corresponding pseudospectrum consists of n distinct intervals, all located in the negative half-axis. Thus matrices $A + B\Delta C, ||\Delta|| \leq \varepsilon$ remain aperiodic.

We conjecture that the lower bound in the Theorem coincides with the upper bound, i.e. equality holds in (19.33), (19.34).

Of course the above result can be easily extended to *discrete-time aperiodicity* when eigenvalues λ_i of A are assumed to be real, distinct and $-1 < \lambda_i < 0$. Then we should include $\phi(-1)$ in the right hand side of (19.33), (19.34).

Also we can obtain results on robust aperiodicity of polynomials by using the same technique as at the end of the previous Section.

19.7 Conclusions

In the paper we presented an unified approach to analysis of perturbations for typical problems of linear algebra (solving systems of equations, checking nonsingularity, finding of eigenvalues). However this approach was restricted

with real perturbations and real eigenvalues only. While the transition to complex perturbations often simplifies the research, the case of complex eigenvalues and real perturbations is much harder. For instance, the challenging problem of finding real stability radius has been solved just recently [20] for spectral norms only. From technical point of view the difference is that critical perturbations are rank-one matrices in the real case (as in this paper) and rank-two matrices in complex case (as in [20]). Of course the presented approach allows to obtain various bounds for the stability radius and related problems with different norms (including the famous problem of robust stability of interval matrices). This is the direction for future work.

19.8 References

[1] Zhou K., Doyle J.C., Glover K. *Robust and Optimal Control,* Upper Saddle River, Prentice Hall, 1996.

[2] Bhattacharyya S.P., Chapellat H., Keel L.H. *Robust Control: the Parametric Approach,* Upper Saddle River, Prentice Hall, 1995.

[3] Barmish B.R. *New Tools for Robustness of Linear Systems,* New York, Macmillan, 1994.

[4] Higham N.J. *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[5] Golub G.H., Van Loan C.F. *Matrix Computations,* Baltimore, John Hopkins University Press, 1989.

[6] Ben-Tal A., Nemirovski A. On approximating matrix norms, *Math. Progr.,* (submitted).

[7] Nesterov Yu. Semidefinite relaxation and nonconvex quadratic optimization, *Optimization and Software,* 1998, **9,** 141–160.

[8] Nesterov Yu. Nonconvex quadratic optimization via conic relaxation, in: *Handbook on Semidefinite Programming,* R.Saigal, H.Wolkowicz and L.Vandenberghe eds., Kluwer, 2000, pp. 363–387.

[9] Nemirovski A., Polyak B. Radius of nonsingularity under structured perturbations (under preparation).

[10] Oettli W., Prager W. Compatability of approximate solutions of linear equations with given error bounds for coefficients and right hand side, *Numer. Math.,* 1964, **6,** 405–409.

[11] Neumaier A. *Interval Methods for Systems of Equations,* Cambridge, Cambridge University Press, 1990.

[12] Kahan W. Numerical linear algebra, *Canadian Math. Bull.,* 1966, **9,** 757–801.

[13] Pseudospectra gateway: `www.comlab.ox.ac.uk/pseudospectra`.

[14] Trefethen L.N. Pseudospectra of matrices, in: *Numerical analysis,* D.F.Griffith and G.A. Watson eds., Harlow, Longman, 1992, pp. 234–266.

[15] Trefethen L.N. Pseudospectra of linear operators, *SIAM Review,* 1997, **30**, 383–406.

[16] Soh C.B., Berger C.S. Strict aperiodic property of polynomials with perturbed coefficients, *IEEE Trans. Autom. Contr.,* 1989, **34**, 546–549.

[17] Polyak B.T., Tsypkin Ya.Z. Robust aperiodicity, *Phys. Dokl.*, 1994, **39**, 149–152.

[18] Meerov M.V., Jury E.I. On aperiodicity robustness, *Intern. Journ. Contr.*, 1998, **70**, 193–201.

[19] Rohn J., Stability of interval matrices: the real eigenvalue case, *IEEE Trans. Autom. Contr.*, 1992, **37**, 1604–1605.

[20] Qiu L., Bernhardson B., Rantzer A., Davison E.J., Young P.M., Doyle J.C. A formula for computation of the real stability radius, *Automatica,* 1995, **31,** 879-890.

20

On Homogeneous Density Functions

Stephen Prajna *Anders Rantzer*

Abstract

We consider homogeneous density functions for proving almost global attractivity of the zero equilibrium in a homogeneous system. It is shown that the existence of such a function is guaranteed when the equilibrium is asymptotically stable, or in the more general case, when there exists a nonhomogeneous density function for the same system satisfying some reasonable conditions. Results related to robustness under nonhomogenizing perturbations are also presented.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 261-274, 2003.
© Springer-Verlag Berlin Heidelberg 2003

20.1 Introduction

A condition for almost global attractivity of the zero equilibrium in a dynamical system, based on the existence of the so-called density functions, has been recently introduced [7]. This condition can be regarded as a dual of the Lyapunov stability theorem, and proves that *almost all* trajectories of the system converge to the origin. Numerous results related to this convergence criterion have also been obtained. For example, controller synthesis has been considered in [9, 10], and a converse theorem has been derived in [8].

The present paper focuses on *homogeneous* dynamical systems [3, 5, 12, 6, 1] whose equilibrium at the origin has the almost global attractivity property. It particularly addresses the question of existence of homogeneous density functions for such a system. As a matter of fact, on the Lyapunov side, a homogeneous Lyapunov function is guaranteed to exist for a homogeneous system whose zero equilibrium is asymptotically stable [11] (see also [2, 4]). Therefore it is natural to expect that a similar result holds in the case of density functions. We show in this paper that the existence of a homogeneous density function can be guaranteed when the origin is asymptotically stable. For the more general case, we prove that if there exists a nonhomogeneous density function, then there exists also a homogeneous density function for the same system, under some reasonable assumptions on the decay of the nonhomogeneous density function and the positivity of the corresponding divergence condition.

Study of homogeneous systems is appealing since homogeneous systems are natural approximation for more arbitrary class of systems. The simplest example is the use of a linear vector field for approximating a nonlinear one. Furthermore, some properties of a homogeneous system are usually preserved (in a local sense) when the vector field is perturbed by nonhomogenizing terms. Issues related to this will also be addressed in the present paper.

The outline of the paper is as follows. Some preliminaries on density functions and homogeneous dynamical systems are presented in Section 20.2. In Section 20.3, we will first provide three motivating examples which hint at the existence of homogeneous density functions for homogeneous systems. Then the main results of the paper are established. Finally, we consider the effect of higher and lower order term perturbations on the attractivity property in Section 20.4.

20.2 Preliminaries

Throughout the paper, we consider systems of the form $\dot{x} = f(x)$, with $x \in \mathbb{R}^n$, $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, and $f(0) = 0$, unless noted otherwise. The flow of such a system is denoted by $\phi_t(x_0)$, with the initial condition $x(0) = x_0$.

First, a condition for almost global attractivity of the origin is provided by the following theorem.

THEOREM 20.1—[7]
Given a system $\dot{x} = f(x)$, $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, suppose there exists $\rho \in C^1(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$ such that

(i) $\rho(x) \geq 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$,
(ii) $\nabla \cdot (\rho f) > 0$ almost everywhere,

(iii) $\frac{\rho(x)f(x)}{\|x\|}$ is integrable on $\{x \in \mathbb{R}^n : \|x\| \geq 1\}$,

then for almost all initial states $x(0)$ the trajectory $x(t)$ exists for $t \in [0, \infty)$ and tends to zero as $t \to \infty$. Moreover, if the origin is stable, then the conclusion remains valid even if ρ takes negative values. □

We also have the following lemma, which is related to Liouville's theorem.

LEMMA 20.2—[7, 8]
Consider a system $\dot{x} = f(x)$, $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Assume that the integral

$$\rho(x) = \int_0^\infty \psi(\phi_{-t}(x)) \left| \frac{\partial \phi_{-t}(x)}{\partial x} \right| dt$$

be well-defined on $\mathbb{R}^n \setminus \{0\}$, and that ρ is integrable on an open set $D \subset \mathbb{R}^n$. Then for all measurable $Z \subset D$ and $t \geq 0$ such that $\phi_\tau(Z) \subset D$, $\forall \tau \in [0, t]$, the following holds:

$$\int_{\phi_t(Z)} \rho(x)dx - \int_Z \rho(x)dx = \int_0^t \int_{\phi_\tau(Z)} \psi(x)dxd\tau.$$

Furthermore, if $\rho \in C^1(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$, then $[\nabla \cdot (\rho f)](x) = \psi(x)$ almost everywhere. □

We will now review some basic definitions and properties of homogeneous functions and homogeneous dynamical systems that will be needed in the subsequent sections. We have the following definitions, which are quite standard.

DEFINITION 20.3
A function $g : \mathbb{R}^n \to \mathbb{R}$ is homogeneous of degree $k \in \mathbb{R}$ with respect to the one-parameter dilation $\Delta_\lambda^r : (x_1, ..., x_n) \mapsto (\lambda^{r_1}x_1, ..., \lambda^{r_n}x_n)$, where $r_i \in [1, \infty)$, $i = 1, ..., n$, if $\forall x \in \mathbb{R}^n \setminus \{0\}$ and $\forall \lambda > 0$,

$$g(\Delta_\lambda^r x) = \lambda^k g(x). \tag{20.1}$$

□

DEFINITION 20.4
A vector field $f : \mathbb{R}^n \to \mathbb{R}^n$ with components $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, ..., n$, is homogeneous of degree $\ell \in \mathbb{R}$ with respect to the one-parameter dilation Δ_λ^r if f_i is homogeneous of degree $\ell + r_i$ with respect to Δ_λ^r. □

DEFINITION 20.5
A continuous map $x \mapsto \|x\|_h$ from $\mathbb{R}^n \to \mathbb{R}$ is called a homogeneous norm with respect to Δ_λ^r if $\|x\|_h \geq 0$, $\|x\|_h = 0 \Leftrightarrow x = 0$, and $\|\Delta_\lambda^r x\|_h = \lambda \|x\|_h$ for $\lambda > 0$. □

REMARK 20.6
Without loss of generality, we have scaled the dilation exponent r in Definitions 20.3, 20.4, and 20.5 such that $r_i \geq 1$ for all i. By this scaling, we can always find for any homogeneous norm $\|.\|_h$ a constant C such that $\frac{\|x\|_h}{\|x\|} \leq C$ on $\{x \in \mathbb{R}^n : \|x\| \geq 1\}$, implying that Theorem 20.1 is still valid when the integrability condition is replaced by

(iii)′ $\frac{\rho(x)f(x)}{\|x\|_h}$ is integrable on $\{x \in \mathbb{R}^n : \|x\|_h \geq 1\}$.

In particular, the above condition is a sufficient condition for (iii) in Theorem 20.1. □

Some properties of homogeneous functions are given below.

PROPOSITION 20.7
If $g : \mathbb{R}^n \to \mathbb{R}$ is differentiable and homogeneous of degree k with respect to Δ_λ^r, then $\frac{\partial g}{\partial x_i}$ is homogeneous of degree $k - r_i$ with respect to Δ_λ^r. □

Proof Differentiate both sides of (20.1) with respect to x_i. □

PROPOSITION 20.8
Let $S^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$. The map $\xi : (0, \infty) \times S^{n-1} \to \mathbb{R}^n \setminus \{0\}$ defined by

$$\xi : (\lambda, x) \mapsto \Delta_\lambda^r x$$

is a bijection. Furthermore, a continuous homogeneous function $g : \mathbb{R}^n \to \mathbb{R}$ is entirely determined by its values on S^{n-1}. □

Proof For a fixed $\tilde{x} \in \mathbb{R}^n \setminus \{0\}$, the map $\lambda \mapsto \sum_{i=1}^n \frac{\tilde{x}_i^2}{\lambda^{2r_i}}$ from $(0, \infty)$ into itself is decreasing and onto. Therefore, given any such $\tilde{x}$, we can find a unique λ such that $\sum_{i=1}^n \frac{\tilde{x}_i^2}{\lambda^{2r_i}} = 1$. Now choose $x_i = \frac{\tilde{x}_i}{\lambda^{r_i}}$. It follows that $x = (x_1, ..., x_n) \in S^{n-1}$ and $\xi(\lambda, x) = \tilde{x}$. This shows that ξ is a bijection.

The proof of the second statement follows directly from (20.1) and the fact that ξ is surjective. □

Finally, the flow of a homogeneous dynamical systems has the following property.

PROPOSITION 20.9
Consider a dynamical system $\dot{x} = f(x)$, where f is continuous and homogeneous of degree ℓ with respect to Δ_λ^r, and let $\phi_t(x_0)$ be the flow of the system starting at x_0. Then we have

$$\phi_t(\Delta_\lambda^r x_0) = \Delta_\lambda^r \phi_{\lambda^\ell t}(x_0). \tag{20.2}$$

□

Proof We only need to show that $\Delta_\lambda^r \phi_{\lambda^\ell t}(x_0)$ is the flow of $\dot{x} = f(x)$ starting at $\Delta_\lambda^r x_0$. For $t = 0$, it is clear that $\Delta_\lambda^r \phi_{\lambda^\ell t}(x_0) = \Delta_\lambda^r x_0$. Furthermore, we have

$$\begin{aligned} \frac{\partial \Delta_\lambda^r \phi_{\lambda^\ell t}(x_0)}{\partial t} &= \Delta_\lambda^r \lambda^\ell \frac{\partial \phi_{(\lambda^\ell t)}}{\partial (\lambda^\ell t)}(x_0) \\ &= \Delta_\lambda^r \lambda^\ell f(\phi_{\lambda^\ell t}(x_0)) \\ &= f(\Delta_\lambda^r \phi_{\lambda^\ell t}(x_0)), \end{aligned}$$

because f is homogeneous with degree ℓ. This shows that $\Delta_\lambda^r \phi_{\lambda^\ell t}(x_0)$ is indeed the flow of $\dot{x} = f(x)$, and therefore we conclude that $\phi_t(\Delta_\lambda^r x_0) = \Delta_\lambda^r \phi_{\lambda^\ell t}(x_0)$. □

20.3 Homogeneous Density Functions for Homogeneous Systems

Some examples

Existence of homogeneous density functions for homogeneous systems is hinted by the following sequence of examples. Notice that although almost all trajectories of these systems tend to the origin, the origin itself in all the three examples has different stability properties.

EXAMPLE 20.10
The following system is homogeneous w.r.t. the standard dilation:

$$\begin{aligned} \dot{x}_1 &= -x_1^3 + x_1 x_2^2, \\ \dot{x}_2 &= -x_2^3. \end{aligned}$$

The equilibrium of the system is asymptotically stable, and convergence of the trajectories follows directly. A homogeneous density function for this system is given by $\rho(x) = (x_1^2 + x_2^2)^{-6}$. Readers may wonder if a density function for a homogeneous system will necessarily be homogeneous. The answer is negative, as the same system also possesses a nonhomogeneous density function $\rho(x) = (x_1^2 + x_1^4 + x_2^2 + x_2^4)^{-6}$. □

EXAMPLE 20.11
The homogeneous system

$$\begin{aligned} \dot{x}_1 &= -x_1^3 + x_1 x_2^2, \\ \dot{x}_2 &= -x_1^2 x_2 \end{aligned}$$

has a stable equilibrium at the origin, although it is not asymptotically stable. However, almost all trajectories converge to the origin, as can be proven using $\rho(x) = (x_1^2 + 3x_2^2)^{-3}$. A phase portrait of the system is depicted in Figure 20.1. □

EXAMPLE 20.12
Finally, consider the homogeneous system

$$\begin{aligned} \dot{x}_1 &= x_1^2 - x_2^2, \\ \dot{x}_2 &= 2x_1 x_2, \end{aligned}$$

which has an unstable equilibrium at the origin. Convergence of almost all trajectories to the origin can be proven using $\rho(x) = (x_1^2+x_2^2)^{-2}(1-x_1(x_1^2+x_2^2)^{-1/2})$. A phase portrait of the system is shown in Figure 20.2. □

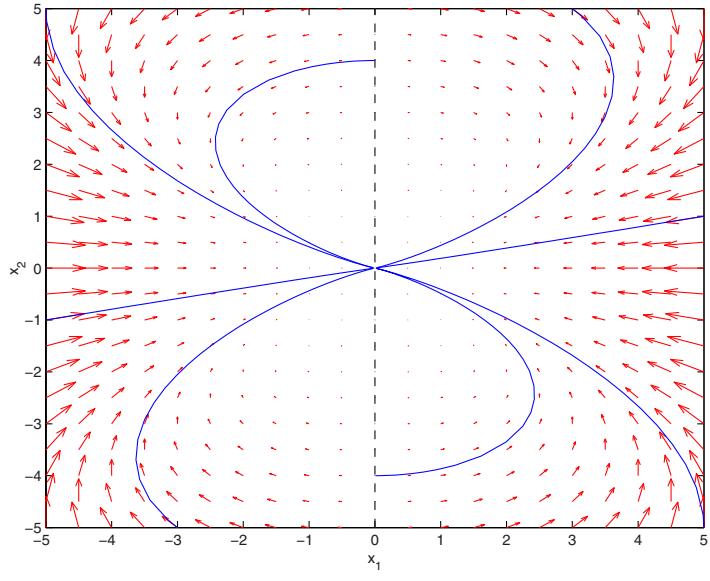


Figure 20.1 Phase portrait of the system in Example 20.11. Solid lines are trajectories of the system. The origin is a stable equilibrium, but not asymptotically stable, as any point on the x_2-axis is an equilibrium of the system.

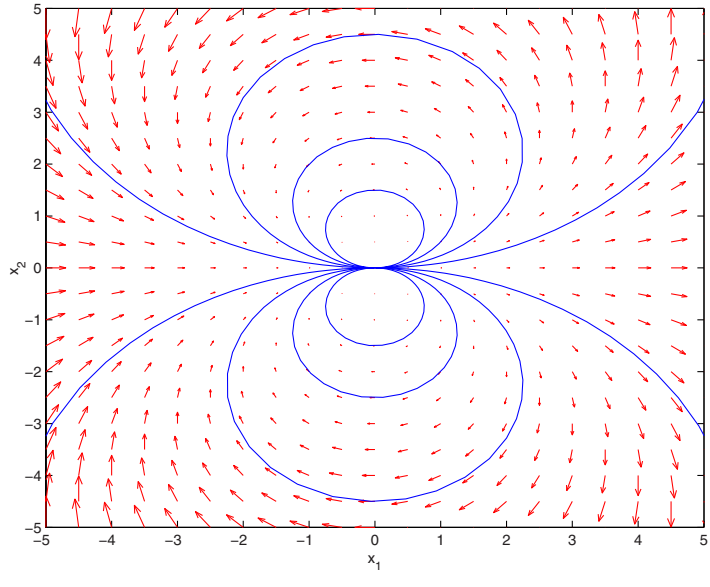


Figure 20.2 Phase portrait of the system in Example 20.12. Solid lines are trajectories of the system. The origin is an unstable equilibrium, yet almost all trajectories converge to it.

Existence of homogeneous density functions

In this subsection, we will first prove that the existence of a homogeneous density function for a homogeneous system is guaranteed in the case of asymptotically stable origin. The more general case will be treated afterwards. The following two lemmas concerning homogeneity of $\nabla \cdot (\bar{\rho} f)$ and integrability of $\frac{\bar{\rho} f}{\|x\|_h}$ will be used in the proofs of the main theorems.

LEMMA 20.13
Assume that $\bar{\rho} : \mathbb{R}^n \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ are differentiable and homogeneous respectively of degree k and ℓ. Then $\nabla \cdot (\bar{\rho} f) \triangleq \bar{\psi} : \mathbb{R}^n \to \mathbb{R}$ is homogeneous of degree $k + \ell$. Conversely, if $\bar{\psi}$ and f are homogeneous of degree k and ℓ, then under the assumption that the integral is well defined,

$$\bar{\rho}(x) = \int_0^\infty \bar{\psi}(\phi_{-t}(x)) \left| \frac{\partial \phi_{-t}(x)}{\partial x} \right| dt \tag{20.3}$$

(cf. Lemma 20.2) will be homogeneous of degree $k - \ell$. □

Proof For the first statement, notice that

$$\begin{aligned}
\bar{\psi}(\Delta_\lambda^r x) &= \nabla \cdot (f\bar{\rho})(\Delta_\lambda^r x) \\
&= \bar{\rho}(\Delta_\lambda^r x) \nabla \cdot f(\Delta_\lambda^r x) + \nabla \bar{\rho}(\Delta_\lambda^r x) \cdot f(\Delta_\lambda^r x) \\
&= \lambda^k \bar{\rho}(x) \lambda^\ell \nabla \cdot f(x) + \Delta_\lambda^{k-r} \nabla \bar{\rho}(x) \cdot \Delta_\lambda^{\ell+r} f(x) \\
&= \lambda^{k+\ell} (\bar{\rho}(x) \nabla \cdot f(x) + \nabla \bar{\rho}(x) \cdot f(x)) \\
&= \lambda^{k+\ell} \bar{\psi}(x).
\end{aligned}$$

For the converse statement, we use the result stated in Proposition 20.9. We have

$$\begin{aligned}
\bar{\rho}(\Delta_\lambda^r x) &= \int_0^\infty \bar{\psi}(\phi_{-t}(\Delta_\lambda^r x)) \left| \frac{\partial \phi_{-t}(\Delta_\lambda^r x)}{\partial (\Delta_\lambda^r x)} \right| dt \\
&= \int_0^\infty \bar{\psi}(\Delta_\lambda^r \phi_{-\lambda^\ell t}(x)) \left| \frac{\partial \Delta_\lambda^r \phi_{-\lambda^\ell t}(x)}{\partial x} \frac{\partial x}{\partial (\Delta_\lambda^r x)} \right| dt \\
&= \int_0^\infty \lambda^k \bar{\psi}(\phi_{-\lambda^\ell t}(x)) \left| \frac{\partial \phi_{-\lambda^\ell t}(x)}{\partial x} \right| dt \\
&= \int_0^\infty \lambda^{k-\ell} \bar{\psi}(\phi_{-\tilde{t}}(x)) \left| \frac{\partial \phi_{-\tilde{t}}(x)}{\partial x} \right| d\tilde{t} \\
&= \lambda^{k-\ell} \bar{\rho}(x).
\end{aligned}$$

□

LEMMA 20.14
Let $\bar{\rho} : \mathbb{R}^n \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ be homogeneous of degree k and ℓ, and $\|.\|_h$ be a homogeneous norm with respect to Δ_λ^r. Then the integral

$$\int_{\|x\|_h \geq 1} \bar{\rho}(x) \frac{f(x)}{\|x\|_h} dx$$

exists, if

$$k + \ell + \max_i r_i + \sum_{i=1}^{n} r_i < 1. \tag{20.4}$$

□

Proof For each $i = 1, ..., n$, we have

$$\begin{aligned}
&\int_{\|x\|_h \geq 1} \bar{\rho}(x) \frac{f_i(x)}{\|x\|_h} dx \\
&= \int_{1 \leq \|x\|_h \leq \lambda} \bar{\rho}(x) \frac{f_i(x)}{\|x\|_h} dx + \int_{\lambda \leq \|x\|_h \leq \lambda^2} \bar{\rho}(x) \frac{f_i(x)}{\|x\|_h} dx + ... \\
&= \int_{1 \leq \|x\|_h \leq \lambda} \bar{\rho}(x) \frac{f_i(x)}{\|x\|_h} dx + \int_{1 \leq \|\Delta^r_{1/\lambda} x\|_h \leq \lambda} \bar{\rho}(x) \frac{f_i(x)}{\|x\|_h} dx + ... \\
&= \int_{1 \leq \|x\|_h \leq \lambda} \bar{\rho}(x) \frac{f_i(x)}{\|x\|_h} dx + \int_{1 \leq \|\tilde{x}\|_h \leq \lambda} \bar{\rho}(\Delta^r_\lambda \tilde{x}) \frac{f_i(\Delta^r_\lambda \tilde{x})}{\|\Delta^r_\lambda \tilde{x}\|_h} d(\Delta^r_\lambda \tilde{x}) + ... \\
&= \int_{1 \leq \|x\|_h \leq \lambda} \bar{\rho}(x) \frac{f_i(x)}{\|x\|_h} dx + \int_{1 \leq \|\tilde{x}\|_h \leq \lambda} \lambda^{(k+\ell+r_i+\sum_i r_i - 1)} \bar{\rho}(\tilde{x}) \frac{f_i(\tilde{x})}{\|\tilde{x}\|_h} d\tilde{x} + ... \\
&= C + \lambda^{(k+\ell+r_i+\sum_i r_i - 1)} C + ...,
\end{aligned}$$

where $\lambda > 1$. The sum is finite when

$$(k + \ell + r_i + \sum_{i=1}^{n} r_i - 1) < 0,$$

and from this we conclude that $\int_{\|x\|_h \geq 1} \rho(x) \frac{f(x)}{\|x\|_h} dx$ exists if (20.4) is satisfied. □

For a homogeneous system whose equilibrium at the origin is asymptotically stable, a homogeneous density function can be constructed from a homogeneous Lyapunov function for the same system, whose existence is guaranteed by the following theorem.

THEOREM 20.15—[11]
Let the zero equilibrium of the system $\dot{x} = f(x)$ be asymptotically stable with $f \in C^0(\mathbb{R}^n, \mathbb{R}^n)$ homogeneous of degree ℓ w.r.t. Δ^r_λ, and let p be a positive integer. Then there exists a homogeneous Lyapunov function $\bar{V} \in C^p(\mathbb{R}^n, \mathbb{R}) \cap C^\infty(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$ with positive degree of homogeneity such that $\bar{V}(0) = 0$ and $\bar{V}(x) > 0$, $\nabla \bar{V}(x) \cdot f(x) < 0$ for all $x \neq 0$. □

The corresponding result for homogeneous density function can now be stated and proven as follows.

THEOREM 20.16
Let the zero equilibrium of the system $\dot{x} = f(x)$ be asymptotically stable, with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ homogeneous of degree ℓ w.r.t. Δ^r_λ. Then there exists a homogeneous function $\bar{\rho} \in C^\infty(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$ with negative degree of homogeneity which satisfies the following properties

(i) $\bar{\rho}(x) > 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\}$,
(ii) $\nabla \cdot (\bar{\rho} f) > 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\}$,
(iii) $\bar{\rho}(x) \frac{f(x)}{\|x\|_h}$ is integrable on $\{x : \|x\|_h \geq 1\}$. □

Proof It can be noted that for homogeneous systems asymptotic stability actually implies global asymptotic stability and hence global attractivity, as can be seen from Proposition 20.9. Let $\bar{V}(x)$ be a homogeneous Lyapunov function as in Theorem 20.15. We will use $\bar{\rho}(x) = \bar{V}(x)^{-\gamma}$ for some $\gamma > 0$ as the homogeneous density function for our system. It immediately follows that $\bar{\rho} \in C^\infty(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$ and property (i) is fulfilled. We will now show that by choosing large enough γ, properties (ii) and (iii) will also be satisfied.

Let k be the degree of $\bar{V}$. Choose γ such that

$$\gamma > \max \left\{ \frac{\ell + \max_i r_i + \sum_i r_i - 1}{k}, \max_{x \in S^{n-1}} \frac{\bar{V}(x)(\nabla \cdot f(x))}{-(\nabla \bar{V}(x) \cdot f(x))} \right\}. \tag{20.5}$$

Since $\bar{\rho}$ has degree $-k\gamma$ and $\gamma > \frac{\ell + \max_i r_i + \sum_i r_i - 1}{k}$, it follows that the condition in Lemma 20.14 is satisfied and therefore (iii) also holds. Next, notice that for $\bar{\rho} = \bar{V}^{-\gamma}$, we have

$$\nabla \cdot (\bar{\rho} f) = (\nabla \cdot f)\bar{V}^{-\gamma} - \gamma \bar{V}^{-(\gamma+1)}(\nabla \bar{V} \cdot f),$$

and furthermore $\nabla \cdot (\bar{\rho} f)$ will also be homogeneous (see Lemma 20.13). Hence we only need to check the positivity of $\nabla \cdot (\bar{\rho} f)$ on S^{n-1}. For $\gamma > \max_{x \in S^{n-1}} \frac{\bar{V}(x)(\nabla \cdot f(x))}{-(\nabla V(x) \cdot f(x))}$, the divergence $\nabla \cdot (\bar{\rho} f)$ will be positive on S^{n-1}, thus proving (ii). □

Unfortunately, for systems whose equilibrium at the origin is not asymptotically stable, the existence of a Lyapunov function is not guaranteed, and for systems with unstable equilibrium such a function cannot even exist. This requires us to proceed in a different way. If for such a system there exists a nonhomogeneous density function ρ with fast decay at large x and positive $\nabla \cdot (\rho f)$, then the homogenization method in the following lemma can be used to construct a homogeneous density function, as shown subsequently in Theorem 20.18.

LEMMA 20.17
Let $\hat{\rho} \in C^1(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$ be a nonnegative function such that $\hat{\rho}(x) = 0$ for $x \in \{x : \|x\| \leq L\}$, for some $L > 0$. In addition, assume that there exists a constant M such that

$$\lim_{\tau \to \infty} \tau^M \hat{\rho}(\Delta_\tau^r x) < \infty, \tag{20.6}$$

$$\lim_{\tau \to \infty} \tau^M \left| \frac{\partial \hat{\rho}}{\partial x_i}(\Delta_\tau^r x) \right| < \infty, \quad i = 1, ..., n, \tag{20.7}$$

for all $x \in S^{n-1}$. Let $k > -M + \max_i r_i$. Then the function $\bar{\rho} : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ given by

$$\bar{\rho}(x) = \int_0^\infty \tau^{-k-1} \hat{\rho}(\Delta_\tau^r x) d\tau$$

is
(i) well-defined and nonnegative,
(ii) homogeneous of degree k with respect to Δ_λ^r,
(iii) of class C^1. □

Proof (i) Since $\hat{\rho}(x) = 0$ when x is close to the origin, for every x we can find $\tilde{\tau}(x) > 0$ such that $\hat{\rho}(\Delta_\tau^r x) = 0$ for all $\tau \leq \tilde{\tau}(x)$. Thus

$$\bar{\rho}(x) = \int_{\tilde{\tau}(x)}^{\infty} \tau^{-k-1} \hat{\rho}(\Delta_\tau^r x) d\tau.$$

From property (20.6) it follows that the integral is well defined, since $k > -M$. It is obvious that $\bar{\rho}(x)$ is nonnegative, since the integrand is also nonnegative.

(ii) We have

$$\begin{aligned}
\bar{\rho}(\Delta_\lambda^r x) &= \int_0^{\infty} \tau^{-k-1} \hat{\rho}(\Delta_\lambda^r \Delta_\tau^r x) d\tau \\
&= \int_0^{\infty} \tau^{-k-1} \hat{\rho}(\Delta_{\lambda\tau}^r x) d\tau \\
&= \int_0^{\infty} \left(\frac{\sigma}{\lambda}\right)^{-k-1} \lambda^{-1} \hat{\rho}(\Delta_\sigma^r x) d\sigma \\
&= \lambda^k \bar{\rho}(x).
\end{aligned}$$

(iii) Since $\bar{\rho}$ is homogeneous, its partial derivatives (if exist) will also be homogeneous (cf. Proposition 20.7) and they will be entirely determined by their values on S^{n-1} (cf. Proposition 20.8). Therefore, to prove that $\bar{\rho}$ is C^1 on $\mathbb{R}^n \backslash \{0\}$, we only need to show that it is C^1 at any $x \in S^{n-1}$. Now, since $\hat{\rho}(x) = 0$ for small x, we can find $\tilde{\tau} > 0$ such that $\hat{\rho}(\Delta_\tau^r x) = 0$ for all $\tau \leq \tilde{\tau}$ and for all $x \in E \triangleq \{x : 1-\varepsilon \leq \|x\| \leq 1+\varepsilon\}$, where ε is a fixed positive number less than one. For all such x, we have

$$\bar{\rho}(x) = \int_{\tilde{\tau}}^{\infty} \tau^{-k-1} \hat{\rho}(\Delta_\tau^r x) d\tau. \tag{20.8}$$

Because of property (20.7), we can find constants $K, T > 0$, and $N > 1$ such that

$$\begin{aligned}
\left| \frac{\partial}{\partial x_i} (\tau^{-k-1} \hat{\rho}(\Delta_\tau^r x)) \right| &= \left| \tau^{-k-1+r_i} \frac{\partial \hat{\rho}}{\partial x_i} (\Delta_\tau^r x) \right| \\
&\leq K \tau^{-N}, \quad i = 1, ..., n,
\end{aligned}$$

for all $x \in E$ and $\tau \geq T$. Rewriting equation (20.8) as

$$\bar{\rho}(x) = \int_{\tilde{\tau}}^{T} \tau^{-k-1} \hat{\rho}(\Delta_\tau^r x) d\tau + \int_T^{\infty} \tau^{-k-1} \hat{\rho}(\Delta_\tau^r x) d\tau,$$

and using comparison test as well as Leibniz's rule, we conclude that $\bar{\rho}$ is C^1 on $\text{int}(E)$ and therefore also at any $x \in S^{n-1}$. By the homogeneity property, $\bar{\rho}$ is C^1 on $\mathbb{R}^n \setminus \{0\}$. □

THEOREM 20.18
Suppose that for a system $\dot{x} = f(x)$, $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, which is homogeneous of degree ℓ w.r.t. $\Delta_\tau^r x$, there exists a (not necessarily homogeneous) density function $\rho(x) \in C^1(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$ that is nonnegative and has a fast decay at large x, i.e., properties (20.6) and (20.7) in Lemma 20.17 hold for all positive M[1]. In addition, suppose that $\nabla \cdot (f\rho) > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$. Then there exists also a homogeneous density function $\bar{\rho}(x) \in C^1(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$ satisfying conditions

(i) $\bar{\rho}(x) \geq 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\}$,
(ii) $\nabla \cdot (\bar{\rho} f) > 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\}$,
(iii) $\bar{\rho}(x) \frac{f(x)}{\|x\|_h}$ is integrable on $\{x : \|x\|_h \geq 1\}$. □

Proof We will construct $\bar{\rho}$ from the nonhomogeneous function ρ. Let $g : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}^+$ be a C^1 function such that $g(x) = 1$ for all $x \in \{\Delta_\tau^r \tilde{x} : \tilde{x} \in S^{n-1}, \tau \geq b\}$ and $g(x) = 0$ for all $x \in \{\Delta_\tau^r \tilde{x} : \tilde{x} \in S^{n-1}, \tau \leq a\}$, $0 < a < b < 1$. The function $\hat{\rho}(x) \triangleq g(x)\rho(x)$ satisfies all the conditions in Lemma 20.17, and therefore we can use the lemma to construct a nonnegative $\bar{\rho}(x) \in C^1(\mathbb{R}^n \setminus \{0\}, \mathbb{R})$ that is homogeneous of degree k, automatically satisfying condition (i). Because of the fast decay, this condition will hold for arbitrary negative k. We will now show that by choosing negative enough k, conditions (ii) and (iii) will also be satisfied.

Since $\bar{\rho}$ is homogeneous, $\nabla \cdot (\bar{\rho} f)$ will also be homogeneous (Lemma 20.13), and therefore we only need to check the positivity condition on S^{n-1}. On this set, $\nabla \cdot (\bar{\rho} f)$ can be written as

$$\begin{aligned}\nabla \cdot (\bar{\rho} f)(x) &= \int_a^\infty \tau^{-k-1-\ell} \nabla \cdot (\hat{\rho} f)(\Delta_\tau^r x) d\tau \\ &= \int_a^b \tau^{-k-1-\ell} \nabla \cdot (\hat{\rho} f)(\Delta_\tau^r x) d\tau + \int_b^\infty \tau^{-k-1-\ell} \nabla \cdot (\hat{\rho} f)(\Delta_\tau^r x) d\tau\end{aligned}$$

The multiplication by g will cause the divergence $\nabla \cdot (\hat{\rho} f)(\Delta_\tau^r x)$ to be negative for some $\tau \in (a, b)$. However, notice also that $\nabla \cdot (\hat{\rho} f)(\Delta_\tau^r x) = \nabla \cdot (\rho f)(\Delta_\tau^r x) > 0$ for $\tau > b$. Define

$$\begin{aligned}C_1 &= \inf\{\nabla \cdot (\hat{\rho} f)(\Delta_\tau^r x) : x \in S^{n-1}, 1 \leq \tau \leq 2\} > 0, \\ C_2 &= \inf\{\nabla \cdot (\hat{\rho} f)(\Delta_\tau^r x) : x \in S^{n-1}, a \leq \tau \leq b\} < 0,\end{aligned}$$

and assume that $-k - 1 - \ell > 0$. Then it follows that

$$\begin{aligned}\int_a^b \tau^{-k-1-\ell} \nabla \cdot (\hat{\rho} f)(\Delta_\tau^r x) d\tau &\geq C_2 (b - a) b^{-k-1-\ell} \\ \int_b^\infty \tau^{-k-1-\ell} \nabla \cdot (\hat{\rho} f)(\Delta_\tau^r x) d\tau &\geq C_1,\end{aligned}$$

from which it can be seen that by making k negative enough, equivalently $-k-1-\ell$ positive enough, we will obtain $\nabla \cdot (\hat{\rho} f)(x) > 0$ on S^{n-1}.

Finally, to guarantee that condition (iii) is satisfied, choose k in accordance with Lemma 20.14. □

[1]This condition is actually too strict. It is enough to require that (20.6) and (20.7) hold up to some large enough M (cf. the proof of the theorem). Unfortunately, a bound for this is not available a priori.

20.4 Perturbations by Higher and Lower Order Terms

Unlike the Lyapunov stability, where asymptotic stability in a homogeneous system is preserved under higher order term perturbations [11], in the case of density function a robustness property is not immediately obtained. In other words, a homogeneous Lyapunov function for the homogeneous system will still serve as a Lyapunov function for the perturbed system, whereas the corresponding homogeneous density function will in general no longer be a density function for the perturbed system. This is obvious, since the existence of a density function proves a *global* property, which is lost under such a perturbation. However, with some appropriate assumptions, local attractivity of the origin is still preserved.

PROPOSITION 20.19
Let $\dot{x} = f(x)$ be a homogeneous dynamical system of degree ℓ with $\bar{\rho}$ a homogeneous density function for it, and let $\hat{\delta}f$ be a higher order perturbing vector field. Denote by $\mathcal{D}$ a neighborhood of the origin where $\nabla \cdot (\bar{\rho}(f + \hat{\delta}f))$ is almost everywhere positive. If $I \subset \mathcal{D}$ is an invariant set strictly containing the origin, then almost all trajectories in I will converge to the origin. □

Proof Note that if for the original vector field we have $\nabla \cdot (\bar{\rho}f) > 0$ on $\mathbb{R}^n \setminus \{0\}$, then $\mathcal{D}$ is guaranteed to exist (cf. proof of Theorem 3 in [11]). The proof of the proposition proceeds in the same way as the proof of Theorem 20.1 (see [7, page 166]), using (with the same notations as in [7]) $X = I$, $P = \{x \in I : \|x\| > r\}$, and $D = \{x \in I : \|x\| > \varepsilon\}$. □

REMARK 20.20
For a system whose zero equilibrium is asymptotically stable, a homogeneous density function $\bar{\rho}$ with $\nabla \cdot (\bar{\rho}f) > 0$ on $\mathbb{R}^n \setminus \{0\}$ can always be constructed (cf. Theorem 20.16). This guarantees the existence of $\mathcal{D}$ in Proposition 20.19. In addition, an invariant set I can always be found. In this context, the proposition actually corresponds to the fact that an asymptotically stable zero equilibrium of a homogeneous system is (locally) asymptotically stable when the system is perturbed by higher order terms. □

On the other hand, when perturbed by lower order terms, in general even local attractivity will be lost. Yet a property related to trajectories far from the origin can be shown to hold as follows.

PROPOSITION 20.21
Let $\dot{x} = f(x)$ be a homogeneous dynamical system of degree ℓ with $\bar{\rho}$ a homogeneous density function for it, and let $\check{\delta}f$ be a lower order perturbing vector field. Assume that $\nabla \cdot (\bar{\rho}(f + \check{\delta}f))$ is positive almost everywhere on $\mathcal{E} = \{x : \|x\| \geq r\}$, where r is some positive constant. Then the set of trajectories with $\lim_{t\to\infty} \phi_t(x) = \infty$ is a set of zero measure. □

Proof Analogous to Proposition 20.19, note that if for the original vector field we have $\nabla \cdot (\bar{\rho}f) > 0$ on $\mathbb{R}^n \setminus \{0\}$, then we can always find a positive r such that $\nabla \cdot (\bar{\rho}(f + \check{\delta}f))$ is positive almost everywhere on $\mathcal{E} = \{x : \|x\| \geq r\}$. Also notice

that if $\frac{\bar{\rho}f}{\|x\|}$ is integrable on $\mathcal{E}$, then so is $\frac{\bar{\rho}(f+\check{\delta}f)}{\|x\|}$. Without loss of generality we may assume that $\bar{\rho}$ is integrable on $\mathcal{E}$ and $\frac{\|(f+\check{\delta}f)(x)\|}{\|x\|}$ is bounded, since otherwise we can introduce a new density function and an equivalent system (modulo a shift of the time axis) that have the required property [7, page 162]. Now define $Z = \{x_0 \in \mathcal{E} : \|\phi_t(x_0)\| > r \text{ for } t \geq 0\}$. The set of trajectories with $\lim_{t\to\infty} \phi_t(x) = \infty$ is contained in Z. From Lemma 20.2 (using $D = \{x : \|x\| > r\}$) and the fact that $\bar{\rho}$ is nonnegative, it follows that

$$0 \geq \int_{\phi_t(Z)} \bar{\rho}(x)dx - \int_Z \bar{\rho}(x)dx = \int_0^t \int_{\phi_\tau(Z)} [\nabla \cdot (\bar{\rho}(f + \hat{\delta}f))](x)dxd\tau.$$

However, $\nabla \cdot (\bar{\rho}(f + \hat{\delta}f))$ is positive for almost all $x \in \mathcal{E}$, and thus Z must have zero measure. This implies that the set of trajectories with $\lim_{t\to\infty} \phi_t(x) = \infty$ is also a set of zero measure. □

20.5 Conclusions

We have proven the existence of a homogeneous density function for a homogeneous system whose equilibrium at the origin is asymptotically stable. When the equilibrium is not asymptotically stable, we show that a homogeneous density function exists if there exists also a nonhomogeneous density function with fast decay and satisfying the positive divergence criterion.

The effects of perturbations on the system have been addressed subsequently. In particular, we show that under some appropriate assumptions local attractivity of the zero equilibrium is preserved when the system is perturbed by higher order terms, and that almost all trajectories cannot escape to infinity, in the case of perturbation by lower order terms.

Acknowledgments

The collaboration between the authors was supported by an exchange grant from the Swedish Foundation for International Cooperation in Research and Higher Education.

20.6 References

[1] L. Grüne. Homogeneous state feedback stabilization of homogenous systems. *SIAM J. Control Optim.*, 38(4):1288–1308, 2000.

[2] W. Hahn. *Stability of Motion*. Springer-Verlag New York, Inc., New York, 1967.

[3] H. Hermes. Nilpotent and high-order approximations of vector field systems. *SIAM Rev.*, 33(2):238–264, 1991.

[4] M. Kawski. Stabilization and nilpotent approximations. In *Proceedings of IEEE Conference on Decision and Control (CDC)*, pages 1244–1248, 1988.

[5] M. Kawski. Geometric homogeneity and applications to stabilization. In *Proceedings IFAC Symposium on Nonlinear Control Systems Design (NOLCOS)*, pages 147–152, 1995.

[6] R. T. M'Closkey and R. M. Murray. Exponential stabilization of driftless nonlinear control systems using homogeneous feedback. *IEEE Trans. Automat. Control*, 42(5):614–628, 1997.

[7] A. Rantzer. A dual to Lyapunov's stability theorem. *Systems and Control Letters*, 42(3):161–168, 2001.

[8] A. Rantzer. A converse theorem for density functions. Accepted for publication at the IEEE Conference on Decision and Control (CDC), 2002.

[9] A. Rantzer and F. Ceragioli. Smooth blending of nonlinear controllers using density functions. In *Proceedings of European Control Conference*, 2001.

[10] A. Rantzer and P. A. Parrilo. On convexity in stabilization of nonlinear systems. In *Proceedings of IEEE Conference on Decision and Control (CDC)*, 2000.

[11] L. Rosier. Homogeneous Lyapunov function for homogeneous continuous vector field. *Systems and Control Letters*, 19:467–473, 1992.

[12] R. Sepulchre and D. Aeyels. Homogeneous Lyapunov functions and necessary conditions for stabilization. *Math. Control Signals Systems*, 9:34–58, 1996.

21

Stabilization by Collocated Feedback

Olof J. Staffans

Abstract

Recently Guo and Luo (and independently Weiss and Tucsnak) were able to prove that a certain damped second order system can be interpreted as a continuous time (well-posed and stable) scattering conservative system. We show that this is a special case of the following more general result: if we apply the so called diagonal transform (which is a particular rescaled feedback/feedforward transform) to an arbitrary continuous time impedance conservative system, then we always get a scattering conservative system. In the particular case mentioned above the corresponding impedance conservative system is as a undamped second order system with collocated actuators and sensors.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 275-292, 2003.
© Springer-Verlag Berlin Heidelberg 2003

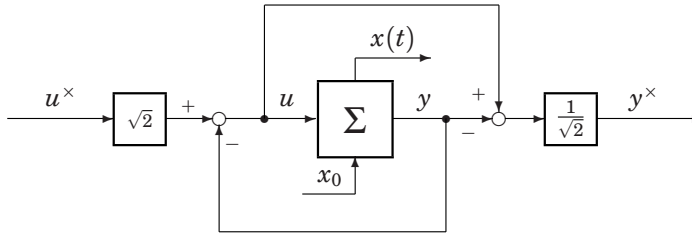


Figure 21.1 The diagonal transform

21.1 Introduction

In two recent articles Guo and Luo [1] and Weiss and Tucsnak [15] study the abstract second order system of differential equations

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2} z(t) + A_0 z(t) &= -\frac{1}{2} C_0^* \frac{\mathrm{d}}{\mathrm{d}t} C_0 z(t) + C_0^* u(t), \\ y(t) &= -\frac{\mathrm{d}}{\mathrm{d}t} C_0 z(t) + u(t), \end{aligned} \tag{21.1}$$

with input u, state $\begin{bmatrix} \sqrt{A_0}\, z \\ \dot{z} \end{bmatrix}$, and output y. Here A_0 is an arbitrary positive (unbounded) self-adjoint operator on a Hilbert space Z with a bounded inverse. We define the fractional powers of A_0 in the usual way, and denote $Z_{1/2} = \mathcal{D}(\sqrt{A_0})$ and $Z_{-1/2} = (Z_{1/2})^*$ (where we identify Z with its dual). Thus, $Z_{1/2} \subset Z \subset Z_{-1/2}$, with continuous and dense injections, and A^{-1} maps $Z_{-1/2}$ onto $Z_{1/2}$. The operator C is an arbitrary bounded linear operator from $Z_{1/2}$ to another Hilbert space U. Guo and Luo showed in [1] and Weiss and Tucsnak showed in [15] (independently of each other) that the above system may be interpreted as a continuous time (well-posed and energy stable) *scattering conservative* system with input u, state $x = \begin{bmatrix} \sqrt{A_0}\, z \\ \dot{z} \end{bmatrix}$, and output y. The input and output spaces are both U, and the state space is $X = \begin{bmatrix} Z \\ Z \end{bmatrix}$ $(= Z \times Z)$.

Formally, the system (21.1) is equivalent to the *diagonally transformed system*

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2} z(t) + A_0 z(t) &= \frac{1}{\sqrt{2}} C_0^* u^\times(t), \\ y^\times(t) &= \frac{1}{\sqrt{2}} \frac{\mathrm{d}}{\mathrm{d}t} C_0 z(t), \end{aligned} \tag{21.2}$$

which we get from (21.1) by replacing u and y in (21.1) by $u^\times = \frac{1}{\sqrt{2}}(u + y)$ respectively $y^\times = \frac{1}{\sqrt{2}}(u - y)$. We can formally get back to (21.1) by repeating the same transform: we replace $u^\times$ and $y^\times$ in (21.2) by $u = \frac{1}{\sqrt{2}}(u^\times + y^\times)$ respectively $y = \frac{1}{\sqrt{2}}(u^\times - y^\times)$. This transform, drawn in Figure 21.1, is simply a rescaled feedback/feedforward connection.

The purpose of this article is to show that the above transformations are not just *formal*, but that that they can be mathematically justified, thereby giving

a positive answer to the question posed in [1, Remark 2]. It follows directly from [8, Theorem 4.7] that (21.2) is an impedance conservative system of the type introduced in [8]. According to [9, Theorem 8.2], by applying the diagonal transform to this system we get a scattering passive system. As we shall show below, this scattering passive system is exactly the system described by (21.1).

21.2 Infinite-Dimensional Linear Systems

Many infinite-dimensional linear time-invariant continuous-time systems can be described by the equations

$$\begin{aligned} x'(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \quad t \geq 0, \\ x(0) &= x_0, \end{aligned} \tag{21.3}$$

on a triple of Hilbert spaces, namely, the input space U, the state space X, and the output space Y. We have $u(t) \in U$, $x(t) \in X$ and $y(t) \in Y$. The operator A is supposed to be the generator of a strongly continuous semigroup. The operators A, B and C are usually unbounded, but D is bounded.

By modifying this set of equations slightly we get the class of systems which will be used in this article. In the sequel, we think about the block matrix $S = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ as *one single (unbounded) operator* from $\left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right]$ to $\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$, and write (21.3) in the form

$$\begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} = S \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}, \quad t \geq 0, \quad x(0) = x_0. \tag{21.4}$$

The operator S completely determines the system. Thus, we may identify the system with such an operator, which we call the *node* of the system.

The system nodes that we use have a certain structure which makes it resemble a block matrix operator of the type $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$. To describe this structure we need the notion of *rigged Hilbert spaces*. Let A be the generator of a C_0 semigroup on the Hilbert space X. We denote its domain $\mathcal{D}(A)$ by X_1. We identify the dual of X with X itself, and denote $X_{-1} = \mathcal{D}(A^*)^*$. Then $X_1 \subset X \subset X_{-1}$ with continuous and dense injections. The operator A has a unique extension to an operator in $\mathcal{L}(X; X_{-1})$ which we denote by $A_{|X}$ (thereby indicating that the domain of this operator is all of X). This operator is the generator a C_0 semigroup on X_{-1}, whose restriction to X is the semigroup generated by A.

DEFINITION 21.1
We call S a *system node* on the three Hilbert spaces (U, X, Y) if it satisfies condition (S) below:[1]

(S) $S := \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right] : \left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right] \supset \mathcal{D}(S) \to \left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$ is a closed linear operator. Here $A\&B$ is the restriction to $\mathcal{D}(S)$ of $[A_{|X} \quad B]$, where A is the *generator of a C_0 semigroup* on X (the notations $A_{|X} \in \mathcal{L}(X; X_{-1})$ and X_{-1} were introduced in the text above). The operator B is an arbitrary operator in $\mathcal{L}(U; X_{-1})$,

[1] This definition is equivalent to the corresponding definitions used by Smuljan in [6] and by Salamon in [4, 5].

and $C\&D$ is an arbitrary linear operator from $\mathcal{D}(S)$ to Y. In addition, we require that

$$\mathcal{D}(S) = \left\{ \left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] \in \left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right] \;\middle|\; A_{|X}x + Bu \in X \right\}.$$

□

We shall use the following names of the different parts of the system node $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$. The operator A is the *main operator* or the *semigroup generator*, B is the *control operator*, $C\&D$ is the *combined observation/feedthrough operator*, and the operator C defined by

$$Cx := C\&D \begin{bmatrix} x \\ 0 \end{bmatrix}, \qquad x \in X_1,$$

is the *observation operator* of S.

An easy algebraic computation (see, e.g., [10, Section 4.7] for details) shows that for each $\alpha \in \rho(A) = \rho(A_{|X})$, the operator $\left[\begin{smallmatrix} 1 & (\alpha - A_{|X})^{-1}B \\ 0 & 1 \end{smallmatrix}\right]$ is an boundedly invertible mapping between $\left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right] \to \left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} X_1 \\ U \end{smallmatrix}\right] \to \mathcal{D}(S)$. Since $\left[\begin{smallmatrix} X_1 \\ U \end{smallmatrix}\right]$ is dense in $\left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right]$, this implies that $\mathcal{D}(S)$ is dense in $\left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right]$. Furthermore, since the second column $\left[\begin{smallmatrix} (\alpha - A_{|X})^{-1}B \\ 1 \end{smallmatrix}\right]$ of this operator maps U into $\mathcal{D}(S)$, we can define the *transfer function* of S by

$$\widehat{\mathfrak{D}}(s) := C\&D \begin{bmatrix} (s - A_{|X})^{-1}B \\ 1 \end{bmatrix}, \qquad s \in \rho(A), \tag{21.5}$$

which is a $\mathcal{L}(U;Y)$-valued analytic function on $\rho(A)$. By the resolvent formula, for any two $\alpha, \beta \in \rho(A)$,

$$\begin{aligned} \widehat{\mathfrak{D}}(\alpha) - \widehat{\mathfrak{D}}(\beta) &= C\left[(\alpha - A_{|X})^{-1} - (\beta - A_{|X})^{-1}\right]B \\ &= (\beta - \alpha)C(\alpha - A)^{-1}(\beta - A_{|X})^{-1}B. \end{aligned} \tag{21.6}$$

Let us finally present the class of *compatible* system nodes, originally introduced by Helton [2]). This class can be defined in several different ways, one of which is the following. We introduce an auxiliary Banach space W satisfying $X_1 \subset W \subset X$, fix some $\alpha \in \rho(A)$, and define $W_{-1} = (\alpha - A_{|X})W$ with $|x|_{W_{-1}} = |(\alpha - A_{|X})^{-1}x|_W$ (defined in this way the norm in W_{-1} depends on α, but the space itself does not). Thus

$$X_1 \subset W \subset X \subset W_{-1} \subset X_{-1}.$$

The embeddings $W \subset X$ and $W_{-1} \subset X_{-1}$ are always dense, but the embeddings $X_1 \subset W$ and $X \subset W_{-1}$ need not be dense. The system is *compatible* if $\mathcal{R}(B) \subset W_{-1}$ and C has an extension to an operator $C_{|W} \in \mathcal{L}(W;Y)$ (this extension is not unique unless the embedding $X_1 \subset W$ is dense). Thus, in this case the operator $C_{|W}(\alpha - A_{|X})^{-1}B \in \mathcal{L}(U;Y)$ for all $\alpha \in \rho(A)$. If we fix some $\alpha \in \rho(A)$ and define

$$D := \widehat{\mathfrak{D}}(\alpha) - C_{|W}(\alpha - A_{|X})^{-1}B,$$

then $D \in \mathcal{L}(U;Y)$, and it follows from (21.6) that D *does not depend on* α, only on A, B, $C_{|W}$, and $\widehat{\mathfrak{D}}$ (in particular, different extensions of C to W give different operators D). Clearly, the above formula means that $\widehat{\mathfrak{D}}$ can be written in the simple form

$$\widehat{\mathfrak{D}}(s) = C_{|W}(s - A_{|X})^{-1}B + D, \quad s \in \rho(A). \tag{21.7}$$

Another way of describing compatibility is to say that the system node S can be extended to a bounded linear operator $\left[\begin{smallmatrix} A_{|W} & B \\ C_{|W} & D \end{smallmatrix}\right] \in \mathcal{L}(\left[\begin{smallmatrix} W \\ U \end{smallmatrix}\right]; \left[\begin{smallmatrix} W_{-1} \\ U \end{smallmatrix}\right])$, where $A_{|W}$ is the restriction of $A_{|X}$ to W. Thus

$$\begin{bmatrix} A\&B \\ C\&D \end{bmatrix} = \begin{bmatrix} A_{|W} & B \\ C_{|W} & D \end{bmatrix}_{|\mathcal{D}(S)}.$$

We shall refer to the operator $\left[\begin{smallmatrix} A_{|W} & B \\ C_{|W} & D \end{smallmatrix}\right]$ as a (possibly non-unique) *compatible representation of* S *over the space* W. There is always a minimal space W which can be used in this representation, namely $W := (\alpha - A)^{-1}W_{-1}$ where $\alpha \in \rho(A)$ and $W_{-1} := (X + BU)$, but it is frequently more convenient to work in some other space W (for example, it may be possible to choose a larger space W for which the embedding $X_1 \subset W$ is dense and the extension is unique).

As shown in [11], the system node S *of a well-posed system* is always compatible, but the converse is not true (an example of a compatible system of the type (21.2) which is not well-posed is given in [13]).

Every system node induces a 'dynamical system' of a certain type:

LEMMA 21.2
Let S be a system node on (U, X, Y). Then, for each $x_0 \in X$ and $u \in W^{2,1}_{loc}(\mathbb{R}^+;U)$ with $\left[\begin{smallmatrix} x_0 \\ u(0) \end{smallmatrix}\right] \in \mathcal{D}(S)$, the equation

$$\begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} = S \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}, \quad t \geq 0, \quad x(0) = x_0, \tag{21.8}$$

has a unique solution (x, y) satisfying $\left[\begin{smallmatrix} x(t) \\ u(t) \end{smallmatrix}\right] \in \mathcal{D}(S)$ for all $t \geq 0$, $x \in C^1(\mathbb{R}^+;X)$, and $y \in C(\mathbb{R}^+;Y)$. □

This lemma is proved in [3] (and also in [10]).[2]

So far we *defined* Σ^t_0 *only for the class of smooth data* given in Lemma 21.2. It is possible to allow arbitrary initial states $x_0 \in X$ and input functions $u \in L^1_{loc}(\mathbb{R}^+;U)$ in Lemma 21.2 by allowing the state to take values in the larger space X_{-1} instead of in X, and by allowing y to be a distribution. Rather than presenting this result in its full generality, let us observe the following fact.

[2]Well-posed versions of this lemma (see Definition 21.4) are (implicitly) found in [4] and [6] (and also in [11]). In the well-posed case we need less smoothness of u: it suffices to take $u \in W^{1,2}_{loc}(\mathbb{R}^+;U)$. In addition y will be smoother: $y \in W^{1,2}_{loc}(\mathbb{R}^+;Y)$.

LEMMA 21.3
Let $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ be a system node on (U, X, Y). Let $x_0 \in X$, and $u \in L^1_{loc}(\mathbb{R}^+; U)$, and let x and y be the state trajectory and output of S with initial state x_0, and input function u. If $x \in W^{1,1}_{loc}(\mathbb{R}^+; X)$, then $\left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] \in L^1_{loc}(\mathbb{R}^+; \mathcal{D}(S))$, $y \in L^1_{loc}(\mathbb{R}^+; Y)$, and $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ is the unique solution with the above properties of the equation

$$\begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} = S \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \text{ for almost all } t \geq 0, \quad x(0) = x_0. \tag{21.9}$$

If $u \in C(\mathbb{R}^+; U)$ and $x \in C^1(\mathbb{R}^+; X)$, then $\left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] \in C(\mathbb{R}^+; \mathcal{D}(S))$, $y \in C(\mathbb{R}^+; Y)$, and the equation (21.9) holds for all $t \geq 0$. □

See [10, Section 4.7] for the proof.

Many system nodes are *well-posed*:

DEFINITION 21.4
A system node S is *well-posed* if, for some $t > 0$, there is a finite constant $K(t)$ such that the solution (x, y) in Lemma 21.2 satisfies

$$|x(t)|^2 + \|y\|^2_{L^2(0,t)} \leq K(t)\big(|x_0|^2 + \|u\|^2_{L^2(0,t)}\big). \tag{WP}$$

It is *energy stable* if there is some $K < \infty$ so that, for all $t \in \mathbb{R}^+$, the solution (x, y) in Lemma 21.2 satisfies

$$|x(t)|^2 + \|y\|^2_{L^2(0,t)} \leq K\big(|x_0|^2 + \|u\|^2_{L^2(0,t)}\big). \tag{ES}$$

□

For more details, explanations and examples we refer the reader to [3] and [7, 8, 9, 10] (and the references therein).

21.3 Passive and Conservative Scattering and Impedance Systems

The following definitions are slightly modified versions of the definitions in the two classical papers [16, 17] by Willems (although we use a slightly different terminology: our *passive* is the *same as Willems' dissipative*, and we use Willems' *storage function* as the norm in the state space).

DEFINITION 21.1
A system node S is *scattering passive* if, for all $t > 0$, the solution (x, y) in Lemma 21.2 satisfies

$$|x(t)|^2 - |x_0|^2 \leq \|u\|^2_{L^2(0,t)} - \|y\|^2_{L^2(0,t)}. \tag{SP}$$

It is *scattering energy preserving* if the above inequality holds in the form of an equality: for all $t > 0$, the solution (x, y) in Lemma 21.2 satisfies

$$|x(t)|^2 - |x_0|^2 = \|u\|^2_{L^2(0,t)} - \|y\|^2_{L^2(0,t)}. \tag{SE}$$

Finally, it is *scattering conservative* if both S and S^* are scattering energy preserving.[3] □

Thus, *every scattering passive system is well-posed and energy stable*: the passivity inequality **(SP)** implies the energy stability inequality **(ES)**.

DEFINITION 21.2
A system node S on (U, X, U) (note that $Y = U$) is *impedance passive* if, for all $t > 0$, the solution (x, y) in Lemma 21.2 satisfies

$$|x(t)|_X^2 - |x_0|_X^2 \leq 2 \int_0^t \Re\langle y(t), u(t)\rangle_U dt. \qquad \textbf{(IP)}$$

It is *impedance energy preserving* if the above inequality holds in the form of an equality: for all $t > 0$, the solution (x, y) in Lemma 21.2 satisfies

$$|x(t)|_X^2 - |x_0|_X^2 = 2 \int_0^t \Re\langle y(t), u(t)\rangle_U dt. \qquad \textbf{(IE)}$$

Finally, S is *impedance conservative* if both S and the dual system node S^* are impedance energy preserving. □

Note that in this case *well-posedness is neither guaranteed, nor relevant*.

Physically, *passivity* means that *there are no internal energy sources*. An energy preserving system has neither any internal energy sources nor any sinks. Other types of passivity have also been considered in the literature; in particular *transmission (or chain scattering)* passive or conservative systems.

Both in the scattering and in the impedance setting, the property of being passive is conserved under the passage from a system node S to its dual. See [8] for details.

The following theorem can be used to test if a system node is impedance passive or energy preserving or conservative:

THEOREM 21.3—[8, THEOREMS 4.2, 4.6, AND 4.7]
Let $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ be a system node on (U, X, U).

(i) S is impedance passive if and only if the system node $\left[\begin{smallmatrix} A\&B \\ -C\&D \end{smallmatrix}\right]$ is dissipative, i.e, for all $\left[\begin{smallmatrix} x_0 \\ u_0 \end{smallmatrix}\right] \in \mathcal{D}(S)$,

$$\Re\left\langle \begin{bmatrix} x_0 \\ u_0 \end{bmatrix}, \begin{bmatrix} A\&B \\ -C\&D \end{bmatrix} \begin{bmatrix} x_0 \\ u_0 \end{bmatrix} \right\rangle_{\left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right]} \leq 0. \qquad (21.10)$$

(ii) S is impedance energy preserving if and only if the system node $\left[\begin{smallmatrix} A\&B \\ -C\&D \end{smallmatrix}\right]$ is skew-symmetric, i.e., $\mathcal{D}(S) = \mathcal{D}(\left[\begin{smallmatrix} A\&B \\ -C\&D \end{smallmatrix}\right]) \subset \mathcal{D}(\left[\begin{smallmatrix} A\&B \\ -C\&D \end{smallmatrix}\right]^*)$, and

$$\begin{bmatrix} A\&B \\ -C\&D \end{bmatrix}^* \begin{bmatrix} x_0 \\ u_0 \end{bmatrix} = -\begin{bmatrix} A\&B \\ -C\&D \end{bmatrix} \begin{bmatrix} x_0 \\ u_0 \end{bmatrix}, \qquad \begin{bmatrix} x_0 \\ u_0 \end{bmatrix} \in \mathcal{D}(S). \qquad (21.11)$$

[3] If S is a system node on (U, X, Y), then its adjoint S^* is a system node on (Y, X, U). See, e.g., [3].

(iii) S is impedance conservative if and only if the system node $\left[\begin{smallmatrix} A\&B \\ -C\&D \end{smallmatrix}\right]$ is skew-adjoint, i.e.,

$$\begin{bmatrix} A\&B \\ -C\&D \end{bmatrix}^* = -\begin{bmatrix} A\&B \\ -C\&D \end{bmatrix}. \tag{21.12}$$

Equivalently, S is impedance conservative if and only $A^* = -A$, $B^* = C$, and $\widehat{\mathfrak{D}}(\alpha) + \widehat{\mathfrak{D}}(-\overline{\alpha})^* = 0$ for some (or equivalently, for all) $\alpha \in \rho(A)$ (in particular, this identity is true for all α with $\Re\alpha \neq 0$).

□

Many impedance passive systems are well-posed. There is a simple way of characterizing such systems:

THEOREM 21.4
An impedance passive system node is well-posed if and only if its transfer function $\widehat{\mathfrak{D}}$ is bounded on some (or equivalently, on every) vertical line in $\mathbb{C}^+$. When this is the case, the growth bound of the system is zero, and, in particular, $\widehat{\mathfrak{D}}$ is bounded on every right half-plane $\mathbb{C}_\varepsilon^+ = \{s \in \mathbb{C} \mid \Re s > \varepsilon\}$ with $\varepsilon > 0$. □

This is [8, Theorem 5.1]. It can be used to show that many systems with *collocated actuators and sensors* are well-posed.

EXAMPLE 21.5
To get the system described by (21.2) we take the state to be $x = \left[\begin{smallmatrix} \sqrt{A_0}\, z \\ \dot{z} \end{smallmatrix}\right]$, the input to be u, and the output to be y. The input and output spaces are U, the state space is $\left[\begin{smallmatrix} Z \\ Z \end{smallmatrix}\right]$, and, in compatibility notion with $W = Z_{1/2}$ and $W_{-1/2} = Z_{-1/2}$, the extended system node is given by

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{cc|c} 0 & \sqrt{A_0} & 0 \\ -\sqrt{A_0} & 0 & \frac{1}{\sqrt{2}} C_0^* \\ \hline 0 & \frac{1}{\sqrt{2}} C_0 & 0 \end{array}\right]$$

(the first element in the middle row stands for an extended version of $\sqrt{A_0}$). The domain of the system node itself consists of those $\left[\begin{smallmatrix} x_1 \\ x_2 \\ u \end{smallmatrix}\right] \in \left[\begin{smallmatrix} Z \\ Z \\ U \end{smallmatrix}\right]$ which satisfy $x_1 - A_0^{-1/2} C_0^* u \in Z_{1/2}$ and $x_2 \in Z_{1/2}$, and its transfer function is

$$\widehat{\mathfrak{D}}(s) = C_0\Big(s + \frac{1}{s} A_0\Big)^{-1} C_0^* \qquad \Re s \neq 0$$

(where the inverse maps $Z_{-1/2}$ onto $Z_{1/2}$). By Theorem 21.3, this system node is impedance conservative. □

EXAMPLE 21.6
Also the system described by (21.1) can be formulated as a system node with the same input, state, and output as in Example 21.5. This time we take the extended

system node to be (in the notation below we have anticipated the fact, which will be proved later, that this example is the diagonal transform of Example 21.5[4])

$$\left[\begin{array}{c|c} A^\times & B^\times \\ \hline C^\times & D^\times \end{array}\right] = \left[\begin{array}{cc|c} 0 & \sqrt{A_0} & 0 \\ -\sqrt{A_0} & \frac{1}{2}C_0^*C_0 & C_0^* \\ \hline 0 & -C_0 & 1 \end{array}\right]$$

(again the first element in the middle row stands for an extended version of $\sqrt{A_0}$). The domain of the system node itself consists of those $\left[\begin{smallmatrix} x_1 \\ x_2 \\ u \end{smallmatrix}\right] \in \left[\begin{smallmatrix} Z \\ Z \\ U \end{smallmatrix}\right]$ which satisfy $x_1 - A_0^{-1/2}\left(\frac{1}{2}C_0^*C_0x_2 + C_0^*u\right) \in Z_{1/2}$ and $x_2 \in Z_{1/2}$, and its transfer function is

$$\widehat{\mathfrak{D}}(s) = 1 - C_0\Big(s + \frac{1}{2}C_0^*C_0 + \frac{1}{s}A_0\Big)^{-1}C_0^* \qquad \Re s \neq 0.$$

□

It is not obvious that Example 21.6 is scattering conservative (hence well-posed and energy stable). That this is, indeed, the case is the main result of [15]. Here we shall rederive that result by a completely different method, appealing to the following general result.

THEOREM 21.7—[9, THEOREM 8.2]
A system node $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ on (U, X, U) is impedance passive (or energy preserving or conservative) if and only if it is diagonally transformable,[5] and the diagonally transformed system node $S^\times = \left[\begin{smallmatrix} [A\&B]^\times \\ [C\&D]^\times \end{smallmatrix}\right]$ is scattering passive (or energy preserving, or conservative) (in particular, it is well-posed and energy stable). The system node $S^\times$ can be determined from its main operator $A^\times$, control operator $B^\times$, observation operator $C^\times$, and transfer function $\widehat{\mathfrak{D}}^\times$, which can be computed from the following formulas, valid for all $\alpha \in \rho(A) \cap \rho(A^\times)$,[6]

$$\begin{aligned} &\left[\begin{array}{cc} (\alpha - A^\times)^{-1} & \frac{1}{\sqrt{2}}(\alpha - A^\times_{|X})^{-1}B^\times \\ \frac{1}{\sqrt{2}}C^\times(\alpha - A^\times)^{-1} & \frac{1}{2}(1 + \widehat{\mathfrak{D}}^\times(\alpha)) \end{array}\right] \\ &\quad = \left(\left[\begin{array}{cc} \alpha & 0 \\ 0 & 1 \end{array}\right] - \left[\begin{array}{c} A\&B \\ -C\&D \end{array}\right]\right)^{-1} \\ &\quad = \left[\begin{array}{cc} (\alpha - A)^{-1} & 0 \\ 0 & 0 \end{array}\right] \\ &\qquad + \left[\begin{array}{c} (\alpha - A_{|X})^{-1}B \\ 1 \end{array}\right](1 + \widehat{\mathfrak{D}}(\alpha))^{-1}\left[\begin{array}{cc} -C(\alpha - A)^{-1} & 1 \end{array}\right] \end{aligned} \tag{21.13}$$

In particular, $1 + \widehat{\mathfrak{D}}(\alpha)$ is invertible and $\widehat{\mathfrak{D}}^\times(\alpha) = (1 - \widehat{\mathfrak{D}}(\alpha))(1 + \widehat{\mathfrak{D}}(\alpha))^{-1}$ for all $\alpha \in \rho(A) \cap \rho(A^\times)$. □

[4] We denote the identity operator (on any Hilbert space) by 1.

[5] This notion will be defined in Section 21.5.

[6] $A^\times_{|X}$ is the extension of $A^\times$ to an operator in $\mathcal{L}(X; X^\times_{-1})$, where $X^\times_{-1}$ is the analogue of X_{-1} with A replaced by $A^\times$.

Thus, in order to show that Example 21.6 is scattering conservative, it suffices to show that it is the diagonal transform of Example 21.5. This can be achieved via a lengthy computation based on formula (21.13), but instead of doing this we shall derive an alternative formula to (21.13) which is valid (only) for *compatible* systems. See Corollary 21.2 and Remark 21.4.

21.4 Flow-Inversion

In order to get a compatible version of (21.13) we need to develop a version of the diagonal transform which is more direct than the one presented in [9] (there this transformation was defined as a Cayley transform, followed by a discrete time diagonal transform, followed by an inverse Cayley transform). Instead of using this lengthy chain of transformations we here want to use a (non-well-posed) system node version of the approach used in [8, Section 5]. That approach used the theory of *flow-inversion* of a well-posed system developed in [12], so we have to start by first extending the notion of flow-inversion to a general system node.[7]

DEFINITION 21.1
Let $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ be a system node on (U, X, Y). We call S *flow-invertible* if there exists another system node $S^\times = \left[\begin{smallmatrix} [A\&B]^\times \\ [C\&D]^\times \end{smallmatrix}\right]$ on (Y, X, U) which together with S satisfies the following conditions: the operator $\left[\begin{smallmatrix} 1 & 0 \\ C\&D \end{smallmatrix}\right]$ maps $\mathcal{D}(S)$ continuously onto $\mathcal{D}(S^\times)$, its inverse is $\left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$, and

$$\begin{aligned} S^\times = \begin{bmatrix} [A\&B]^\times \\ [C\&D]^\times \end{bmatrix} &= \begin{bmatrix} A\&B \\ 0 \quad 1 \end{bmatrix} \begin{bmatrix} 1 \quad 0 \\ C\&D \end{bmatrix}^{-1}, \\ S = \begin{bmatrix} A\&B \\ C\&D \end{bmatrix} &= \begin{bmatrix} [A\&B]^\times \\ 0 \quad 1 \end{bmatrix} \begin{bmatrix} 1 \quad 0 \\ [C\&D]^\times \end{bmatrix}^{-1}. \end{aligned} \tag{21.14}$$

In this case we call S and $S^\times$ *flow-inverses* of each other. □

Obviously, the flow-inverse of a node S in unique (when it exists). Furthermore, by [12, Corollary 5.3], in the well-posed case this notion agrees with the notion of flow-inversion introduced in [12].

The following theorem lists a number of alternative characterizations for the flow-invertibility of a system node.[8]

THEOREM 21.2
Let $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ be a system node on (U, X, Y), with main operator A, control operator B, observation operator C, and transfer function $\mathfrak{D}$, and let $S^\times = \left[\begin{smallmatrix} [A\&B]^\times \\ [C\&D]^\times \end{smallmatrix}\right]$ be a system node on (Y, X, U), with main operator $A^\times$, control operator $B^\times$, observation operator $C^\times$, and transfer function $\mathfrak{D}^\times$. We denote $\mathcal{D}(A) = X_1$, $(\mathcal{D}(A^*))^* = X_{-1}$, $\mathcal{D}(A^\times) = X_1^\times$, and $(\mathcal{D}((A^\times)^*))^* = X_{-1}$. Then the following conditions are equivalent:

[7] Flow-inversion can be interpreted as a special case of output feedback, and conversely, output feedback can be interpreted as a special case of flow-inversion. See [12, Remark 5.5].

[8] In this list we have not explicitly included the equivalent discrete time eigenvalue conditions that can be derived from the alternative characterization of continuous time flow-inversion as a Cayley transform, followed by a discrete time flow inversion, followed by an inverse Cayley transform.

(i) S and $S^\times$ are flow-inverses of each other.

(ii) The operator $\left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$ maps $\mathcal{D}(S^\times)$ one-to-one onto $\mathcal{D}(S)$, and

$$\begin{bmatrix} [A\&B]^\times \\ 0 \quad 1 \end{bmatrix} = \begin{bmatrix} A\&B \\ C\&D \end{bmatrix} \begin{bmatrix} 1 & 0 \\ [C\&D]^\times \end{bmatrix} \text{ (on } \mathcal{D}(S^\times)). \tag{21.15}$$

(iii) For all $\alpha \in \rho(A^\times)$, the operator $\left[\begin{smallmatrix} \alpha & 0 \\ 0 & 0 \end{smallmatrix}\right] - S$ maps $\mathcal{D}(S)$ one-to-one onto $\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$, and its (bounded) inverse is given by

$$\left(\begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix} - S \right)^{-1} = \begin{bmatrix} (\alpha - A^\times)^{-1} & -(\alpha - A^\times_{|X})^{-1} B^\times \\ C^\times (\alpha - A^\times)^{-1} & -\widehat{\mathfrak{D}}^\times(\alpha) \end{bmatrix}. \tag{21.16}$$

(iv) For some $\alpha \in \rho(A^\times)$, the operator $\left[\begin{smallmatrix} \alpha & 0 \\ 0 & 0 \end{smallmatrix}\right] - S$ maps $\mathcal{D}(S)$ one-to-one onto $\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$ and (21.16) holds.

(v) For all $\alpha \in \rho(A) \cap \rho(A^\times)$, $\widehat{\mathfrak{D}}(\alpha)$ is invertible and the following operator identity holds in $\mathcal{L}\left(\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]; \mathcal{D}(S)\right)$:

$$\begin{aligned} \begin{bmatrix} (\alpha - A^\times)^{-1} & -(\alpha - A^\times_{|X})^{-1} B^\times \\ C^\times (\alpha - A^\times)^{-1} & -\widehat{\mathfrak{D}}^\times(\alpha) \end{bmatrix} &= \begin{bmatrix} (\alpha - A)^{-1} & 0 \\ 0 & 0 \end{bmatrix} \\ - \begin{bmatrix} (\alpha - A_{|X})^{-1} B \\ 1 \end{bmatrix} &[\widehat{\mathfrak{D}}(\alpha)]^{-1} \begin{bmatrix} C(\alpha - A)^{-1} & 1 \end{bmatrix}. \end{aligned} \tag{21.17}$$

(vi) For some $\alpha \in \rho(A) \cap \rho(A^\times)$, $\widehat{\mathfrak{D}}(\alpha)$ is invertible and (21.17) holds.

When these equivalent conditions hold, then $\left[\begin{smallmatrix} 1 \\ C \end{smallmatrix}\right]$ maps X_1 into $\mathcal{D}(S^\times)$, $\left[\begin{smallmatrix} 1 \\ C^\times \end{smallmatrix}\right]$ maps $X_1^\times$ into $\mathcal{D}(S)$, and

$$\begin{aligned} A &= A^\times_{|X_1} + B^\times C, & A^\times &= A_{|X_1^\times} + BC^\times, \\ 0 &= [C\&D]^\times \begin{bmatrix} 1 \\ C \end{bmatrix}, & 0 &= C\&D \begin{bmatrix} 1 \\ C^\times \end{bmatrix}. \end{aligned} \tag{21.18}$$

□

Proof We begin by observing that (21.18), which is equivalent to

$$\begin{bmatrix} [A\&B]^\times \\ [C\&D]^\times \end{bmatrix} \begin{bmatrix} 1 \\ C \end{bmatrix} = \begin{bmatrix} A \\ 0 \end{bmatrix}, \quad \begin{bmatrix} A\&B \\ C\&D \end{bmatrix} \begin{bmatrix} 1 \\ C^\times \end{bmatrix} = \begin{bmatrix} A^\times \\ 0 \end{bmatrix}, \tag{21.19}$$

follows from (i) and (21.14) since $\left[\begin{smallmatrix} X_1 \\ 0 \end{smallmatrix}\right] \in \mathcal{D}(S)$ and $\left[\begin{smallmatrix} X_1^\times \\ 0 \end{smallmatrix}\right] \in \mathcal{D}(S_\times)$.

(i) $\Rightarrow$ (ii): This is obvious (see Definition 21.1).

(ii) $\Rightarrow$ (i): Suppose that (ii) holds. Then $\left[\begin{smallmatrix} 1 & 0 \\ C\&D \end{smallmatrix}\right] \left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ on $\mathcal{D}(S^\times)$ (by assumption, $C\&D \left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right] = [0 \quad 1]$, and we have $[1 \quad 0] \left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right] = [1 \quad 0]$). Thus, $\left[\begin{smallmatrix} 1 & 0 \\ C\&D \end{smallmatrix}\right]$ is a left-inverse of $\left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$. However, as (by assumption) $\left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$ is both one-to-one and onto, it is invertible, so the left inverse is also a right

inverse, i.e., the inverse of $\left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$ is $\left[\begin{smallmatrix} 1 & 0 \\ C\&D \end{smallmatrix}\right]$. Multiplying (21.15) to the right by $\left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]^{-1}$ we get the second identity in (21.14). The first identity in (21.14) can equivalently be written in the form $\left[\begin{smallmatrix} [A\&B]^\times \\ [C\&D]^\times \end{smallmatrix}\right] = \left[\begin{smallmatrix} A\&B & 0 \\ 0 & 1 \end{smallmatrix}\right] \left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$. The top part $[A\&B]^\times = A\&B \left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$ of this identity is contained in (21.15)), and the bottom part $[C\&D]^\times = [0 \quad 1] \left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$ is always valid. We conclude that (ii) $\Rightarrow$ (i).

(ii) $\Rightarrow$ (iii): Let $\alpha \in \mathbb{C}$ be arbitrary. Clearly, (ii) is equivalent to the requirement that $\left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$ maps $\mathcal{D}(S^\times)$ one-to-one onto $\mathcal{D}(S)$, combined with the identity (note that $[\alpha \quad 0] \left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right] = [\alpha \quad 0]$)

$$\left(\begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix} - S\right) \begin{bmatrix} 1 & 0 \\ [C\&D]^\times \end{bmatrix} = \left(\begin{bmatrix} \alpha & 0 \\ 0 & -1 \end{bmatrix} - \begin{bmatrix} [A\&B]^\times \\ 0 \end{bmatrix}\right) \text{ (on } \mathcal{D}(S^\times)).$$

If $\alpha \in \rho(A^\times)$, then $\left[\begin{smallmatrix} (\alpha-A^\times)^{-1} & (\alpha-A^\times_{|X})^{-1}B^\times \\ 0 & 1 \end{smallmatrix}\right]$ maps $\left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right]$ one-to-one onto $\mathcal{D}(S^\times)$, so we may multiply the above identity by this operator to the right to get the equivalent identity

$$\left(\begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix} - S\right) \begin{bmatrix} (\alpha - A^\times)^{-1} & (\alpha - A^\times_{|X})^{-1}B^\times \\ C^\times(\alpha - A^\times)^{-1} & \widehat{\mathfrak{D}}^\times(\alpha) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

which is now valid on all of $\left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right]$. This can alternatively be written as (multiply by $\left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ to the right)

$$\left(\begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix} - S\right) \begin{bmatrix} (\alpha - A^\times)^{-1} & -(\alpha - A^\times_{|X})^{-1}B^\times \\ C^\times(\alpha - A^\times)^{-1} & -\widehat{\mathfrak{D}}^\times(\alpha) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

By tracing the history of the second factor on the left-hand side we find that it maps $\left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right]$ one-to-one onto $\mathcal{D}(S)$. Thus, $\left[\begin{smallmatrix} \alpha & 0 \\ 0 & 0 \end{smallmatrix}\right] - S$ is the left-inverse of an invertible operator, hence invertible, and (21.16) holds.

(iii) $\Rightarrow$ (iv): This is obvious.

(iv) $\Rightarrow$ (ii): This is the same computation that we did in the proof of the implication (ii) $\Rightarrow$ (iii) done backwards, for one particular value of $\alpha \in \rho(A^\times)$. Observe, in particular, that $\left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times \end{smallmatrix}\right]$ maps $\mathcal{D}(S^\times)$ one-to-one onto $\mathcal{D}(S)$ if and only if the operator on the right-hand side of (21.16) maps $\left[\begin{smallmatrix} X \\ U \end{smallmatrix}\right]$ one-to-one onto $\mathcal{D}(S)$.

(iii) $\Rightarrow$ (v): This follows from the easily verified identity

$$\begin{aligned} &\left(\begin{bmatrix} \alpha & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} A\&B \\ C\&D \end{bmatrix}\right) \\ &\quad = \begin{bmatrix} 1 & 0 \\ -C(\alpha - A)^{-1} & 1 \end{bmatrix} \begin{bmatrix} \alpha - A & 0 \\ 0 & -\widehat{\mathfrak{D}}(\alpha) \end{bmatrix} \begin{bmatrix} 1 & -(\alpha - A_{|X})^{-1}B \\ 0 & 1 \end{bmatrix}. \end{aligned} \tag{21.20}$$

valid for all $\alpha \in \rho(A)$.

(v) $\Rightarrow$ (vi): This is obvious.

(vi) $\Rightarrow$ (iv): Argue as in the proof of the implication (iii) $\Rightarrow$ (v). □

The original idea behind the flow-inversion of a well-posed system introduced in [12, Section 5] was to interchange the roles of the input and output. A similar interpretation is valid for the flow-inversion of system nodes, too.

THEOREM 21.3
Let $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ be a flow-invertible system node on (Y, X, U), whose flow-inverse $S^\times$ is also a system node (on (U, X, Y)). Let x and y be the state trajectory and output of S with initial state $x_0 \in X$ and input function $u \in L^1_{loc}(\mathbb{R}^+; U)$, and suppose that $x \in W^{1,1}_{loc}(\mathbb{R}^+; X)$. Then $y \in L^1_{loc}(\mathbb{R}^+; Y)$, and x and u are the state trajectory and output of $S^\times$ with initial state x_0 and input function y. □

Proof By Lemma 21.3, $\left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] \in L^1_{loc}(\mathbb{R}^+; \mathcal{D}(S))$, $y \in L^1_{loc}(\mathbb{R}^+; Y)$, and $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ is the unique solution with the above properties of the equation

$$\begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} = S \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \text{ for almost all } t \geq s, \quad x(s) = x_s.$$

$\left[\begin{smallmatrix} 1 & 0 \\ C\&D & \end{smallmatrix}\right]$ maps $\mathcal{D}(S)$ continuously on $\mathcal{D}(S^\times)$, so $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 & 0 \\ C\&D & \end{smallmatrix}\right] \left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] \in L^1_{loc}(\mathbb{R}^+; \mathcal{D}(S^\times))$. Moreover, since $\left[\begin{smallmatrix} 1 & 0 \\ C\&D & \end{smallmatrix}\right]^{-1} = \left[\begin{smallmatrix} 1 & 0 \\ [C\&D]^\times & \end{smallmatrix}\right]$, we have for almost all $t \geq s$,

$$\begin{aligned} \begin{bmatrix} x'(t) \\ u(t) \end{bmatrix} &= \begin{bmatrix} A\&B \\ 0 \quad 1 \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} A\&B \\ 0 \quad 1 \end{bmatrix} \begin{bmatrix} 1 \quad 0 \\ [C\&D]^\times \end{bmatrix} \begin{bmatrix} 1 \quad 0 \\ C\&D \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \\ &= \begin{bmatrix} [A\&B]^\times \\ [C\&D]^\times \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}. \end{aligned}$$

By Lemma 21.3, this implies that x and u are the state and output function of $S^\times$ with initial time s, initial state x_s, and input function y. □

Our next theorem shows that compatibility is preserved under flow-inversion in most cases.

THEOREM 21.4
Let $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ be a compatible system node on (Y, X, U), and let $\left[\begin{smallmatrix} A_{|W} & B \\ C_{|W} & D \end{smallmatrix}\right] \in \mathcal{L}(\left[\begin{smallmatrix} W \\ U \end{smallmatrix}\right]; \left[\begin{smallmatrix} W_{-1} \\ Y \end{smallmatrix}\right])$ be a compatible extension of S (here $X_1 \subset W \subset X$ and W_{-1} is defined as in Section 21.2). Suppose that S if flow-invertible. Denote the flow-inverted system node by $S^\times = \left[\begin{smallmatrix} [A\&B]^\times \\ [C\&D]^\times \end{smallmatrix}\right]$, let $X_1^\times$ and $X_{-1}^\times$ be the analogues of X_1 and X_{-1} for $S^\times$, and let $W_{-1}^\times$ be the analogue of W_{-1} for $S^\times$ (i.e., $W_{-1}^\times = (\alpha - A^\times_{|W})W$ for some $\alpha \in \rho(A^\times)$).

(i) If D has a left inverse $D^{-1}_{left} \in \mathcal{L}(Y; U)$, then $X_1^\times \subset W$ and $S^\times$ is compatible with extended observation operator $C^\times_{|W} : W \to U$ and corresponding feedthrough operator $D^\times$ given by

$$\begin{aligned} C^\times_{|W} &= -D^{-1}_{left} C_{|W}, \\ D^\times &= D^{-1}_{left}, \end{aligned} \tag{21.21}$$

and the the main operator $A^\times$ of $S^\times$ is given by

$$A^\times = \left(A_{|X} - B D^{-1}_{left} C_{|W}\right)_{|X_1^\times}.$$

In this case the space W_{-1} can be identified with a closed subspace of $W_{-1}^\times$, so that $X \subset W_{-1} \subset X_{-1} \cap X_{-1}^\times$. With this identification,

$$A_{|W} = A^\times_{|W} + B^\times C_{|W}, \qquad B = B^\times D$$

(where we by $A_{|W}$ and $A^\times_{|W}$ mean the restrictions of $A_{|X}$ and $A^\times_{|X}$ to W).

(ii) If D is invertible (with a bounded inverse), then $W_{-1} = W_{-1}^\times$, $A^\times W \subset W_{-1}$, $B^\times U \subset W_{-1}$, and the operator $\left[\begin{smallmatrix} A^\times_{|W} & B^\times \\ C^\times_{|W} & D^\times \end{smallmatrix}\right] \in \mathcal{L}\left(\left[\begin{smallmatrix} W \\ U \end{smallmatrix}\right]; \left[\begin{smallmatrix} W_{-1} \\ Y \end{smallmatrix}\right]\right)$ defined by

$$\begin{aligned}\begin{bmatrix} A^\times_{|W} & B^\times \\ C^\times_{|W} & D^\times \end{bmatrix} &= \begin{bmatrix} A_{|W} - BD^{-1}C_{|W} & BD^{-1} \\ -D^{-1}C_{|W} & D^{-1} \end{bmatrix} \\ &= \begin{bmatrix} A_{|W} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B \\ 1 \end{bmatrix} D^{-1} \begin{bmatrix} -C_{|W} & 1 \end{bmatrix} \\ &= \begin{bmatrix} A_{|W} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B \\ 1 \end{bmatrix} \begin{bmatrix} C^\times_{|W} & 1 \end{bmatrix} \\ &= \begin{bmatrix} A_{|W} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} B^\times \\ 1 \end{bmatrix} \begin{bmatrix} -C_{|W} & 1 \end{bmatrix}\end{aligned}$$

is a compatible extension of $S^\times$.

□

Proof (i) Take $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] \in \mathcal{D}(S^\times)$, and define $u = [C\&D]^\times \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$. Then $\left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] \in \mathcal{D}(S)$ and $y = C\&D \left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] = C_{|W}x + Du$. Multiplying the above identity by D_{left}^{-1} to the left we get for all $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] \in \mathcal{D}(S^\times)$,

$$u = [C\&D]^\times \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] = -D_{left}^{-1}C_{|W}x + D_{left}^{-1}y.$$

The right-hand side is defined (and continuous) on all of $W \times Y$. By (21.17), for all $y \in Y$ and all $\alpha \in \rho(A) \cap \rho(A^\times)$,

$$(\alpha - A^\times_{|X})^{-1}B^\times y = (\alpha - A_{|X})^{-1}B\widehat{\mathfrak{D}}^\times(\alpha)y \in W,$$

so $\mathcal{R}(B^\times) \in W_{-1}^\times$. This implies that $\left[\begin{smallmatrix} A^\times_{|W} & B^\times \\ C^\times_{|W} & D^\times \end{smallmatrix}\right]$ is a compatible extension of $S^\times$, with $C^\times_{|W} = -D_{left}^{-1}C_{|W}$ and $D^\times = D_{left}^{-1}$. By (21.18), for all $x \in X_1^\times$, we have $A^\times x = (A_{|X} + BC^\times)x = (A_{|X} - BD_{left}^{-1}C_{|W})x$, as claimed.

Next we construct an embedding operator $J : W_{-1} \to W_{-1}^\times$. This operator is required to be one-to-one, and its restriction to X should be the identity operator. We define

$$\begin{aligned} J &= (\alpha - A^\times_{|W} - B^\times C_{|W})(\alpha - A_{|W})^{-1}, \\ J^\times &= (\alpha - A_{|W} - BC^\times_{|W})(\alpha - A^\times_{|W})^{-1}. \end{aligned} \tag{21.22}$$

The compatibility of S and $S^\times$ implies $J \in \mathcal{L}(W_{-1}; W_{-1}^\times)$ and $J^\times \in \mathcal{L}(W_{-1}^\times; W_{-1})$ and by (21.18), both J and $J^\times$ reduce to the identity operator on X.

We claim that $J^\times \in \mathcal{L}(W_{-1}^\times; W_{-1})$ is a left inverse of $J \in \mathcal{L}(W_{-1}; W_{-1}^\times)$, or equivalently, that $(\alpha - A_{|W})^{-1}J^\times J(\alpha - A_{|W})$ is the identity on W. To see that this

is the case we use (21.22), (21.21), (21.17), and (21.7) (in this order) to compute

$$\begin{aligned}
&(\alpha - A_{|W})^{-1} J^{\times} J (\alpha - A_{|W}) \\
&\quad = (\alpha - A_{|W})^{-1} (\alpha - A_{|W} - B C_{|W}^{\times}) \\
&\qquad \times (\alpha - A_{|W}^{\times})^{-1} (\alpha - A_{|W}^{\times} - B^{\times} C_{|W}) \\
&\quad = (1 - (\alpha - A_{|W})^{-1} B C_{|W}^{\times})(1 - (\alpha - A_{|W}^{\times})^{-1} B^{\times} C_{|W}) \\
&\quad = (1 + (\alpha - A_{|W})^{-1} B D_{left}^{-1} C_{|W})(1 - (\alpha - A_{|W})^{-1} B \widehat{\mathfrak{D}}^{-1}(\alpha) C_{|W}) \\
&\quad = 1 + (\alpha - A_{|W})^{-1} B \left[D_{left}^{-1} - \widehat{\mathfrak{D}}^{-1}(\alpha) - D_{left}^{-1} C_{|W} (\alpha - A_{|W})^{-1} B \widehat{\mathfrak{D}}^{-1}(\alpha) \right] C_{|W} \\
&\quad = 1 + (\alpha - A_{|W})^{-1} B D_{left}^{-1} \left[\widehat{\mathfrak{D}}(\alpha) - D - C_{|W} (\alpha - A_{|W})^{-1} B \right] \widehat{\mathfrak{D}}^{-1}(\alpha) C_{|W} \\
&\quad = 1.
\end{aligned}$$

This implies that the operator J is one-to-one; hence it defines a (not necessarily dense) embedding of W_{-1} into $W_{-1}^{\times}$. In the sequel we shall identify W_{-1} with the range of J. That W_{-1} is closed in $W_{-1}^{\times}$ follows from the fact that J has a bounded left inverse.

The identification of W_{-1} with a subspace of $W_{-1}^{\times}$ means that the embedding operator $J = (\alpha - A_{|W}^{\times} - B^{\times} C_{|W})(\alpha - A_{|W})^{-1}$ becomes the identity on W_{-1}, and hence, with this identification, $(\alpha - A_{|W}) = (\alpha - A_{|W}^{\times} - B^{\times} C_{|W})$, or equivalently,

$$A_{|W} = A_{|W}^{\times} + B^{\times} C_{|W}.$$

The remaining identity $B = B^{\times} D$ can verified as follows. By (21.17) and the fact that $A_{|W}^{\times} = A_{|W} - B^{\times} C_{|W}$,

$$\begin{aligned}
B^{\times} \widehat{\mathfrak{D}}(\alpha) &= (\alpha - A_{|W}^{\times})(\alpha - A_{|W})^{-1} B \\
&= (\alpha - A_{|W} + B^{\times} C_{|W})(\alpha - A_{|W})^{-1} B \\
&= (B + B^{\times} C_{|W} (\alpha - A_{|W})^{-1} B) \\
&= (B + B^{\times} (\widehat{\mathfrak{D}}(\alpha) - D)) \\
&= B^{\times} \widehat{\mathfrak{D}}(\alpha) + B - B^{\times} D.
\end{aligned}$$

Thus $B = B^{\times} D$.

(ii) Part (ii) follows from part (i) if we interchange S and $S^{\times}$. (This will also interchange W_{-1} with $W_{-1}^{\times}$ and J with $J^{\times}$.) □

21.5 The Diagonal Transform

With the theory that we developed in the preceding section at our disposal we can now proceed in the same way as we did in [8, Section 5] to investigate the continuous time diagonal transform. First of all, by comparing (21.13) and (21.17) we observe that it is possible to reduce the continuous time diagonal transform to flow-inversion in the following way.

DEFINITION 21.1
Let $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ be a system node on (U, X, U) (note that $Y = U$). We call S *diagonally transformable* if the system node $\widetilde{S} = \left[\begin{smallmatrix} A\&B \\ \widetilde{C\&D} \end{smallmatrix}\right]$ is flow-invertible, where

$$\widetilde{C\&D} = \frac{1}{\sqrt{2}}\Big(C\&D + [\,0 \quad 1\,]\Big).$$

Denote the flow-inverse of this system node by $\widetilde{S}^\times = \left[\begin{smallmatrix} [A\&B]^\times \\ [\widetilde{C\&D}]^\times \end{smallmatrix}\right]$. Then the *diagonal transform* of S is the system node $S^\times = \left[\begin{smallmatrix} [A\&B]^\times \\ [C\&D]^\times \end{smallmatrix}\right]$, where

$$[C\&D]^\times = \sqrt{2}\,[\widetilde{C\&D}]^\times - [\,0 \quad 1\,].$$

□

The diagonal transform can be computed more explicitly as follows.

COROLLARY 21.2
Let $S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right]$ be a diagonally transformable system node on (U, X, U). Then the diagonal transform $S^\times = \left[\begin{smallmatrix} [A\&B]^\times \\ [C\&D]^\times \end{smallmatrix}\right]$ of S satisfies

$$S^\times + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} A\&B \\ 0 \quad 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \\ C\&D \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{2} \end{bmatrix}.$$

If S is compatible with a compatible extension $\left[\begin{smallmatrix} A_{|W} & B \\ C_{|W} & D \end{smallmatrix}\right] \in \mathcal{L}(\left[\begin{smallmatrix} W \\ U \end{smallmatrix}\right]; \left[\begin{smallmatrix} W_{-1} \\ U \end{smallmatrix}\right])$ where $1 + D$ invertible, then $S^\times$ is also compatible, with the compatible extension (over the same space W)

$$\begin{aligned} \begin{bmatrix} A^\times_{|W} & B^\times \\ C^\times_{|W} & D^\times \end{bmatrix} &= \begin{bmatrix} A_{|W} & 0 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} B \\ \sqrt{2} \end{bmatrix} (1+D)^{-1} [\,-C_{|W} \quad \sqrt{2}\,] \\ &= \begin{bmatrix} A_{|W} - B(1+D)^{-1}C_{|W} & \sqrt{2}B(1+D)^{-1} \\ -\sqrt{2}(1+D)^{-1}C_{|W} & (1-D)(1+D)^{-1} \end{bmatrix}. \end{aligned} \tag{21.23}$$

□

This follows directly from Definition 21.1 and Theorems 21.3 and 21.4.

COROLLARY 21.3
Example 21.6 is a scattering conservative system node. □

This follows from Theorem 21.7 and Corollary 21.2.

REMARK 21.4
By applying the same theory to other examples of impedance passive or conservative systems we can create many more examples of continuous time scattering passive or conservative systems. One particularly interesting class is the one which is often referred to as 'systems with collocated actuators and sensors', discussed in, e.g., [1], [13], and [14]. □

21.6 References

[1] B.-Z. Guo and Y.-H. Luo. Controllability and stability of a sencond-order hyperbolic system with collocated sensor/actuator. *Systems Control Lett.*, 46:45–65, 2002.

[2] J. W. Helton. Systems with infinite-dimensional state space: the Hilbert space approach. *Proceedings of the IEEE*, 64:145–160, 1976.

[3] J. Malinen, O. J. Staffans, and G. Weiss. When is a linear system conservative? In preparation, 2002.

[4] D. Salamon. Infinite dimensional linear systems with unbounded control and observation: a functional analytic approach. *Trans. Amer. Math. Soc.*, 300:383–431, 1987.

[5] D. Salamon. Realization theory in Hilbert space. *Math. Systems Theory*, 21:147–164, 1989.

[6] Y. L. Smuljan. Invariant subspaces of semigroups and the Lax-Phillips scheme. Dep. in VINITI, N 8009-1386, Odessa, 49pp., 1986.

[7] O. J. Staffans. J-energy preserving well-posed linear systems. *Int. J. Appl. Math. Comput. Sci.*, 11:1361–1378, 2001.

[8] O. J. Staffans. Passive and conservative continuous time impedance and scattering systems. Part I: Well-posed systems. *Math. Control Signals Systems*, 2002. To appear.

[9] O. J. Staffans. Passive and conservative infinite-dimensional impedance and scattering systems (from a personal point of view). In *Mathematical Systems Theory in Biology, Communication, Computation, and Finance*, J. Rosenthal and D. S. Gilliam, eds., IMA Volumes in Mathematics and its Applications, New York, 2002. Springer Verlag.

[10] O. J. Staffans. *Well-Posed Linear Systems: Part I*. Book manuscript, available at http://www.abo.fi/˜staffans/, 2002.

[11] O. J. Staffans and G. Weiss. Transfer functions of regular linear systems. Part II: the system operator and the Lax-Phillips semigroup. *Trans. Amer. Math. Soc.*, 354:3229–3262, 2002.

[12] O. J. Staffans and G. Weiss. Transfer functions of regular linear systems. Part III: inversions and duality. Submitted, 2002.

[13] G. Weiss. Optimal control of systems with a unitary semigroup and with colocated control and observation. *Systems Control Lett.*, 2002.

[14] G. Weiss and R. F. Curtain. Exponential stabilization of vibrating systems by collocated feedback. In *Proceedings of the 7th IEEE Mediterranean Conference on Control and Systems*, pages 1705–1722, CD-ROM, Haifa, Israel, July 28–30 1999.

[15] G. Weiss and M. Tucsnak. How to get a conservative well-posed linear system out of thin air. Part I: well-posedness and energy balance. Submitted, 2001.

[16] J. C. Willems. Dissipative dynamical systems Part I: General theory. *Arch. Rational Mech. Anal.*, 45:321–351, 1972.

[17] J. C. Willems. Dissipative dynamical systems Part II: Linear systems with quadratic supply rates. *Arch. Rational Mech. Anal.*, 45:352–393, 1972.

22

High-Order Open Mapping Theorems

Héctor J. Sussmann

Abstract

The well-known finite-dimensional first-order open mapping theorem says that a continuous map with a finite-dimensional target is open at a point if its differential at that point exists and is surjective. An identical result, due to Graves, is true when the target is infinite-dimensional, if "differentiability" is replaced by "strict differentiability." We prove general theorems in which the linear approximations involved in the usual concept of differentiability is replaced by an approximation by a map which is homogeneous relative to a one-parameter group of dilations, and the error bound in the approximation involves a "homogeneous pseudonorm" or a "homogeneous pseudodistance," rather than the ordinary norm. We outline how these results can be used to derive sufficient conditions for openness involving higher-order derivatives, and carry this out in detail for the second-order case.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 293-316, 2003.
© Springer-Verlag Berlin Heidelberg 2003

22.1 Introduction

Open mapping theorems (abbr. OMTs) are key tools in the derivation of necessary conditions for an optimum in optimization theory, and in particular in optimal control. For example, the usual Lagrange multipliers rule is a trivial corollary of the following OMT.

THEOREM 22.1
Let X, Y be normed spaces, let Ω be an open subset of X, and let $F : \Omega \mapsto Y$ be a continuous map. Let $x_* \in \Omega$ be such that F is Fréchet differentiable at x_* and the differential $DF(x_*)$ is a surjective linear map from X to Y. Assume that Y is finite-dimensional. Then F is open at x_*, that is, F maps neighborhoods of x_* to neighborhoods of $F(x_*)$. □

(To deduce the Lagrange rule from Theorem 22.1, let us consider the problem of minimizing the quantity $f_0(x)$ subject to finitely many equality constraints $f_1(x) = \cdots = f_m(x) = 0$, where $f_0, f_1, \ldots, f_m$ are continuous real-valued functions on an open subset Ω of a normed space X, and assume that x_* is a solution. We then define $F : \Omega \mapsto \mathbb{R}^{m+1}$ by letting

$$F(x) = (f_0(x), f_1(x), \ldots, f_m(x)) \quad \text{for} \quad x \in \Omega\,,$$

and observe that, if we let $y_*(\varepsilon) = (f_0(x_*) - \varepsilon, 0, \ldots, 0) \in \mathbb{R}^{m+1}$ for $\varepsilon \in \mathbb{R}$, then $F(x_*) = y_*(0)$, and $y_*(\varepsilon) \notin F(\Omega)$ whenever $\varepsilon > 0$, so F is not open at x_*. If $f_0, f_1, \ldots, f_m$ are Fréchet differentiable at x_*, then F is Fréchet differentiable at x_*, so Theorem 22.1 implies that $DF(x_*)$ is not surjective. Therefore the $m+1$ vectors $\nabla f_0(x_*), \nabla f_1(x_*), \ldots, \nabla f_m(x_*)$ are not linearly independent. It follows that there exist numbers $\lambda_0, \lambda_1, \ldots, \lambda_m$ that are not all zero and satisfy $\sum_{i=0}^m \lambda_i \nabla f_i(x_*) = 0$.)

Theorem 22.1 is only valid if Y is finite-dimensional. (Precisely: if Y is any normed space, then Y is finite-dimensional if and only if it has the property that whenever $F : Y \mapsto Y$ is a continuous map such that $F(0) = 0$, F is Fréchet differentiable at 0, and the differential $DF(0)$ is the identity map of Y, it follows that F is open at 0.) On the other hand, L.M. Graves proved in [4] (cf. also Dontchev [3]) the following infinite-dimensional result.

THEOREM 22.2
Let X, Y be Banach spaces, let Ω be an open subset of X, and let $F : \Omega \mapsto Y$ be a continuous map. Let $x_* \in \Omega$ be such that F is strictly differentiable at x_* and the differential $DF(x_*)$ is a surjective linear map from X to Y. Then F is open at x_*. □

(Graves actually proved a stronger result, in which strict differentiability with a surjective differential is replaced by the weaker condition that there exist a surjective bounded linear map $L : X \mapsto Y$ such that

$$\limsup_{x \to x_*, x' \to x_*, x \neq x'} \frac{\|F(x) - F(x') - L(x - x')\|}{\|x - x'\|} < \|L^{-1}\|^{-1}\,, \tag{22.1}$$

where $\|L^{-1}\|$ is the infimum of the numbers C such that for every $y \in Y$ there exists $x \in X$ such that $L(x) = y$ and $\|x\| \leq C\|y\|$. The definition of "strict

differentiability" is as follows: F is strictly differentiable at x_* with differential L if the left-hand side of (22.1) vanishes.)

The purpose of this note is to present "higher-order" sufficient conditions for openness at a point, generalizing Theorems 22.1 and 22.2.

REMARK 22.3
Stronger results can also be proved, in which

1. in the infinite-dimensional case, the analogue of strict differentiability is replaced by the analogue of condition (22.1);

2. in the finite-dimensional case, the analogue of differentiability is replaced by the analogue of the condition that

$$\limsup_{x\to x_*, x\neq x_*} \frac{\|F(x)-F(x_*)-L(x-x_*)\|}{\|x-x_*\|} < \|L^{-1}\|^{-1};$$

3. F is only required to be defined on a "conic neighborhood" of x_*, i.e., a set S of the form

$$S = S_X(x_*,\varepsilon,C) \stackrel{\text{def}}{=} \{x : \|x-x_*\| < \varepsilon \wedge x - x_* \in C\},$$

where $\varepsilon > 0$ and C is a convex cone in X with nonempty interior such that $0 \in C$;

4. the conclusion says that the map F is "directionally open" at x_* in the direction of a given vector $v_* \in Y$, provided that $v_* \in \operatorname{int} LC$. (Precisely: for every positive ε there exists a positive δ such that the image $F(S_X(x_*,\varepsilon,C))$ contains the set $S_Y(F(x_*),\delta,D)$ for some convex cone D in Y such that $v_* \in \operatorname{int} D$.)

In this paper, however, we will only carry out the simpler task of generalizing the non-directional theorems 22.1 and 22.2. □

Naturally, "higher-order" means "involving higher-oder derivatives." A map F of class C^m has a Taylor approximation

$$F(x_*+h) \sim F(x_*) + P_1(h) + \frac{1}{2}P_2(h,h) + \cdots + \frac{1}{m!}P_m(h,h,\ldots,h),$$

where $P_j = D^jF(x_*)$ for $j = 1,\ldots,m$, so each P_j is a Y-valued, continuous, symmetric, multilinear map defined on the Cartesian product X^j of j copies of X. Openness at x_* follows from the first-order theorems if the linear map P_1 is surjective. When P_1 is not surjective, the high-order results will give openness if the missing directions in the image P_1X can somehow be realized as image directions using higher-order derivatives. On the other hand, if the "first-order effect" $P_1(h)$ of a particular direction h is nonzero, then this effect will dominate whatever contributions h may make through higher-order terms. This suggests that one should only look at the second-order effects $P_2(h,h)$ of vectors h belonging

to the kernel K_2 of P_1. So, if $m = 2$, we take $K_1 = X$, $K_2 = \ker P_1$, and consider the approximation

$$F(x_* + h_1 + h_2) \sim F(x_*) + P_1(h_1) + P_1(h_2) + \frac{1}{2}P_2(h_1, h_1) + \frac{1}{2}P_2(h_2, h_2) + P_2(h_1, h_2)\,,$$

where $h_1 \in K_1$, $h_2 \in K_2$. The remainder is clearly $o(\|h_1\|^2 + \|h_2\|^2)$, which is in particular $o(\|h_1\| + \|h_2\|^2)$. Furthermore, the terms $\frac{1}{2}P_2(h_1, h_1)$ and $P_2(h_1, h_2)$ are $o(\|h_1\|)$, so they are $o(\|h_1\| + \|h_2\|^2)$ as well, and can be absorbed into the remainder. Finally, $P_1(h_2)$ vanishes, because $h_2 \in \ker P_1$, and we end up with the approximation

$$F(x_* + h_1 + h_2) = F(x_*) + P_1(h_1) + \frac{1}{2}P_2(h_2, h_2) + o(\|h_1\| + \|h_2\|^2) \quad \text{as } h_1 \to 0,\ h_2 \to 0,\ h_1 \in K_1,\ h_2 \in K_2\,. \tag{22.2}$$

Such an approximation is valid under more general situations, even if F is not of class C^2 near x_*. (For example, if $F : \mathbb{R}^3 \mapsto \mathbb{R}$ is given by

$$F(x, y, z) = x + |x|^{3/2}(1 + |y|^{1/2}) + y^2 - z^2\,,$$

then $F(x, y, z) = x + y^2 - z^2 + o(|x| + y^2 + z^2)$, so (22.2) holds with $K_1 = \mathbb{R}^3$, $K_2 = \{0\} \times \mathbb{R}^2$, and obvious choices of P_1 and P_2.) If we define $\mathcal{X} = K_1 \times K_2$, and write

$$\begin{aligned} \mathcal{F}(h_1, h_2) &= F(x_* + h_1 + h_2) - F(x_*)\,, \\ \mathcal{G}(h_1, h_2) &= P_1(h_1) + \frac{1}{2}P_2(h_2, h_2)\,, \\ \nu(h_1, h_2) &= \|h_1\| + \|h_2\|^2\,, \end{aligned}$$

for $h_1 \in K_1$, $h_2 \in K_2$, then (22.2) says that

$$\mathcal{F}(\xi) = \mathcal{G}(\xi) + o(\nu(\xi)) \qquad \text{as } \xi \to 0\,, \quad \xi \in \mathcal{X}\,. \tag{22.3}$$

Moreover, if we let

$$\delta_t(h_1, h_2) = (th_1, t^{1/2}h_2) \qquad \text{for} \qquad h_1 \in K_1,\ h_2 \in K_2,\ t > 0\,,$$

then

(HA.1) $\boldsymbol{\delta} = \{\delta_t\}_{t>0}$ *is a continuous one-parameter group of dilations on* $\mathcal{X}$,

(HA.2) $\mathcal{G}$ *is* δ*-homogeneous, in the sense that* $\mathcal{G}(\delta_t(\xi)) = t\mathcal{G}(\xi)$ *whenever* $\xi \in \mathcal{X}$ *and* $t > 0$,

(HA.3) ν *is* δ*-homogeneous, in the sense that* $\nu(\delta_t(\xi)) = t\nu(\xi)$ *whenever* $\xi \in \mathcal{X}$ *and* $t > 0$,

(HA.4) ν *is a "pseudonorm on* $\mathcal{X}$*," that is, a continuous nonnegative real-valued function that satisfies the conditions (i)* $\nu(\xi) = 0 \Longleftrightarrow \xi = 0$ *and (ii)* $\lim_{\|\xi\| \to +\infty} \nu(\xi) = +\infty$.

So, when we rewrite (22.2) as (22.3), we find that we are really dealing with an *approximation of a map $\mathcal{F}$ by another map $\mathcal{G}$ which is homogeneous with respect to a continuous one-parameter group of dilations*. Furthermore, *the approximation is of exactly the same kind as the approximation by a linear map involved in the usual concept of differentiability, except that the ordinary norm on X is replaced by a dilation-homogeneous pseudonorm*.

Similar considerations apply to higher-order approximations. For example, if $F \in C^3$, and we choose three linear subspaces K_1, K_2, K_3 of X, then

$$\begin{aligned} F(x_* + h_1 + h_2 + h_3) = {} & F(x_*) + P_1(h_1) + P_1(h_2) + P_1(h_3) \\ & + \frac{1}{2}\Big(P_2(h_1,h_1) + P_2(h_2,h_2) + P_2(h_3,h_3)\Big) \\ & + P_2(h_1,h_2) + P_2(h_1,h_3) + P_2(h_2,h_3) + P_3(h_1,h_2,h_3) \\ & + \frac{1}{6}\Big(P_3(h_1,h_1,h_1) + P_3(h_2,h_2,h_2) + P_3(h_3,h_3,h_3)\Big) \\ & + \frac{1}{2}\Big(P_3(h_1,h_2,h_2) + P_3(h_1,h_1,h_2) + P_3(h_1,h_3,h_3) \\ & \qquad + P_3(h_1,h_1,h_3) + P_3(h_2,h_3,h_3) + P_3(h_2,h_2,h_3)\Big) \\ & + o(\|h_1\|^3 + \|h_2\|^3 + \|h_3\|^3) \\ & \text{as } h_1 \to 0,\ h_2 \to 0,\ h_3 \to 0,\ h_1 \in K_1,\ h_2 \in K_2,\ h_3 \in K_3\,. \end{aligned}$$

If we choose the K_i such that $K_1 = X$, and K_2 and K_3 are subsets of $\ker P_1$ (so that $P_1(h_2) = P_1(h_3) = 0$), and absorb into the remainder all the terms that are obviously $o(\|h_1\| + \|h_2\|^2 + \|h_3\|^3)$), we find

$$\begin{aligned} F(x_* + h_1 + h_2 + h_3) = {} & F(x_*) + P_1(h_1) + \frac{1}{2}\Big(P_2(h_2,h_2) + P_2(h_3,h_3)\Big) \\ & + P_2(h_2,h_3) + \frac{1}{6}P_3(h_3,h_3,h_3) + o(\|h_1\| + \|h_2\|^2 + \|h_3\|^3)\,. \end{aligned}$$

(The term $P_3(h_2,h_3,h_3)$ is eliminated because

$$\begin{aligned} \|P_3(h_2,h_3,h_3)\| &\le \|P_3\|\,\|h_2\|\,\|h_3\|^2 \le \|P_3\|(\|h_2\|^{7/3} + (\|h_3\|^2)^{7/4}) \\ &= \|P_3\|(\|h_2\|^{7/3} + \|h_3\|^{7/2}) = o(\|h_2\|^2 + \|h_3\|^3)\,, \end{aligned}$$

using the fact that the inequality $ab \le a^p + b^q$ holds if $a \ge 0$, $b \ge 0$, $p > 1$, $q > 1$, and $(1/p) + (1/q) = 1$, and taking $p = 7/3$, $q = 7/4$.)

If we require in addition that "$K_3 \subseteq \ker P_2$," in the precise sense that every $h \in K_3$ must satisfy $P_2(h,h') = 0$ whenever $h' \in K_2$, then we find

$$\begin{aligned} F(x_* + h_1 + h_2 + h_3) = {} & F(x_*) + P_1(h_1) + \frac{1}{2}P_2(h_2,h_2) \\ & + \frac{1}{6}P_3(h_3,h_3,h_3) + o(\|h_1\| + \|h_2\|^2 + \|h_3\|^3) \\ & \text{as } h_1 \to 0,\ h_2 \to 0,\ h_3 \to 0,\ h_1 \in K_1,\ h_2 \in K_2,\ h_3 \in K_3\,. \end{aligned}$$

So, once again, we find that (22.3) holds, if we define $\mathcal{X} = K_1 \times K_2 \times K_3$, and let

$$\begin{aligned} \mathcal{F}(h_1, h_2, h_3) &= F(x_* + h_1 + h_2 + h_3) - F(x_*)\,, \\ \mathcal{G}(h_1, h_2, h_3) &= P_1(h_1) + \frac{1}{2}P_2(h_2, h_2) + \frac{1}{6}P_3(h_3, h_3, h_3)\,, \\ \nu(h_1, h_2, h_3) &= \|h_1\| + \|h_2\|^2 + \|h_3\|^3\,, \end{aligned}$$

for $h_1 \in K_1$, $h_2 \in K_2$, $h_3 \in K_3$. In addition, if we let

$$\delta_t(h_1, h_2, h_3) = (th_1, t^{1/2}h_2, t^{1/3}h_3) \quad \text{for} \quad h_1 \in K_1,\, h_2 \in K_2,\, h_3 \in K_3,\, t > 0\,,$$

then the four homogeneous approximation conditions HA.1-4 hold.

Our higher-order generalization of Theorem 22.1 involves a condition closely resembling the usual formula $\mathcal{F}(\xi) = \mathcal{L}(\xi) + o(\|\xi\|)$ that characterizes ordinary differentiability of $\mathcal{F}$ at 0. The crucial difference is that (a) the linear map $\mathcal{L}$ is replaced by a map $\mathcal{G}$ which is homogeneous with respect to some suitable group of dilations, and (b) a pseudonorm ν is substituted for the ordinary norm. The requirement that $\mathcal{L}$ be surjective, which occurs in the first-order theorems, will be replaced by the condition that $\mathcal{G}$ have a surjective differential at some point $\xi_* \in \mathcal{X}$ such that $\mathcal{G}(\xi_*) = 0$. We will show that these conditions imply that $\mathcal{F}$ is open at 0 if the target space $\mathcal{Y}$ of $\mathcal{F}$ is finite-dimensional, generalizing Theorem 22.1. To get the generalization of Theorem 22.2, in which $\mathcal{Y}$ is allowed to be infinite-dimensional, we will have use the appropriate "strict" analogues of differentiability: the map $\mathcal{G}$ will have to be assumed to be strictly differentiable at ξ_*, and the approximation formula (22.3) will have to be replaced by

$$\mathcal{E}(\xi) - \mathcal{E}(\xi') = o(\nu(\xi, \xi')) \qquad \text{as } \xi \to 0\,,\ \xi' \to 0\,,\quad \xi \in \mathcal{X}\,,\ \xi' \in \mathcal{X}\,, \tag{22.4}$$

where the error $\mathcal{E}$ is defined by $\mathcal{E}(\xi) = \mathcal{F}(\xi) - \mathcal{G}(\xi)$, and ν is a "homogeneous Lipschitz-bounded pseudodistance," in a sense to be defined below.

From the general theorems involving homogeneous approximations, it will then be possible to derive high-order open mapping theorems by means of the construction sketched above, using the obvious fact that, when $\mathcal{F}$ is defined in terms of F as we have done in our second- and third-order examples, then if $\mathcal{F}$ is open at 0 it follows that F is open at x_*.

This will be done here for the second-order case in §22.5, where it will be shown that our sufficient conditions for openness apply in particular when the Hessian of the map has a regular zero, and in the cases considered by Avakov in [1, 2].

22.2 Preliminaries

Metric spaces, normed spaces, balls, openness at a point. If X is a metric space with distance d, x is a point of X, and $r \geq 0$, we will use $\mathbb{B}_X(x, r)$ to denote the closed r-ball with center x, i.e., the set $\{x' \in X : d(x', x) \leq r\}$.

We use the word "neighborhood" in the usual sense of point set topology: if X is a topological space, and $x_* \in X$, a *neighborhood* of x_* in X is a subset U of X such that x_* is an interior point of U.

DEFINITION 22.4
If X, Y are topological spaces, $x_* \in X$, and $F : X \mapsto Y$ is a map, then F is said to be *open at* x_* if whenever U is a neighborhood of x_* in X, it follows that the image $F(U)$ is a neighborhood of 0 in Y. □

All linear spaces will be over $\mathbb{R}$, the field of real numbers. If X is a normed linear space, then we will write $\mathbb{B}_X(r)$ instead of $\mathbb{B}_X(0,r)$. If X, Y are normed linear spaces, then $\mathrm{Lin}(X,Y)$ will denote the space of all bounded linear maps from X to Y. If $A \in \mathrm{Lin}(X,Y)$ and $x \in X$, then we will use interchangeably the expressions Ax, $A \cdot x$ and $A(x)$ to denote the value of A at x. Convergence in $\mathrm{Lin}(X,Y)$ is uniform convergence, i.e., convergence relative to the operator norm $\mathrm{Lin}(X,Y) \ni A \mapsto \|A\| \stackrel{\mathrm{def}}{=} \sup\{\|Ax\| : x \in X, \|x\| \le 1\}$.

Differentiability and strict differentiability. The word "differentiable" will only be used to refer to maps between normed spaces, and will mean "Fréchet differentiable." That is, if we assume that

(A) *X and Y are normed spaces, Ω is an open subset of X, $x_* \in \Omega$, $F : \Omega \mapsto Y$ is a map, and $L : X \mapsto Y$ is a bounded linear map,*

then we say that *F is differentiable at x_* with differential L* if

$$\lim_{x \to x_*, x \neq x_*} \frac{\|F(x) - F(x_*) - L.(x - x*)\|}{\|x - x_*\|} = 0\,. \tag{22.5}$$

A stronger concept of differentiability is "strict differentiability," defined as follows.

DEFINITION 22.5
Let X, Y, Ω, x_*, F, L be such that (A) above holds. We say that *F is strictly differentiable at x_* with strict differential L* if the equality

$$\lim_{x \to x_*, x' \to x_*, x \neq x'} \frac{\|F(x) - F(x') - L.(x - x')\|}{\|x - x'\|} = 0 \tag{22.6}$$

holds. □

Clearly, if F is strictly differentiable at x_* with strict differential L, then F is differentiable at x_* with differential L, since we can obtain (22.5) by taking $x' = x_*$ in (22.6).

Dilations, pseudonorms, pseudodistances, homogeneous maps.

DEFINITION 22.6
Assume that $\mathcal{X}$ is a normed real linear space. A *continuous one-parameter group of dilations* on $\mathcal{X}$ is a family $\boldsymbol{\delta} = \{\delta_t\}_{t>0}$ of bounded linear maps $\delta_t : \mathcal{X} \mapsto \mathcal{X}$ such that

D1. δ_1 is the identity map $\mathrm{id}_{\mathcal{X}}$ of $\mathcal{X}$,

D2. $\delta_t \circ \delta_s = \delta_{ts}$ whenever $t > 0$ and $s > 0$,

D3. the map $]0, +\infty[\ni t \mapsto \delta_t \in \mathrm{Lin}(\mathcal{X}, \mathcal{X})$ is continuous with respect to the uniform operator norm on $\mathrm{Lin}(\mathcal{X}, \mathcal{X})$ (that is, $\lim_{t \to s} \|\delta_t - \delta_s\| = 0$ whenever $s > 0$),

D4. $\lim_{t\downarrow 0} \|\delta_t\| = 0$.

□

LEMMA 22.7
If $\mathcal{X}$ is a normed linear space and $\boldsymbol{\delta} = \{\delta_t\}_{t>0}$ is a continuous one-parameter group of dilations on $\mathcal{X}$, then

$$\lim_{t\to+\infty} \|\delta_t(x)\| = +\infty \qquad \text{for all} \qquad x \in \mathcal{X}\backslash\{0\}\,, \tag{22.7}$$

and

$$\text{the map }]\,0, +\infty\,[\ni t \mapsto \delta_t(x) \in \mathcal{X} \text{ is one-to-one whenever } x \in X\backslash\{0\}\,. \tag{22.8}$$

□

Proof If $x \neq 0$ and it is not true that $\lim_{t\to+\infty} \|\delta_t(x)\| = +\infty$, then there exists a sequence $\{t_j\}_{j=1}^{\infty}$ such that $\lim_{j\to\infty} t_j = +\infty$ while, on the other hand, the set $S = \{\delta_{t_j}(x) : j \in \mathbb{N}\}$ is bounded. Then, if we let $\xi_j = \delta_{t_j}(x)$, we can pick a constant C such that $\|\xi_j\| \leq C$ for all j, and conclude that $x = \delta_{t_j^{-1}}(\xi_j)$, so $\|x\| \leq C\|\delta_{t_j^{-1}}\|$. Since $\|\delta_{t_j^{-1}}\| \to 0$—because $t_j^{-1} \to 0$—it follows that $x = 0$. This completes the proof of (22.7).

To prove (22.8), we observe that if $x \neq 0$, $\delta_t(x) = \delta_s(x)$, and $0 < t < s$, then $\delta_\tau(x) = x$, if $\tau = s/t$. Then $\tau > 1$, and $\delta_{\tau^k}(x) = x$ for $k = 1, 2, \ldots$, so the norm $\|\delta_{\tau^k}(x)\|$ does not go to $+\infty$ as $k \to \infty$ but $\tau^k \to +\infty$ as $k \to \infty$, contradicting (22.7). □

DEFINITION 22.8
Assume that $\mathcal{X}$, $\boldsymbol{\delta}$ are as in Definition 22.6. A *$\boldsymbol{\delta}$-pseudonorm* is a continuous map $\nu : \mathcal{X} \mapsto \mathbb{R}$ such that (1) $\nu(x) \geq 0$ for all $x \in \mathcal{X}$, (2) $\nu(x) = 0 \Longleftrightarrow x = 0$, (3) $\lim_{\|x\|\to+\infty} \nu(x) = +\infty$, and (4) $\nu(\delta_t(x)) = t\nu(x)$ whenever $x \in \mathcal{X}$, $t \in \mathbb{R}$, and $t > 0$. □

DEFINITION 22.9
Assume that the space $\mathcal{X}$ and the dilation group $\boldsymbol{\delta}$ are as in Definition 22.6. A *$\boldsymbol{\delta}$-pseudodistance* is a continuous map $\nu : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$ such that (1) $\nu(x, x') = \nu(x', x) \geq 0$ for all $x, x' \in \mathcal{X}$, (2) $\nu(x, x') = 0 \Longleftrightarrow x = x'$, (3) $\lim_{\|x\|\to+\infty} \nu(x, 0) = +\infty$, and (4) $\nu(\delta_t(x), \delta_t(x')) = t\nu(x, x')$ whenever $x, x' \in \mathcal{X}$, $t \in \mathbb{R}$, and $t > 0$. □

If $\nu : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a δ-pseudodistance on $\mathcal{X}$, we define, for each positive number R,

$$\kappa_\nu(R) \stackrel{\text{def}}{=} \sup\left\{ \frac{\nu(x, x')}{\|x - x'\|} : x, x' \in \mathbb{B}_{\mathcal{X}}(R)\,,\ x \neq x' \right\}, \tag{22.9}$$

Then

$$\kappa_\nu(R) < \infty \text{ for some positive } R \Longleftrightarrow \kappa_\nu(R) < \infty \text{ whenever } R > 0. \tag{22.10}$$

(Indeed, if R_0 is such that $R_0 > 0$ and $\kappa_\nu(R_0) < \infty$, then there exists a positive $\bar{t}$ such that $R_0\|\delta_{\bar{t}}\| < R$, since $\lim_{t\downarrow 0}\|\delta_t\| = 0$. Then $\delta_{\bar{t}}(\mathbb{B}_{\mathcal{X}}(R)) \subseteq \mathbb{B}_{\mathcal{X}}(R_0)$. If $x, x' \in \mathbb{B}_{\mathcal{X}}(R)$, then

$$\begin{aligned}\nu(x,x') &= \bar{t}^{-1}\nu(\delta_{\bar{t}}(x), \delta_{\bar{t}}(x')) \\ &\le \bar{t}^{-1}\kappa_\nu(R_0)\|\delta_{\bar{t}}(x) - \delta_{\bar{t}}(x')\| \\ &\le \bar{t}^{-1}\kappa_\nu(R_0)\|\delta_{\bar{t}}\|\,\|x - x'\|\,,\end{aligned}$$

so $\kappa_\nu(R) \le \bar{t}^{-1}\kappa_\nu(R_0)\|\delta_{\bar{t}}\| < +\infty$.)

DEFINITION 22.10
A pseudodistance $\nu : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is *Lipschitz-bounded* if $\kappa_\nu(R) < \infty$ for some (and hence every) positive number R. □

DEFINITION 22.11
Assume that $\mathcal{X}$, $\boldsymbol{\delta}$ are as in Definition 22.6, $\mathcal{Y}$ is a normed linear space, and $\mathcal{G} : \mathcal{X} \mapsto \mathcal{Y}$ is a map. We say that $\mathcal{G}$ is $\boldsymbol{\delta}$*-homogeneous* if $\mathcal{G}(\delta_t(x)) = t\,\mathcal{G}(x)$ whenever $x \in \mathcal{X}$ and $t \ge 0$. □

Regular zeros. If X, Y are normed spaces, Ω is open in X, and F is a map from Ω to Y, then a *regular zero* (resp. a *strictly regular zero*) of F is a point $x_* \in \Omega$ such that $F(x_*) = 0$, F is Fréchet differentiable (resp. strictly Fréchet differentiable) at x_*, and the differential $DF(x_*) : X \mapsto Y$ is surjective.

22.3 The Finite-Dimensional Theorem

THEOREM 22.12
Assume that $\mathcal{X}$ and $\mathcal{Y}$ are real linear normed spaces, $\mathcal{Y}$ is finite-dimensional, $\boldsymbol{\delta} = \{\delta_t\}_{t>0}$ is a continuous one-parameter group of dilations of $\mathcal{X}$, $\nu : \mathcal{X} \mapsto \mathbb{R}_+$ is a $\boldsymbol{\delta}$-pseudonorm, and $\mathcal{G} : \mathcal{X} \mapsto \mathcal{Y}$ is a continuous $\boldsymbol{\delta}$-homogeneous map. Let Ω be an open subset of $\mathcal{X}$ such that $0 \in \Omega$, and let $\mathcal{F} : \Omega \mapsto \mathcal{Y}$ be a continuous map such that $\mathcal{F}(0) = 0$ and

$$\lim_{\xi\to 0} \frac{\|\mathcal{F}(\xi) - \mathcal{G}(\xi)\|}{\nu(\xi)} = 0\,. \tag{22.11}$$

Assume that $\mathcal{G}$ has a regular zero. Then $\mathcal{F}$ is open at 0. □

Proof Let V be a neighborhood of 0 in $\mathcal{X}$ such that $V \subseteq \Omega$. Let R be such that $R > 0$ and $\mathbb{B}_{\mathcal{X}}(2R) \subseteq V$. Let ξ_* be a regular zero of $\mathcal{G}$. For $t > 0$, let $\xi_{*,t} = \delta_t(\xi_*)$. Let $\bar{t}$ be such that $\bar{t} > 0$ and the inequalities

$$\|\delta_t\| \le 1\,, \qquad \|\xi_{*,t}\| \le R\,,$$

hold whenever $0 < t \le \bar{t}$. Let $\mathcal{L} = D\mathcal{G}(\xi_*)$, so $\mathcal{L}$ is surjective. Let $\mathcal{M} : \mathcal{Y} \mapsto \mathcal{X}$ be a linear map such that $\mathcal{L} \circ \mathcal{M} = \mathrm{id}_{\mathcal{Y}}$. Then $\mathcal{M}$ is bounded, because $\mathcal{Y}$ is finite-dimensional. Let r be such that $0 < r < R$ and

$$\|\mathcal{G}(\xi) - \mathcal{G}(\xi_*) - \mathcal{L}(\xi - \xi_*)\| \le \frac{\|\xi - \xi_*\|}{4\|\mathcal{M}\|} \qquad \text{whenever} \qquad \|\xi - \xi_*\| \le r\,.$$

Let s be such that $s > 0$, $s\|\mathcal{M}\| \leq r$, and $\bar{\nu} < \infty$, where

$$\bar{\nu} = \sup \Big\{ \nu(\xi) : \xi \in \mathbb{B}_{\mathcal{X}}(\xi_*; s\|\mathcal{M}\|) \Big\} .$$

Let Ψ be the map from $\mathbb{B}_{\mathcal{Y}}(s)$ to $\mathcal{Y}$ given by $\Psi(y) = \mathcal{G}(\xi_* + \mathcal{M} \cdot y)$. Then, if $y \in \mathbb{B}_{\mathcal{Y}}(s)$, $\mathcal{M}(y)$ belongs to $\mathbb{B}_{\mathcal{X}}(r)$, so

$$\begin{aligned} \|\Psi(y) - y\| &= \|\mathcal{G}(\xi_* + \mathcal{M} \cdot y) - \mathcal{L}(\mathcal{M}(y))\| \\ &= \|\mathcal{G}(\xi_* + \mathcal{M} \cdot y) - \mathcal{G}(\xi_*) - \mathcal{L}(\mathcal{M}(y))\| \\ &\leq \frac{1}{4\|\mathcal{M}\|} \|\mathcal{M}\| \cdot \|y\| \\ &\leq \frac{s}{4} , \end{aligned}$$

using the fact that $\mathcal{G}(\xi_*) = 0$. Next, we define $W_t = \delta_t\Big(\mathbb{B}_{\mathcal{X}}(\xi_*, s\|\mathcal{M}\|)\Big)$, and observe that $W_t \subseteq \mathbb{B}_{\mathcal{X}}(\|\xi_{*,t}\| + s\|\delta_t\| \, \|\mathcal{M}\|)$. In particular,

$$\lim_{t \downarrow 0} \Big(\sup\{\|\xi\| : \xi \in W_t\} \Big) = 0 ,$$

$$0 < t \leq \bar{t} \implies W_t \subseteq \mathbb{B}_{\mathcal{X}}(2R) ,$$

and

$$\lim_{t \downarrow 0} \lambda_t = 0 ,$$

where

$$\lambda_t \stackrel{\text{def}}{=} \sup \Big\{ \frac{\|\mathcal{F}(\xi) - \mathcal{G}(\xi)\|}{\nu(\xi)} : \xi \in W_t \Big\} .$$

Define

$$\Phi_t(y) = t^{-1}(\mathcal{F} \circ \delta_t)(\xi_* + \mathcal{M} \cdot y) \qquad \text{for} \qquad y \in \mathbb{B}_{\mathcal{Y}}(s) , \quad 0 < t \leq \bar{t} .$$

(The definition is possible because, if $y \in \mathbb{B}_{\mathcal{Y}}(s)$ and $0 < t \leq \bar{t}$, then

$$\|\delta_t(\xi_* + \mathcal{M} \cdot y)\| = \|\xi_{*,t} + \delta_t(\mathcal{M} \cdot y)\| \leq \|\xi_{*,t}\| + s\|\delta_t\| \, \|\mathcal{M}\| \leq R + r < 2R ,$$

so $\delta_t(\xi_* + \mathcal{M} \cdot y) \in \mathbb{B}_{\mathcal{X}}(2R)$ and then $(\mathcal{F} \circ \delta_t)(\xi_* + \mathcal{M} \cdot y)$ is well defined.)

Then, if we let

$$\begin{aligned} \mu(t, y) &= \|(\mathcal{F} \circ \delta_t)(\xi_* + \mathcal{M} \cdot y) - (\mathcal{G} \circ \delta_t)(\xi_* + \mathcal{M} \cdot y)\| , \\ \tilde{\mu}(t, y) &= \frac{\mu(t, y)}{\nu(\delta_t(\xi_* + \mathcal{M} \cdot y))} , \\ \mu_t &= \sup \Big\{ \mu(t, y) : y \in \mathbb{B}_{\mathcal{Y}}(s) \Big\} , \end{aligned}$$

we have

$$\tilde{\mu}(t, y) \leq \sup \Big\{ \frac{\|\mathcal{F}(\xi) - \mathcal{G}(\xi)\|}{\nu(\xi)} : \xi \in W_t \Big\} = \lambda_t ,$$

so that

$$\begin{aligned}\mu_t &= \sup\left\{\tilde{\mu}(t,y)\cdot\nu(\delta_t(\xi_*+\mathcal{M}\cdot y)) : y\in\mathbb{B}_{\mathcal{Y}}(s)\right\}\\ &\le \lambda_t \sup\left\{\nu(\delta_t(\xi)) : \xi\in\mathbb{B}_{\mathcal{X}}(\xi_*; s\|\mathcal{M}\|)\right\}\\ &= \lambda_t \sup\left\{t\nu(\xi) : \xi\in\mathbb{B}_{\mathcal{X}}(\xi_*; s\|\mathcal{M}\|)\right\}\\ &= t\,\lambda_t \sup\left\{\nu(\xi) : \xi\in\mathbb{B}_{\mathcal{X}}(\xi_*; s\|\mathcal{M}\|)\right\}\\ &= t\,\lambda_t\,\bar{\nu}\,.\end{aligned}$$

Furthermore,

$$\begin{aligned}\|\Phi_t(y)-y\| &\le t^{-1}\mu(t,y)+\|t^{-1}\mathcal{G}(\delta_t(\xi_*+\mathcal{M}\cdot y))-y\|\\ &\le t^{-1}\mu(t,y)+\|\mathcal{G}(\xi_*+\mathcal{M}\cdot y)-y\|\\ &\le t^{-1}\mu_t+\|\Psi(y)-y\|\\ &\le \lambda_t\,\bar{\nu}+\frac{s}{4}\,.\end{aligned}$$

Now choose t such that $\bar{\nu}\lambda_t \le \frac{s}{4}$. Then $\|\Phi_t(y)-y\| \le \frac{s}{2}$. Let $B=\mathbb{B}_{\mathcal{Y}}(\frac{s}{2})$, $\tilde{B}=\mathbb{B}_{\mathcal{Y}}(s)$. If $y\in B$, define a map $\zeta_y : \tilde{B}\mapsto\mathcal{Y}$ by letting $\zeta_y(y')=y'-\Phi_t(y')+y$ if $y'\in\tilde{B}$. Then, if $y'\in\tilde{B}$, the inequalities

$$\|\zeta_y(y')\| \le \|y'-\Phi_t(y')\|+\|y\| \le \frac{s}{2}+\frac{s}{2}=s$$

imply that $\zeta_y(y')\in\tilde{B}$ as well. Hence ζ_y is a continuous map from $\tilde{B}$ to $\tilde{B}$, so ζ_y has a fixed point by Brouwer's theorem. If $\bar{y}'$ is a fixed point of ζ_y, then $\bar{y}'-\Phi_t(\bar{y}')+y=\zeta_y(\bar{y}')=\bar{y}'$, so $\Phi_t(\bar{y}')=y$. Then

$$t^{-1}\mathcal{F}\Big(\delta_t(\xi_*+\mathcal{M}\cdot\bar{y}')\Big)=y\,,$$

and $\delta_t(\xi_*+\mathcal{M}\cdot\bar{y}')\in W_t$. So, if we let $z=\delta_t(\xi_*+\mathcal{M}\cdot\bar{y}')$, it follows that $z\in V$ (because $W_t\subseteq\mathbb{B}_{\mathcal{X}}(2R)\subseteq V$) and $\mathcal{F}(z)=ty$. Since y is an arbitrary member of B, we have shown that

$$tB\subseteq\mathcal{F}(V)\,.$$

Therefore tB is a neighborhood of 0 in $\mathcal{Y}$, and $tB\subseteq\mathcal{F}(V)$. This completes our proof. □

22.4 The Infinite-Dimensional Theorem

THEOREM 22.13
Assume that $\mathcal{X}$ and $\mathcal{Y}$ are real Banach spaces, $\boldsymbol{\delta}=\{\delta_t\}_{t>0}$ is a continuous one-parameter group of dilations of $\mathcal{X}$, $\nu : \mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}_+$ is a Lipschitz-bounded $\boldsymbol{\delta}$-pseudodistance, and $\mathcal{G}:\mathcal{X}\mapsto\mathcal{Y}$ is a continuous $\boldsymbol{\delta}$-homogeneous map. Let Ω be

an open subset of $\mathcal{X}$ such that $0 \in \Omega$, and let $\mathcal{F} : \Omega \mapsto \mathcal{Y}$ be a continuous map such that $\mathcal{F}(0) = 0$ and

$$\lim_{\xi \to 0, \xi' \to 0, \xi \neq \xi'} \frac{\Big\| \Big(\mathcal{F}(\xi) - \mathcal{G}(\xi) \Big) - \Big(\mathcal{F}(\xi') - \mathcal{G}(\xi') \Big) \Big\|}{\nu(\xi, \xi')} = 0 \,. \tag{22.12}$$

Assume that $\mathcal{G}$ has a strictly regular zero. Then $\mathcal{F}$ is open at 0. □

Proof We define the error functions $\mathcal{E} : \Omega \mapsto \mathcal{Y}$, $\tilde{\mathcal{E}} : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{Y}$, by letting

$$\begin{aligned} \mathcal{E}(\xi) &= \mathcal{F}(\xi) - \mathcal{G}(\xi) \ \text{ for } \ \xi \in \Omega \,, \\ \tilde{\mathcal{E}}(\xi, \xi') &= \mathcal{G}(\xi) - \mathcal{G}(\xi') - \mathcal{L}(\xi - \xi') \ \text{ for } \ \xi, \xi' \in \mathcal{X} \,. \end{aligned}$$

Let V be a neighborhood of 0 in $\mathcal{X}$ such that $V \subseteq \Omega$. Let R be such that $R > 0$ and $\mathbb{B}_{\mathcal{X}}(2R) \subseteq V$. Let ξ_* be a strictly regular zero of $\mathcal{G}$. For $t > 0$, let $\xi_{*,t} = \delta_t(\xi_*)$. Let $\bar{t}$ be such that $\bar{t} > 0$ and the inequalities

$$\|\delta_t\| \leq 1 \,, \qquad \|\xi_{*,t}\| \leq R \,,$$

hold whenever $0 < t \leq \bar{t}$. Let $\mathcal{L} = D\mathcal{G}(\xi_*)$, so $\mathcal{L}$ is a bounded, surjective linear map from $\mathcal{X}$ to $\mathcal{Y}$. Then the Banach open mapping theorem implies that there is a positive constant C such that

(#) *for every $y \in \mathcal{X}$ there exists a $\xi \in \mathcal{X}$ such that $\mathcal{L}(\xi) = y$ and $\|\xi\| \leq C\|y\|$.*

Using the strict differentiability of $\mathcal{G}$ at ξ_*, we choose r such that $0 < r < R$ and

$$\|\tilde{\mathcal{E}}(\xi, \xi')\| \leq \frac{\|\xi - \xi'\|}{4C} \text{ whenever } \|\xi - \xi_*\| \leq r \text{ and } \|\xi' - \xi_*\| \leq r \,.$$

Let s be such that $s > 0$, $sC \leq r$, and $\bar{\nu} < \infty$, where

$$\bar{\nu} = \sup \Big\{ \nu(\xi, 0) : \xi \in \mathbb{B}_{\mathcal{X}}(\xi_*; sC) \Big\} \,.$$

Define

$$\begin{aligned} W_t \ &= \ \delta_t \Big(\mathbb{B}_{\mathcal{X}}(\xi_*, sC) \Big) \,, \\ \omega_t \ &= \ \sup \Big\{ \|\xi\| : \xi \in W_t \Big\} \,, \end{aligned}$$

and observe that $\omega_t \leq \|\xi_{*,t}\| + s\|\delta_t\| \, C$ for every t. In particular,

$$\lim_{t \downarrow 0} \omega_t = 0 \,,$$

$$0 < t \leq \bar{t} \implies W_t \subseteq \mathbb{B}_{\mathcal{X}}(\omega_t) \subseteq \mathbb{B}_{\mathcal{X}}(2R) \,,$$

and

$$\lim_{t \downarrow 0} \lambda_t = 0 \,,$$

where

$$\lambda_t \stackrel{\text{def}}{=} \sup\left\{ \frac{\left\|\mathcal{E}(\xi) - \mathcal{E}(\xi')\right\|}{\nu(\xi,\xi')} : \xi \in \mathbb{B}_{\mathcal{X}}(\omega_t),\, \xi' \in \mathbb{B}_{\mathcal{X}}(\omega_t),\, \xi \neq \xi' \right\}.$$

The Lipschitz-boundedness of ν implies that the constant $\bar{\kappa} \stackrel{\text{def}}{=} \kappa_\nu(R + \|\xi_*\|)$ is finite. For each t, we let

$$\alpha_t \stackrel{\text{def}}{=} \frac{1}{4} + C\lambda_t\bar{\kappa}\,.$$

We now fix a t such that

$$0 < t \leq \bar{t}\,, \qquad 4\lambda_t\bar{\nu} \leq s\,, \qquad \text{and} \qquad 2\alpha_t \leq 1\,, \tag{22.13}$$

write $\tau = 1/t$, and prove that the ball $\mathbb{B}_{\mathcal{Y}}(t\lambda_t\bar{\nu})$ is contained in $\mathcal{F}(W_t)$. For this purpose, we fix a $y \in \mathcal{Y}$ such that $\|y\| \leq t\lambda_t\bar{\nu}$, and construct a $\xi \in W_t$ such that $\mathcal{F}(\xi) = y$. This ξ will be the limit of a sequence $\{\xi^j\}_{j=0}^{\infty}$ of points of W_t.

To begin with, we let $\xi^0 = \xi_{*,t}$, and observe that $\xi^0 \in W_t$, and the error $e^0 \stackrel{\text{def}}{=} \mathcal{F}(\xi^0) - y$ satisfies

$$e^0 = \mathcal{F}(\xi_{*,t}) - \mathcal{G}(\xi_{*,t}) - (\mathcal{F}(0) - \mathcal{G}(0)) - y\,,$$

since $\mathcal{G}(\xi_{*,t}) = \mathcal{G}(0) = \mathcal{F}(0) = 0$. So

$$\begin{aligned}
\|e^0\| &\leq \|\mathcal{F}(\xi_{*,t}) - \mathcal{G}(\xi_{*,t}) - (\mathcal{F}(0) - \mathcal{G}(0))\| + \|y\| \\
&\leq \lambda_t\nu(\xi_{*,t}, 0) + \|y\| \\
&= \lambda_t\nu(\delta_t(\xi_*), \delta_t(0)) + \|y\| \\
&= t\lambda_t\nu(\xi_*, 0) + \|y\| \\
&\leq t\lambda_t\bar{\nu} + \|y\| \\
&= 2t\lambda_t\bar{\nu}\,.
\end{aligned}$$

We then choose $\zeta^1 \in \mathcal{X}$ such that $\mathcal{L}(\zeta^1) = -t^{-1}e^0$ and $\|\zeta^1\| \leq t^{-1}C\|e^0\|$, and define $\eta^1 = \delta_t(\zeta^1)$, $\xi^1 = \xi^0 + \eta^1 = \delta_t(\xi_* + \zeta^1)$. Then $t\mathcal{L}(\delta_\tau(\eta^1)) = -e^0$, and $\|\zeta^1\| \leq 2C\lambda_t\bar{\nu}$. Therefore $\|\delta_\tau(\xi^1) - \xi_*\| = \|\zeta^1\| \leq 2C\lambda_t\bar{\nu} \leq 4C\lambda_t\bar{\nu} \leq sC$, from which it follows that $\delta_\tau(\xi^1) \in \mathbb{B}_{\mathcal{X}}(\xi_*, sC)$, so that $\xi^1 \in W_t$.

We then let $e^1 = \mathcal{F}(\xi^1) - y$. It follows that

$$\begin{aligned}
e^1 &= \mathcal{E}(\xi^1) + \Big(\mathcal{G}(\xi^1) - y\Big) \\
&= \Big(\mathcal{E}(\xi^1) - \mathcal{E}(\xi^0)\Big) + \Big(\mathcal{F}(\xi^0) - \mathcal{G}(\xi^0)\Big) + \Big(\mathcal{G}(\xi^1) - y\Big) \\
&= \mathcal{E}(\xi^1) - \mathcal{E}(\xi^0) + \Big(\mathcal{F}(\xi^0) - y\Big) + \Big(\mathcal{G}(\xi^1) - \mathcal{G}(\xi^0)\Big) \\
&= \mathcal{E}(\xi^1) - \mathcal{E}(\xi^0) + e^0 + t\Big(\mathcal{G}(\delta_\tau(\xi^1)) - \mathcal{G}(\delta_\tau(\xi^0))\Big) \\
&= \mathcal{E}(\xi^1) - \mathcal{E}(\xi^0) + e^0 + t\tilde{\mathcal{E}}(\delta_\tau(\xi^1), \delta_\tau(\xi^0)) + t\mathcal{L}(\delta_\tau(\eta^1)) \\
&= \mathcal{E}(\xi^1) - \mathcal{E}(\xi^0) + t\tilde{\mathcal{E}}(\delta_\tau(\xi^1), \delta_\tau(\xi^0))\,.
\end{aligned}$$

On the other hand,

$$\begin{aligned}\|\mathcal{E}(\xi^1) - \mathcal{E}(\xi^0)\| &\leq \lambda_t \nu(\xi^1, \xi^0)\\ &= t\lambda_t \nu(\delta_\tau(\xi^1), \delta_\tau(\xi^0))\\ &\leq t\lambda_t \bar\kappa \|\delta_\tau(\xi^1) - \delta_\tau(\xi^0))\|\\ &= t\lambda_t \bar\kappa \|\delta_\tau(\eta^1)\|\\ &= t\lambda_t \bar\kappa \|\zeta^1\|\\ &\leq 2Ct\lambda_t^2 \bar\kappa \bar\nu\,,\end{aligned}$$

and

$$\|\tilde{\mathcal{E}}(\delta_\tau(\xi^1), \delta_\tau(\xi^0))\| \leq \frac{\|\delta_\tau(\xi^1) - \delta_\tau(\xi^0)\|}{4C} = \frac{\|\zeta^1\|}{4C} \leq \frac{2C\lambda_t\bar\nu}{4C} = \frac{1}{2}\lambda_t\bar\nu\,.$$

Therefore

$$\|e^1\| \leq 2Ct\lambda_t^2\bar\kappa\bar\nu + \frac{t}{2}\lambda_t\bar\nu = t\lambda_t\bar\nu\Big(2C\lambda_t\bar\kappa + \frac{1}{2}\Big) = 2t\lambda_t\bar\nu\alpha_t\,.$$

Next, we choose $\zeta^2 \in \mathcal{X}$ such that $\mathcal{L}(\zeta^2) = -t^{-1}e^1$ and $\|\zeta^2\| \leq t^{-1}C\|e^1\|$, and define $\eta^2 = \delta_t(\zeta^2)$, $\xi^2 = \xi^1 + \eta^2 = \xi^1 + \delta_t(\zeta^2) = \delta_t(\xi_* + \zeta^1 + \zeta^2)$. Then $t\mathcal{L}(\delta_\tau(\eta^2)) = -e^1$, and $\|\zeta^2\| \leq 2C\lambda_t\bar\nu\alpha_t$. Therefore

$$\|\delta_\tau(\xi^2) - \xi_*\| = \|\zeta^1 + \zeta^2\| \leq 2C\lambda_t\bar\nu(1+\alpha_t) \leq 4C\lambda_t\bar\nu \leq sC\,,$$

from which it follows that $\delta_\tau(\xi^2) \in \mathbb{B}_{\mathcal{X}}(\xi_*, sC)$, so that $\xi^2 \in W_t$.

We then let $e^2 = \mathcal{F}(\xi^2) - y$. It follows that

$$\begin{aligned}e^2 &= \mathcal{E}(\xi^2) + \Big(\mathcal{G}(\xi^2) - y\Big)\\ &= \Big(\mathcal{E}(\xi^2) - \mathcal{E}(\xi^1)\Big) + \Big(\mathcal{F}(\xi^1) - \mathcal{G}(\xi^1)\Big) + \Big(\mathcal{G}(\xi^2) - y\Big)\\ &= \mathcal{E}(\xi^2) - \mathcal{E}(\xi^1) + \Big(\mathcal{F}(\xi^1) - y\Big) + \Big(\mathcal{G}(\xi^2) - \mathcal{G}(\xi^1)\Big)\\ &= \mathcal{E}(\xi^2) - \mathcal{E}(\xi^1) + e^1 + t\Big(\mathcal{G}(\delta_\tau(\xi^2)) - \mathcal{G}(\delta_\tau(\xi^1))\Big)\\ &= \mathcal{E}(\xi^2) - \mathcal{E}(\xi^1) + e^1 + t\tilde{\mathcal{E}}(\delta_\tau(\xi^2), \delta_\tau(\xi^1)) + t\mathcal{L}(\delta_\tau(\eta^2))\\ &= \mathcal{E}(\xi^2) - \mathcal{E}(\xi^1) + t\tilde{\mathcal{E}}(\delta_\tau(\xi^2), \delta_\tau(\xi^1))\,.\end{aligned}$$

On the other hand,

$$\begin{aligned}\|\mathcal{E}(\xi^2) - \mathcal{E}(\xi^1)\| &\leq \lambda_t \nu(\xi^2, \xi^1)\\ &= t\lambda_t \nu(\delta_\tau(\xi^2), \delta_\tau(\xi^1))\\ &\leq t\lambda_t \bar\kappa \|\delta_\tau(\xi^2) - \delta_\tau(\xi^1))\|\\ &= t\lambda_t \bar\kappa \|\delta_\tau(\eta^2)\|\\ &= t\lambda_t \bar\kappa \|\zeta^2\|\\ &\leq t\lambda_t \bar\kappa (2C\lambda_t\bar\nu\alpha_t)\\ &\leq 2Ct\lambda_t^2 \bar\kappa\bar\nu\alpha_t\,,\end{aligned}$$

and

$$\|\mathcal{E}(\delta_\tau(\xi^2),\delta_\tau(\xi^1))\| \le \frac{\|\delta_\tau(\xi^2)-\delta_\tau(\xi^1)\|}{4C} = \frac{\|\zeta^2\|}{4C} \le \frac{2C\lambda_t\bar\nu\alpha_t}{4C} = \frac{1}{2}\lambda_t\bar\nu\alpha_t\,.$$

Therefore

$$\|e^2\| \le 2Ct\lambda_t^2\bar\kappa\bar\nu\alpha_t + \frac{t}{2}\lambda_t\bar\nu\alpha_t = 2t\lambda_t\bar\nu\Big(C\lambda_t\bar\kappa+\frac{1}{4}\Big)\alpha_t = 2t\lambda_t\bar\nu\alpha_t^2\,.$$

We continue this construction inductively. Suppose we have defined

$$\xi^0,\dots,\xi^k \in W_t\,,\quad e^0,\dots,e^k\in\mathcal{Y}\,,\quad \zeta^1,\dots,\zeta^k\in\mathcal{X}\,,\quad \eta^1,\dots,\eta^k\in\mathcal{Y}\,,$$

such that

$$\begin{aligned}
\xi^j &= \xi^{j-1}+\eta^j && \text{for } j=1,\dots,k\,,\\
e^j &= \mathcal{F}(\xi^j)-y && \text{for } j=0,\dots,k\,,\\
t\mathcal{L}(\zeta^j) &= -e^{j-1} && \text{for } j=1,\dots,k\,,\\
\eta^j &= \delta_t(\zeta_j) && \text{for } j=1,\dots,k\,,\\
\|\zeta^j\| &\le 2C\lambda_t\bar\nu\alpha_t^{j-1} && \text{for } j=1,\dots,k\,,\\
\|e^j\| &\le 2t\lambda_t\bar\nu\alpha_t^j && \text{for } j=0,\dots,k\,.
\end{aligned}$$

We then choose $\zeta^{k+1}\in\mathcal{X}$ such that

$$\mathcal{L}(\zeta^{k+1}) = -t^{-1}e^k \;\text{ and }\; \|\zeta^{k+1}\| \le t^{-1}C\|e^k\|\,,$$

and define

$$\begin{aligned}
\eta^{k+1} &= \delta_t(\zeta^{k+1})\,,\\
\xi^{k+1} &= \xi^k+\eta^{k+1} = \xi^k+\delta_t(\zeta^{k+1}) = \delta_t(\xi_*+\zeta^1+\ldots+\zeta^{k+1})\,.
\end{aligned}$$

Then $t\mathcal{L}(\delta_\tau(\eta^{k+1}))=-e^k$, and

$$\|\zeta^{k+1}\| \le t^{-1}C\|e^k\| \le 2C\lambda_t\bar\nu\alpha_t^k\,.$$

Therefore

$$\|\delta_\tau(\xi^{k+1})-\xi_*\| = \|\zeta^1+\ldots+\zeta^{k+1}\| \le 2C\lambda_t\bar\nu\sum_{j=0}^{k}\alpha_t^j = \frac{2C\lambda_t\bar\nu}{1-\alpha_t} \le 4C\lambda_t\bar\nu\,,$$

so $\|\delta_\tau(\xi^{k+1})-\xi_*\|\le sC$, from which it follows that $\delta_\tau(\xi^{k+1})\in\mathbb{B}_{\mathcal{X}}(\xi_*,sC)$, so that $\xi^{k+1}\in W_t$.

We then let $e^{k+1} = \mathcal{F}(\xi^{k+1}) - y$. It follows that

$$\begin{aligned}
e^{k+1} &= \mathcal{E}(\xi^{k+1}) + \Big(\mathcal{G}(\xi^{k+1}) - y\Big) \\
&= \Big(\mathcal{E}(\xi^{k+1}) - \mathcal{E}(\xi^k)\Big) + \Big(\mathcal{F}(\xi^k) - \mathcal{G}(\xi^k)\Big) + \Big(\mathcal{G}(\xi^{k+1}) - y\Big) \\
&= \mathcal{E}(\xi^{k+1}) - \mathcal{E}(\xi^k) + \Big(\mathcal{F}(\xi^k) - y\Big) + \Big(\mathcal{G}(\xi^{k+1}) - \mathcal{G}(\xi^k)\Big) \\
&= \mathcal{E}(\xi^{k+1}) - \mathcal{E}(\xi^k) + e^k + t\Big(\mathcal{G}(\delta_\tau(\xi^{k+1})) - \mathcal{G}(\delta_\tau(\xi^k))\Big) \\
&= \mathcal{E}(\xi^{k+1}) - \mathcal{E}(\xi^k) + e^1 + t\tilde{\mathcal{E}}(\delta_\tau(\xi^{k+1}), \delta_\tau(\xi^k)) + t\mathcal{L}(\delta_\tau(\eta^{k+1})) \\
&= \mathcal{E}(\xi^{k+1}) - \mathcal{E}(\xi^k) + t\tilde{\mathcal{E}}(\delta_\tau(\xi^{k+1}), \delta_\tau(\xi^k))\,.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|\mathcal{E}(\xi^{k+1}) - \mathcal{E}(\xi^k)\| &\le \lambda_t \nu(\xi^{k+1}, \xi^k) \\
&= t\lambda_t \nu(\delta_\tau(\xi^{k+1}), \delta_\tau(\xi^k)) \\
&\le t\lambda_t \bar{\kappa} \|\delta_\tau(\xi^{k+1}) - \delta_\tau(\xi^k))\| \\
&= t\lambda_t \bar{\kappa} \|\delta_\tau(\eta^{k+1})\| \\
&= t\lambda_t \bar{\kappa} \|\zeta^{k+1}\| \\
&\le t\lambda_t \bar{\kappa} (2C\lambda_t \bar{\nu}\alpha_t) \\
&\le 2Ct\lambda_t^2 \bar{\kappa}\bar{\nu}\alpha_t^k\,,
\end{aligned}$$

and

$$\|\tilde{\mathcal{E}}(\delta_\tau(\xi^{k+1}), \delta_\tau(\xi^k))\| \le \frac{\|\delta_\tau(\xi^{k+1}) - \delta_\tau(\xi^k)\|}{4C} = \frac{\|\zeta^{k+1}\|}{4C} \le \frac{1}{2}\lambda_t \bar{\nu}\alpha_t^k\,.$$

Therefore

$$\|e^{k+1}\| \le 2Ct\lambda_t^2 \bar{\kappa}\bar{\nu}\alpha_t^k + \frac{t}{2}\lambda_t \bar{\nu}\alpha_t^k = 2t\lambda_t \bar{\nu}\Big(C\lambda_t \bar{\kappa} + \frac{1}{4}\Big)\alpha_t^k = 2t\lambda_t \bar{\nu}\alpha_t^{k+1}\,,$$

and our inductive construction is complete.

The bound $\|\zeta^j\| \le 2C\lambda_t \bar{\nu}\alpha_t^{j-1}$ implies—since $2\alpha_t \le 1$—that the series $\sum_{j=1}^\infty \zeta^j$ converges, and the sum ζ of the series satisfies $\|\zeta\| \le 4C\lambda_t \bar{\nu} \le sC$. Therefore the limit $\Xi = \lim_{j\to\infty} \delta_\tau(\xi^j)$ exists and satisfies $\Xi \in \mathbb{B}_X(\xi_*, sC)$. So the limit $\xi = \delta_t(\Xi) = \lim_{j\to\infty} \xi^j$ exists and belongs to W_t. Furthermore, $\mathcal{F}(\xi) - y = \lim_{j\to\infty}(\mathcal{F}(\xi^j) - y) = \lim_{j\to\infty} e^j = 0$, because of the bound $\|e^j\| \le 2t\lambda_t \bar{\nu}\alpha_t^j$. Therefore $\mathcal{F}(\xi) = y$, and our proof is complete. □

22.5 Second-Order Open Mapping Theorems

If X and Y are real linear spaces, and $X \times X \ni (x, x') \mapsto B(x, x') \in Y$ is a symmetric bilinear map, we write Q_B to denote the quadratic map associated with B, i.e., the map $X \ni x \mapsto B(x, x) \stackrel{\text{def}}{=} Q_B(x) \in Y$. It is well known that B is completely determined by Q_B, since

$$B(x, y) = \frac{1}{4}\Big(Q_B(x + y) - Q_B(x - y)\Big)\,. \tag{22.14}$$

A *quadratic map* from X to Y is a map Q such that $Q = Q_B$ for some (unique) bilinear symmetric map $B : X \times X \mapsto Y$. If Q is a quadratic map, then we will use B^Q to denote the corresponding symmetric bilinear map.

If X, Y are normed, then a bilinear map $B : X \times X \mapsto Y$ is continuous if and only if it is bounded, in the sense that there exists a constant C such that $\|B(x,x')\| \leq C\|x\|\,\|x'\|$ for all $x, x' \in X$. It follows from (22.14) that a quadratic map $Q : X \to Y$ is continuous if and only if the bilinear map B^Q is continuous.

DEFINITION 22.14
If X, Y are normed spaces, Ω is open in X, $x_* \in \Omega$, and F is a map from Ω to Y, then a *linear-quadratic approximation* (abbreviated LQA) of F at x_* is a triple $A = (L, K, Q)$ such that

LQA1. L is a bounded linear map from X to Y,

LQA2. K is a closed linear subspace of X,

LQA3. Q is a continuous quadratic map from K to Y.

LQA4. $F(x_* + x + k) = F(x_*) + Lx + \frac{1}{2}Q(k) + o(\|k\|^2 + \|x\|)$ as (x,k) goes to zero via values in $X \times K$.

□

If (L, K, Q) is a LQA of F at x_*, then it is clear that F is Fréchet differentiable at x_* and $DF(x_*) = L$, $K \subseteq \ker L$, and the quadratic map Q is completely determined by K, by means of the formula.

$$Q(k) = 2 \lim_{t \downarrow 0} t^{-2}(F(x_* + tk) - F(x_*)) \qquad \text{for} \qquad x \in K\,.$$

The error bound of LQA4 is important because it gives an estimate of the the error in terms of the function $X \times K \ni (x,k) \mapsto \|x\| + \|k\|^2$, which is positively homogeneous of degree 1 relative to an appropriate group of dilations on $X \times K$. We make this observation precise in the following statement, whose proof is trivial.

LEMMA 22.15
Assume that (a) X, Y, Ω, x_*, F are as in Definition 22.14, and (b) $A = (L, K, Q)$ is a linear-quadratic approximation of F at x_*. Let

$$\mathcal{X} \stackrel{\text{def}}{=} X \times K\,, \qquad \tilde{\Omega} = \Big\{\, (x,k) \in \mathcal{X} : x_* + x + k \in \Omega \,\Big\}\,,$$

and define maps $\nu : \mathcal{X} \mapsto \mathbb{R}$, $\delta_t : \mathcal{X} \mapsto \mathcal{X}$ for $t > 0$, $\mathcal{F} : \tilde{\Omega} \mapsto Y$, $\mathcal{G} : \mathcal{X} \mapsto \mathcal{Y}$, $\mathcal{E} : \tilde{\Omega} \mapsto Y$, by

$$\begin{aligned}
\nu(x,k) &= \|x\| + \|k\|^2 && \text{for} \quad (x,k) \in \mathcal{X}\,,\\
\delta_t(x,k) &= (tx, \sqrt{t}k) && \text{for} \quad (x,k) \in \mathcal{X}\,,\ t > 0\,,\\
\mathcal{F}(x,k) &= F(x_* + x + k) - F(x_*) && \text{for} \quad (x,k) \in \tilde{\Omega}\,,\\
\mathcal{G}_A(x,k) &= Lx + \frac{1}{2}Q(k) && \text{for} \quad (x,k) \in \mathcal{X}\,,\\
\mathcal{E}(x,k) &= F(x_* + x + k) - F(x_*) - \mathcal{G}_A(x,k)\\
&= \mathcal{F}(x,k) - \mathcal{G}_A(x,k) && \text{for} \quad (x,k) \in \tilde{\Omega}\,.
\end{aligned}$$

Then $\boldsymbol{\delta} = \{\delta_t\}_{t>0}$ is a continuous one-parameter group of dilations of $\mathcal{X}$, $\mathcal{G}_A$ is $\boldsymbol{\delta}$-homogeneous, ν is a $\boldsymbol{\delta}$-pseudonorm, and the error bound

$$\lim_{x\to 0,\, k\to 0} \frac{\|\mathcal{E}(x,k)\|}{\nu(x,k)} = 0 \tag{22.15}$$

is satisfied. □

DEFINITION 22.16
Assume that X, Y, Ω, x_*, F are as in Definition 22.14. A *strict linear-quadratic approximation* (abbreviated SLQA) of F at x_* is a triple $A = (L, K, Q)$ that satisfies conditions LQA1,2,3 of Definition 22.14 and is such that

(SLQA) $\Big(F(x_* + x + k) - Lx - \frac{1}{2}Q(k)\Big) - \Big(F(x_* + x' + k') - Lx' - \frac{1}{2}Q(k')\Big)$
$= o\Big(\|x - x'\| + (\sqrt{\|x\| + \|x'\|} + \|k\| + \|k'\|).\|k - k'\|\Big)$ as (x, k, x', k') goes to zero via values in $X \times K \times X \times K$.

□

If $A = (L, K, Q)$ is a SLQA of F at x_*, then A is a LQA of F at x_*, F is strictly Fréchet differentiable at x_*, and $DF(x_*) = L$.

As in the case of (non-strict) LQAs, the error bound of condition (SLQA) is important because it estimates the error in terms of a function which is positively homogeneous of degree 1 relative to an appropriate group of dilations. We make this precise in the following statement, whose proof is trivial.

LEMMA 22.17
Let X, Y, Ω, x_*, F, A, L, K, Q, $\mathcal{X}$, $\tilde{\Omega}$, $\mathcal{F}$, $\mathcal{G}_A$, $\boldsymbol{\delta}$, $\mathcal{E}$ be as in the statement of Lemma 22.15. Assume that A is a strict linear-quadratic approximation of F at x_*. Define a map $\nu : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ by letting

$$\nu(x, k, x', k') = \|x - x'\| + \Big(\sqrt{\|x\| + \|x'\|} + \|k\| + \|k'\|\Big).\|k - k'\| \tag{22.16}$$

for $(x, k, x', k') \in \mathcal{X} \times \mathcal{X}$. Then

1. $\boldsymbol{\delta}$ is a continuous one-parameter group of dilations of $\mathcal{X}$,
2. $\mathcal{G}_A$ is δ-homogeneous,
3. ν is a Lipschitz-bounded $\boldsymbol{\delta}$-pseudodistance,
4. the error bound

$$\lim_{x\to 0, x'\to 0, x'\neq x} \frac{\|\mathcal{E}(x,k) - \mathcal{E}(x',k')\|}{\nu(x,k,x',k')} = 0 \tag{22.17}$$

is satisfied.

□

An important class of maps that necessarily admit strict linear-quadratic approximations consists of the *maps of class C^1 with a differentiable derivative.* Precisely, let us assume that

(#) *X, Y are normed spaces, Ω is an open subset of X, $F : \Omega \mapsto Y$ is a map of class C^1, $x_* \in \Omega$, and the map $\Omega \ni x \mapsto DF(x) \in \mathrm{Lin}(X, Y)$ is differentiable at x_*.*

Then DF is a continuous map from Ω to the space $\mathrm{Lin}(X, Y)$ of bounded linear maps from X to Y, and the second derivative $D(DF)(x_*) = D^2F(x_*)$ is a bounded linear map from X to $\mathrm{Lin}(X, \mathrm{Lin}(X, Y))$. Furthermore, *this map is symmetric*, i.e., $D^2F(x_*)(x) \cdot x' = D^2F(x_*)(x') \cdot x$ for all $x, x' \in X$. (This is true because of the identity

$$D^2F(x_*)(x) \cdot x' = \lim_{\alpha \downarrow 0} \alpha^{-2} \Big(F(x_* + \alpha x + \alpha x') - F(x_* + \alpha x) - F(x_* + \alpha x') + F(x_*) \Big) ,$$

whose right-hand side is clearly symmetric under the interchange of x and x'.)

It follows that we can regard $D^2F(x_*)$ as a bounded symmetric bilinear map $B : X \times X \mapsto Y$, given by $B(x, x') = D^2F(x_*)(x) \cdot x'$. Then $B = B^Q$, where $Q : X \mapsto Y$ is the quadratic map given by $Q(x) = B(x, x)$, so that $Q = Q_B$ and $B = B^Q$. Even more important for us will be the restriction $\mathcal{Q}$ of Q to the kernel $K = \ker L$, where $L = DF(x_*)$.

It turns out that the triple $A = (L, K, \mathcal{Q})$ is a strict linear-quadratic approximation of F at x_*, as we now show.

First, we write

$$\begin{aligned} M(x) &= L \cdot x + \frac{1}{2} Q(x) , \\ \mathcal{G}_A(x, k) &= L \cdot x + \frac{1}{2} \mathcal{Q}(k) , \\ E(x) &= F(x_* + x) - F(x_*) - M(x) , \\ \mathcal{E}(x, k) &= F(x_* + x + k) - F(x_*) - \mathcal{G}_A(x, k) , \end{aligned}$$

for $x \in X$, $k \in K$.

LEMMA 22.18
Let X, Y, Ω, F, L, K, Q, $\mathcal{Q}$, M, $\mathcal{G}_A$, A, E, $\mathcal{E}$, be as above. Then

$$\lim_{x \to 0, x' \to 0, x' \neq x} \frac{\|E(x) - E(x')\|}{(\|x\| + \|x'\|)\|x - x'\|} = 0 , \tag{22.18}$$

and A is a strict linear-quadratic approximation of F at x_*. □

Proof Without loss of generality, we assume that $x_* = 0$ and $F(x_*) = 0$. We fix a positive R such that $\mathbb{B}_X(R) \subseteq \Omega$. For $0 < r \leq R$, let

$$\theta(r) = \sup \left\{ \frac{\|DF(x) - \mathcal{L} - B^Q(x, \cdot)\|}{\|x\|} : 0 < \|x\| \leq r \right\} .$$

Then θ is monotonically nondecreasing, and the differentiability assumption implies that $\lim_{r\downarrow 0}\theta(r) = 0$. Clearly, the bound

$$\|DF(\xi).v - \mathcal{L}.v - B^Q(\xi, v)\| \leq \theta(r)\,\|\xi\|\,\|v\| \tag{22.19}$$

is satisfied whenever $0 < r \leq R$, $\xi \in \mathbb{B}_X(r)$, and $v \in X$.

Then, if $x, x' \in \mathbb{B}_X(R)$, and we write $v = x - x'$ $\xi_s = x' + sv$, we have

$$\begin{aligned}
E(x) - E(x') &= F(x) - F(x') - \mathcal{L}\cdot v - \frac{1}{2}\Big(Q(x) - Q(x')\Big)\\
&= \Big(\int_0^1 DF(\xi_s)\,ds\Big)\cdot v - \mathcal{L}\cdot v - \frac{1}{2}\Big(Q(x) - Q(x')\Big)\\
&= \Big(\int_0^1 (DF(\xi_s) - \mathcal{L})\,ds\Big)\cdot v - \frac{1}{2}\Big(Q(x) - Q(x')\Big)\\
&= \int_0^1 \Big((DF(\xi_s) - \mathcal{L}).v - B^Q(\xi_s, v)\Big)\,ds\\
&\quad + \int_0^1 B^Q(\xi_s, v)\,ds - \frac{1}{2}\Big(B^Q(x,x) - B^Q(x',x')\Big)\\
&= \int_0^1 \Big((DF(\xi_s) - \mathcal{L}).v - B^Q(\xi_s, v)\Big)\,ds\\
&\quad + B^Q(x', v) + \frac{1}{2}B^Q(v,v) - \frac{1}{2}\Big(B^Q(x,x) - B^Q(x',x')\Big)\\
&= \int_0^1 \Big((DF(\xi_s) - \mathcal{L}).v - B^Q(\xi_s, v)\Big)\,ds\,,
\end{aligned}$$

using the identities $B^Q(x', v) + \frac{1}{2}B^Q(v,v) - \frac{1}{2}\Big(B^Q(x,x) - B^Q(x',x')\Big) = 0$ and $\int_0^1 B^Q(\xi_s, v)\,ds = B^Q(x', v) + \frac{1}{2}B^Q(v,v)$. Then (22.19), with $\xi = \xi_s$, yields

$$\|(DF(\xi_s) - \mathcal{L}).v - B^Q(\xi_s, v)\| \leq \theta(\|x\| + \|x'\|)\cdot(\|x\| + \|x'\|)\cdot\|x - x'\|\,,$$

since $\|\xi_s\| \leq \|x\| + \|x'\|$ whenever $0 \leq s \leq 1$. Integrating this inequality, we get the bound

$$\|E(x) - E(x')\| \leq \theta(\|x\| + \|x'\|)\cdot(\|x\| + \|x'\|)\cdot\|x - x'\|\,, \tag{22.20}$$

and (22.18) follows.

To prove that A is a strict linear-quadratic approximation of F at 0, we have to show that

$$\mathcal{E}(x,k) - \mathcal{E}(x',k') = o\Big(\|x - x'\| + (\sqrt{\|x\| + \|x'\|} + \|k\| + \|k'\|).\|k - k'\|\Big)$$

as $(x, x', k, k') \to (0, 0, 0, 0)$. But

$$\begin{aligned}
\mathcal{E}(x,k) - \mathcal{E}(x',k') &= (F(x+k) - \mathcal{G}_A(x,k)) - (F(x'+k') - \mathcal{G}_A(x',k')) \\
&= \Big(F(x+k) - L\cdot x - \frac{1}{2}Q(k)\Big) - \Big(F(x'+k') - L\cdot x' - \frac{1}{2}Q(k')\Big) \\
&= \Big(F(x+k) - L\cdot(x+k) - \frac{1}{2}Q(x+k)\Big) - \Big(F(x'+k') - L\cdot(x'+k') \\
&\quad -\frac{1}{2}Q(x'+k')\Big) + \frac{1}{2}\Big(Q(x+k) - Q(x'+k') - Q(k) + Q(k')\Big) \\
&= E(x+k) - E(x'+k') + \frac{1}{2}\Big(Q(x+k) - Q(x'+k') - Q(k) + Q(k')\Big) \\
&= E(x+k) - E(x'+k') + \frac{1}{2}\Big(Q(x) - Q(x')\Big) + B^Q(x,k) + B^Q(x',k') \\
&= E(x+k) - E(x'+k') \\
&\quad +\frac{1}{2}Q(x-x') + B^Q(x', x'-x) + B^Q(x-x', k) + B^Q(x', k-k')\,.
\end{aligned}$$

Therefore (22.20) implies, if we write $\kappa = \|B^Q\|$, $\Theta(s) = s\theta(s)$, that

$$\begin{aligned}
\|\mathcal{E}(x,k) - \mathcal{E}(x',k')\| &\leq \Theta\Big(\|x\| + \|x'\| + \|k\| + \|k'\|)\Big)\cdot\Big(\|x-x'\| + \|k-k'\|\Big) \\
&\quad +\frac{\kappa}{2}\|x-x'\|^2 + \kappa\|x'\|.\|x-x'\| + \kappa\|x-x'\|.\|k\| + \kappa\|x'\|.\|k-k'\|\,,
\end{aligned}$$

which is clearly $o\Big(\|x-x'\| + (\sqrt{\|x\| + \|x'\|} + \|k\| + \|k'\|).\|k-k'\|\Big)$. □

We can then apply Theorems 22.12 and 22.13 and obtain a number of open mapping theorems. The crucial condition in all of them is, of course, the existence of a regular zero (ξ_*, k_*) of the map $\mathcal{G}_A$ associated to the LQA $A = (L, K, Q)$. (Such a point will then automatically be a strictly regular zero, because $\mathcal{G}_A$ is a polynomial map.) It turns out that this condition is equivalent to a statement in terms of the "Hessian" H_A of A, as we now explain.

Let Z be the quotient space Y/LX, and let π be the canonical projection from Y onto Z. The *Hessian* H_A is the quadratic map

$$K \ni k \mapsto \pi(Q(k,k)) \stackrel{\text{def}}{=} H_A(k) \in Z\,.$$

DEFINITION 22.19
A *regular zero* of H_A is a $k_* \in K$ such that $H_A(k_*, k_*) = 0$ and the map $K \ni k \mapsto \pi(B^Q(k_*, k)) \in Z$ is surjective. □

REMARK 22.20
If LX is a closed subspace of Y then Z is a normed space and the quadratic map $H_A : K \mapsto Z$ is continuous, and hence smooth. In that case, k_* is a regular zero of H_A in the sense defined above if and only if it is a regular zero in the sense defined earlier in §22.2. Here, however, *we are not requiring LX to be closed*, and the concept of a "regular zero" of H_A has to be understood in the purely algebraic sense of Definition 22.19. □

LEMMA 22.21
If $A = (L, K, Q)$ is a LQA of a map F at a point x_*, and $k_* \in K$, then k_* is a regular zero of H_A if and only if there exists $\xi_* \in X$ such that (ξ_*, k_*) is a regular zero of $\mathcal{G}_A$. In particular, H_A has a regular zero if and only if $\mathcal{G}_A$ has a regular zero. □

Proof Recall that a regular zero of $\mathcal{G}_A$ is a pair $(\xi_*, k_*) \in X \times K$ such that $L\xi_* + \frac{1}{2}Q(k_*) = 0$ and the linear map $X \times K \ni (x, k) \mapsto Lx + B^Q(k_*, k) \in Y$ is surjective.

If k_* is a regular zero of the Hessian H_A, then $\pi(Q(k_*)) = 0$, so $Q(k_*)$ belongs to LX, and then there exists $\xi_* \in X$ such that $\frac{1}{2}Q(k_*) + L \cdot \xi_* = 0$. Therefore $\mathcal{G}_A(\xi_*, k_*) = 0$, and $D\mathcal{G}_A(\xi_*, k_*)(x, k) = L \cdot x + B^Q(k_*, k)$ if $x \in X$, $k \in K$. If $y \in Y$, then the surjectivity of the map $K \ni k \mapsto \pi(B^Q(k_*, k)) \in Z$ implies that there exists a $k \in K$ such that $\pi(B^Q(k_*, k)) = \pi(y)$, i.e., that $y - B^Q(k_*, k)$ belongs to LX. It follows that there exists $x \in X$ with $y = B^Q(k_*, k) + Lx = D\mathcal{G}_A(h_*, k_*)(x, k)$. Since y is an arbitrary member of Y, the linear map $D\mathcal{G}_A(\xi_*, k_*) : X \times K \mapsto Y$ is surjective, so (ξ_*, k_*) is a regular zero of $\mathcal{G}_A$.

Conversely, if (ξ_*, k_*) is a regular zero of $\mathcal{G}_A$, then $\frac{1}{2}Q(k_*) + L \cdot \xi_* = 0$, so $Q(k_*) \in LX$, and then $H_A(k_*) = 0$. Furthermore, if $v \in Z$, and $y \in Y$ is such that $\pi(y) = v$, then we can write $y = D\mathcal{G}_A(\xi_*, k_*)(x, k) = Lx + B^Q(k_*, k)$ for some $x \in X$, $k \in K$. But then $v = \pi(y) = \pi(B^Q(k_*, k))$. Therefore the map $K \ni k \mapsto \pi(B^Q(k_*, k)) \in Z$ is surjective, and we have shown that k_* is a regular zero of H_A. □

THEOREM 22.22
Assume that X, Y are normed spaces, Y is finite-dimensional, Ω is open in X, $x_* \in \Omega$, $F : \Omega \mapsto Y$ is a continuous map, and $A = (L, K, Q)$ is a linear-quadratic approximation of F at x_* such that the Hessian H_A has a regular zero. Then F is open at x_*. □

THEOREM 22.23
Assume that X, Y are Banach spaces, Ω is open in X, $x_* \in \Omega$, $F : \Omega \mapsto Y$ is a continuous map and $A = (L, K, Q)$ is a strict linear-quadratic approximation of F at x_* such that the Hessian H_A has a regular zero. Then F is open at x_*. □

In particular, we can take F to be a map of the kind considered in Lemma 22.18. In that case, the *Hessian of F at x_** is the Hessian of the strict LQA (L, K, Q), where $L = DF(x_*)$, $K = \ker L$, and Q is the quadratic map $K \ni k \mapsto D^2F(x_*)(k, k)$.

THEOREM 22.24
Assume that X, Y are Banach spaces, Ω is open in X, $x_* \in \Omega$, $F : \Omega \mapsto Y$ is a map of class C^1 such that the derivative DF is differentiable at x_* and the Hessian H of F at x_* has a regular zero. Then F is open at x_*. □

REMARK 22.25
Theorem 22.24 is very similar to the result proved by Avakov (cf. [1, 2]). Avakov's sufficient condition for openness—called "2-regularity" by some authors [5, 6]—is

fomulated in slightly different terms, but is easily seen to be equivalent to the existence of a regular zero of the Hessian.

Precisely, the algebraic part of Avakov's condition says—using

$$L = DF(x_*)(k_*, k_*)\,, \quad K = \ker L\,,$$

and writing Q for the map $X \ni x \mapsto D^2F(x_*)(x,x) \in Y$, and $\boldsymbol{Q}$ for the restriction of Q to K—that

Av1. $Lk_* = 0$,

Av2. $Q(k_*) \in LX$,

and

Av3. *the map* $X \ni x \mapsto \mathcal{A}(x) \stackrel{\text{def}}{=} (Lx, \pi(B^Q(k_*, x))) \in LX \times Z$ *is surjective.*

LEMMA 22.26
The algebraic part of Avakov's condition holds if and only if k_* is a regular zero of the Hessian. □

Proof If Avakov's condition holds, then of course $k_* \in K$, so $\boldsymbol{Q}(k_*) \in LX$, and then $H(k_*) = 0$. Furthermore, the surjectivity of $\mathcal{A}$ implies, in particular, that given any $z \in Z$ there exists a $k \in X$ such that $(Lk, \pi(B^Q(k_*, k))) = (0, z)$. But then $Lk = 0$, so $k \in K$, and $\pi(B^{\boldsymbol{Q}}(k_*, k)) = z$. This shows that the map $K \ni k \mapsto \pi(B^{\boldsymbol{Q}}(k_*, k)) \in Z$ is surjective, so k_* is a regular zero of H.

Conversely, suppose that k_* is a regular zero of H. Then $\pi(\boldsymbol{Q}(k_*)) = 0$, so $Q(k_*) \in LX$. We now show that $\mathcal{A}$ is surjective. Pick $(y, z) \in LX \times Z$. Write $y = Lx$, $x \in X$. Let $v' = B^Q(k_*, x)$, $z' = \pi(v')$. Then the fact that the map $K \ni k \mapsto \pi(B^{\boldsymbol{Q}}(k_*, k)) \in Z$ is surjective implies that there exists a $k \in K$ such that $\pi(B^{\boldsymbol{Q}}(k_*, k)) = z - z'$. Then $\pi(B^Q(k_*, x + k)) = z' + (z - z') = z$, and $L(x + k) = Lx = y$, since $Lk = 0$. So $\mathcal{A}(x + k) = (y, z)$. Hence $\mathcal{A}$ is surjective, and our proof is complete. □

We point out, however, that in the work of Avakov and Ledzewicz-Schättler it also required, in addition to the algebraic condition described above, that the space LX be closed, whereas we do not need to make that extra requirement, because in our framework the purely algebraic condition on the Hessian suffices to obtain the openness theorem. □

Acknowledgment

Supported in part by NSF Grant DMS01-03901.

22.6 References

[1] Avakov, E.R., "Extremum conditions for smooth problems with equalitytype constraints." (Russian) *Zh. Vychisl. Mat. Mat. Fiz.* **25**, No. 5, 1985, pp. 680-693. Eng. translation in *U.S.S.R. Comput. Maths. Math. Phys.* **25**, No. 3 (Pergamon Press), 1985, pp. 24-32.

[2] Avakov, E.R., "Necessary conditions for an extremum for smooth abnormal problems with constraints of equality and inequality type." (Russian) *Matematicheskie Zametki* **45**, No. 6, 1989, pp. 3-11. Eng. translation in *Math. Notes* **47**, no. 5-6, 1990, pp. 425-432.

[3] Dontchev, A.L., *The Graves theorem revisited.* J. Convex Analysis **3**, 1996, pp. 45-53.

[4] Graves, L. M., *Some mapping theorems.* Duke Math. J. **17**, 1950, pp. 111–114.

[5] Ledzewicz, U., and H. Schättler, "An extended maximum principle." *Nonlinear Analysis, Theory, Methods and Applications* **29**, 1997, pp. 159-183.

[6] Ledzewicz, U., and H. Schättler, "A high-order generalized local maximum principle." *SIAM J. Control Optim.* **38**, 2000, pp. 823-854.

23

New Integrability Conditions for Classifying Holonomic and Nonholonomic Systems

Tzyh-Jong Tarn *Mingjun Zhang* *Andrea Serrani*

Abstract

This paper presents new integrability conditions for classifying holonomic and nonholonomic systems using the Frobenius Theorem in differential forms. Some of the previous results in the literature give sufficient conditions only. The results in this paper give necessary and sufficient conditions. The contribution of this paper is that it shows a new application of the Frobenius Theorem in differential forms for solving an integrability condition problem in systems and control area.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 317-331, 2003.
© Springer-Verlag Berlin Heidelberg 2003

23.1 Introduction

Dynamics of many under-actuated mechanical systems is subject to second order differential constraints. Based on the integrability of differential constraints, dynamic systems can be classified as either holonomic systems or nonholonomic systems [1].

Consider a dynamic system subject to a second order differential constraint as

$$f(\ddot{q}, \dot{q}, q, t) = 0 \tag{23.1}$$

where $t \in \mathbb{R}$ represents time. The variables $q, \dot{q}$, and $\ddot{q} \in \mathbb{R}^n$, $n \in \mathbb{N}$ can be regarded as the position, velocity, and acceleration of the physical system, respectively.

If (23.1) can be integrated completely as a constraint of a form $g(q,t) = 0$, than the dynamic system can be regarded a holonomic system. Otherwise, it is a nonholonomic system defined as a system with constraints that are not integrable. Although most examples of nonholonomic constraints presented in the literature are kinematics, differential constraints of higher order such as second order nonintegrable differential constraints can be considered as nonholonomic constraint in principle. Second order nonintegrable differential constraints are defined as second order nonholonomic constraints in the systems and control area. This definition is adopted from classical mechanics.

It has been shown that many under-actuated mechanical systems are subject to second order nonholonomic constraints [2], and they are nonholonomic systems. As well known in the literature, it is impossible to asymptotically stabilize a large class of nonholonomic systems to an equilibrium state using smooth feedback even locally.

By the definition of holonomic and nonholonomic systems, it is clear that the integrability conditions for differential constraints are very important for classifying holonomic and nonholonomic systems. Some integrability conditions, such as the Exact Integrable Theorem and the Integrable Differential Constraint Theorem, can be easily found in many books, but not so easily be applied to differential constraints. For many physical systems, differential constraints are normally not simple, and it is almost impossible to investigate the integrability properties by finding their closed forms.

Two well known results in the literature about classifying holonomic and nonholonomic systems are given by Oriolo and Nakamura [3], and Wichlund et al. [4]. However, the conditions are only sufficient. New conditions using the Frobenius Theorem in differential forms are given in this paper. The conditions are necessary and sufficient. In general, they can be applied to any order differential constraints.

23.2 Previous Work on Integrability Conditions

The following two conditions about integrability of differential constraints have been widely studied in the literature.

Exact Differential Theorem [5]: The left side of $\sum_{i=1}^{k} f_i(u)du_i = 0$ with continuously differentiable coefficients in two or more variables is exact if and

only if

$$\frac{\partial f_i}{\partial u_j} = \frac{\partial f_j}{\partial u_i}, \quad i, j = 1, ..., k; k \in \mathbb{N} \tag{23.2}$$

for all values of u.

Clearly, the statement of integrability conditions for multi-variable differential constraints, which is the case for many dynamic systems, is much more complex.

Integrable Differential Constraint Theorem [5]: With continuously differentiable coefficients $f_i(u)$, the differential constraint of $\sum_{i=1}^{k} f_i(u)du_i = 0$ in three or more variables, is integrable if and only if

$$f_\gamma \left(\frac{\partial f_\beta}{\partial u_\alpha} - \frac{\partial f_\alpha}{\partial u_\beta} \right) + f_\beta \left(\frac{\partial f_\alpha}{\partial u_\gamma} - \frac{\partial f_\gamma}{\partial u_\alpha} \right) + f_\alpha \left(\frac{\partial f_\gamma}{\partial u_\beta} - \frac{\partial f_\beta}{\partial u_\gamma} \right) = 0$$

for all α, β, and γ, and for all values of u. If the differential constraint satisfies the above equation, then there exists an integrating factor $g(u) \neq 0$, such that the differential form

$$\sum_{i=1}^{k} g(u) f_i(u) du_i = 0, k \in \mathbb{N} \tag{23.3}$$

is exact.

For $k = 3$, there is only one condition to evaluate in equation (23.3). For $k > 3$, however, there are $k(k-1)(k-2)/6$ conditions, which becomes oppressive when k is large. It is not practical to use this condition to identify the integrability for second order dynamic constraints.

Because of the need for conditions that can be applied directly to identify integrability of second order differential constraints, many studies have been conducted since the early part of the last decade [3, 4]. The following are two well-known results in the literature.

Oriolo-Nakamura's Conditions: Oriolo and Nakamura [3] discussed integrability conditions for second order nonholonomic constraints of an under-actuated manipulator with m actuated joints from a total of n joints. The constraint is expressed in the generalized coordinate as

$$M_u(q)\ddot{q} + C_u(q, \dot{q}) + e_u(q) = 0 \tag{23.4}$$

where $q, \dot{q}$ and $\ddot{q} \in \mathbb{R}^n$ are joint variables. $M_u(q)$ is an $(n-m) \times n$ inertia matrix. $C_u(q, \dot{q}) = c_u(q, \dot{q})\dot{q}$ are the corresponding Coriolis and centripetal torques, and $e_u(q)$ is the gravitational term. All of these terms correspond to under-actuated parts.

Conditions for partial integrable and complete integrable are introduced as follows:

The constraint in (23.4) is partially integrable if and only if

- The gravitational torque e_u is constant.
- The unactuated joint variables q_u do not appear in the manipulator inertia matrix $M(q)$.

The constraint in (23.4) is completely integrable (holonomic constraints) if and only if

- It is partially integrable.
- The distribution Δ defined by null space of $M_u(q)$ is involutive.

In summary, the constraint in (23.4) is a second order nonholonomic constraint if and only if it is not even partially integrable. If it is partially integrable, but not completely integrable, then it is a first order nonholonomic constraint. Otherwise, it is a holonomic constraint.

The above results are for constraints (23.4) in the generalized coordinate. It cannot be applied directly to constraints in a moving coordinate, because Oriolo and Nakamura's assumptions of the Coriolis and centripetal matrix are not satisfied in moving coordinate. This is the reason that Wichlund et al. [4] have proposed the following conditions for a constraint in a moving coordinate, a constraint that has a damping term and includes a kinematics transformation from velocity to configuration.

Wichlund's Conditions: Wichlund et al. [4] extended the results of Oriolo and Nakamura by studying the dynamics of under-actuated vehicles. They discussed a model in a moving coordinate as

$$\begin{aligned} M_u\dot{v} + C_u(v)v + D_u(v)v + e_u(q) &= 0 \\ \dot{q} &= J(q)v \end{aligned} \tag{23.5}$$

where $q \in \mathbb{R}^n$ denotes the position and orientation vector with coordinates in the generalized coordinate. And, $v \in \mathbb{R}^n$ denotes the linear and angular velocity vector with coordinates in a moving coordinate. M_u denotes the last $(n-m)$ rows of the inertia matrix, including added mass. It is a constant matrix. C_u denotes the last $(n-m)$ rows of the matrix of Coriolis and centripetal terms, also including added mass. D_u denotes the last $(n-m)$ rows of the damping matrix, and e_u denotes the last $(n-m)$ elements of the vector of gravitational and possible buoyant forces and moments. In (23.5), $\dot{q} = J(q)v$ represents the kinematics equation. This is a general dynamic model for under-actuated vehicles in a moving coordinate.

Wichlund's conditions for partially integrable and completely integrable are given as follows:

The constraint (23.5) is partially integrable if and only if

- The gravity term e_u is constant.
- The Coriolis, centripetal and damping terms $(C_u(v) + D_u(v))$ is a constant matrix.
- The distribution $\Omega^{\perp}$ defined by $\Omega^{\perp} = \ker((C_u + D_u)J^{-1}(q))$ is completely integrable.

The constraint (23.5) is completely integrable (holonomic constraints) if and only if

- It is partially integrable.

- The Coriolis, centripetal and damping terms $C_u + D_u = 0$.
- The distribution Δ defined by $\Delta(q) = \ker(M_u J^{-1}(q))$ is completely integrable.

Just as in Oriolo-Nakamura's Conditions, the constraint (23.5) is a second order nonholonomic constraint if and only if it is not even partially integrable. If it is partially integrable, but not completely integrable, then it is a first order nonholonomic constraint. Otherwise, it is a holonomic constraint.

Oriolo-Nakamura's and Wichlund's Conditions are two well-known conditions for classifying holonomic and nonholonomic constraints in the literature. However, they are only sufficient. This conclusion can be obtained by applying them to same dynamic systems described in different coordinate systems. Let us examine three examples using these conditions.

Example One: Yang [6] considered a special case of an underwater vehicle that is subject to holonomic constraint $m_{u1}\ddot{q}_1 = 0$ in a generalized coordinate. The model is obtained by considering a rigid body whose position and orientation is described by a set of generalized coordinate $q = [q_1^T, q_2^T]^T$, where $q_1 = [x, y, z]^T$, $q_2 = [\phi, \theta, \pi]^T$. Denote v_1 as the translational velocity of the rigid body in the moving coordinate; Then, we have $\dot{q}_1 = J_1(q_2)v_1$, where $J_1 q_2$ is the matrix corresponding to kinematic equation of translational velocity from moving coordinate to generalized coordinate, which is always invertible. Assume none of the generalized velocities are actuated, i.e., q_1 is unactuated, m_{u1} is the mass of the body. Using Lagrangian formalism, the part corresponding q_1 can be expressed as $m_{u1}\ddot{q}_1 = 0$.

Pre-multiplying the nonsingular matrix J_1^T for $m_{u1}\ddot{q}_1 = 0$ and after substituting the kinematic equation $\dot{q}_1 = J_1(q_2)v_1$, we can obtain Equation (6), which has constant matrix for the inertia coefficient; so that Wichlund's condition can be applied to for checking the integrability of the constraint.

$$m_{u1}\dot{v}_1 + C_1(v_2)v_1 = 0 \tag{23.6}$$

By Wichlund's Conditions for integrability of constraint in a moving coordinate, we conclude that the constraint (23.6) is a nonholonomic constraint($C_1(v2)$ is not constant). We know, however, that this is not the case.

Example Two: Consider a rigid body whose position and orientation are described by a set of generalized coordinates $q = [q_1, q_2]^t$, and a set of generalized velocities, respectively. If the body does not have potential energy due to gravity, a simple calculation using Lagrangian Formulation gives the following differential constraint

$$m\ddot{q}_1 + \alpha\dot{q}_1 = 0 \tag{23.7}$$

where m is the mass of the body and $\alpha > 0$ is a coefficient expressing the correspondence of the velocity and friction force term.

Clearly, constraint (23.7) is completely integrable and the constraint is a holonomic constraint, because the solution of the equation is

$$q_1(t) = q_1(0) + \frac{m\dot{q}_1(0)}{\alpha}(1 - e^{\frac{\alpha}{m}t})$$

If we substitute the transformation $\dot{q}_1 = J_1(q_2)v_1$ into the above constraint (23.7), we have a constraint in the moving coordinate as

$$m\dot{v}_1 + C_1(v_2)v_1 + \alpha v_1 = 0 \tag{23.8}$$

where $C_1(v_2) = mv_2 \times v_1$.

By Wichlund's Conditions, one can verify that it is a nonholonomic constraint ($C_1(v2)$ is not constant). However, we know it is not the case either.

Example Three: Consider the following holonomic constraint in a moving coordinate

$$\dot{v}_1 + v_1 = 0 \tag{23.9}$$

which is integrable.

If we transfer it back into the generalized coordinate by using the kinematic transformation equation

$$v_1 = J_1^{-1}(q_2)\dot{q}_1 \tag{23.10}$$

then

$$\dot{v}_1 = J_1^{-1}(q_2)\ddot{q}_1 + \dot{J}_1^{-1}(q_2)\dot{q}_1 \tag{23.11}$$

Substitute (23.10) and (23.11) into (23.9), then

$$J_1^{-1}(q_2)\ddot{q}_1 + \left[\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right]\dot{q}_1 = 0$$

By Oriolo and Nakamura's integrability conditions, it is a nonholonomic constraint. The available integrability conditions, however, lead to a contradiction again.

From Example One and Example Two, we find dynamic systems which are integrable in the generalized coordinate. After transferring the system into a moving coordinate, Wichlund's Conditions lead to opposite conclusions.

On the other hand, Example Three shows a system, which is integrable in a moving coordinate. However, after transferring the system into a generalized coordinate, Oriolo-Nakamura's Conditions fail to identify the integrability.

The reason for the failure is that Oriolo-Nakamura's and Wichlund's Conditions are obtained by performing derivation only for models (23.4) and (23.5), respectively. For Oriolo-Nakamura's Conditions, the Coriolis and centripetal torque terms are also required to satisfy the Christoffel explicit expression

$$C_u(q,\dot{q}) = \dot{H}_u(q)\dot{q} - \frac{1}{2}\left[\frac{\partial(\dot{q}^T H\dot{q})}{\partial q_u}\right]^T$$

which is a standard result of the Euler-Lagrangian equation and does not contain any additional external terms, such as environmental effects and friction forces.

The above condition is an overly restrictive requirement for many practical systems. Example Two also fails for this reason (there is a friction term in the dynamic model).

It turns out as no surprise that Oriolo-Nakamura and Wichlund's conditions all depend on dynamic models. As a result, even though the above conditions are thought to be sufficient and necessary, they are only sufficient conditions.

In conclusion, a coordinate independent integrability condition that can be used to identify integrability properties of dynamic constraints universally is strongly desired. However, to the best of the author's knowledge, no such results are available in the literature.

With this goal in mind, we propose the following new integrability conditions.

23.3 New Integrability Conditions

The Frobenius Theorem [8] gives a necessary and sufficient integrability condition that is coordinate independent for any order differential constraints. One form of the Frobenius Theorem has been used widely in solving nonlinear control systems [7]. Another format of the Frobenius Theorem is expressed in differential forms. The purpose of this paper is to study integrability conditions for dynamic constraints using the Frobenius Theorem in differential forms.

The Frobenius Theorem [8]: supposes that $(\omega_{p+1}, \omega_{p+2}, ..., \omega_n)$ is a system of $n-p$ differential forms of degree 1, of class $\mathbb{C}^1$, in an open $\mathbb{U} \subset \mathbb{R}^n$, such that, at each point $x \in \mathbb{U}$, the rank of the system $(\omega_{p+1}, \omega_{p+2}, ..., \omega_n)$ is equal to $n-p$. Then the differential system $\omega_i = 0, (p+1 \leq i \leq n)$ is completely integrable if, and only if, the differential forms $d\omega_i \wedge \omega_{p+1} \wedge \omega_{p+2} \wedge ... \wedge \omega_n, (p+1 \leq i \leq n)$ vanish identically.

In order to use this theorem, we should first transform our dynamic constraints into proper differential forms. This can be done by transforming the dynamic systems into normal control form, i.e., first order differential equations. Then we can cast the system into differential forms using the Frobenius Theorem. As a result, necessary and sufficient integrability conditions can be obtained.

We will first consider systems described in a generalized coordinate, then systems in a moving coordinate. Finally, we will give new integrability conditions based on the discussions.

Generalized Coordinate Systems: Consider a class of under-actuated mechanical systems in a generalized coordinate, the differential constraint is expressed by

$$M_u(q)\ddot{q} + C_u(q, \dot{q}) + e_u(q) = 0 \tag{23.12}$$

where q is the vector of generalized coordinates. $M_u(q)$ is the $(n-m) \times n$ inertia matrix corresponding to the under-actuated part, $C_u(q, \dot{q}) = c_u(q, \dot{q})\dot{q}$ is the vector of Coriolis and centripetal torques corresponding to the under-actuated part, and $e_u(q)$ is the gravitational term corresponding to the under-actuated part. Here, m is the number of control input from total of n degrees of freedom.

Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, and

$$X_1 = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix} = q,$$

$$X_2 = \begin{bmatrix} x_{n+1} \\ x_{n+2} \\ \vdots \\ x_{2n} \end{bmatrix} = \begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \vdots \\ \dot{q}_n \end{bmatrix} = \dot{q}$$

Then

$$dX_1 = X_2 dt$$
$$M_u(X_1)dX_2 + C_u(X)dt + e_u(X_1)dt = 0$$

where

$$M_u(X_1) = \begin{bmatrix} m_{m+1,1} & m_{m+1,2} & \ldots & m_{m+1,n} \\ m_{m+2,1} & m_{m+2,2} & \ldots & m_{m+2,n} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n,1} & m_{n,2} & \ldots & m_{n,n} \end{bmatrix}$$

$$C_u(X) = \begin{bmatrix} c_{m+1,1} & c_{m+1,2} & \ldots & c_{m+1,n} \\ c_{m+2,1} & c_{m+2,2} & \ldots & c_{m+2,n} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n,1} & c_{n,2} & \ldots & c_{n,n} \end{bmatrix}$$

$$e_u(X_1) = \begin{bmatrix} e_{m+1} \\ e_{m+2} \\ \vdots \\ e_n \end{bmatrix}$$

After simple calculation, let

$$\begin{aligned}
\omega_1 &= dx_1 \\
&\vdots \\
\omega_n &= dx_n \\
\omega_{n+1} &= dx_1 - x_{n+1}dt \\
&\vdots \\
\omega_{2n} &= dx_n - x_{2n}dt \\
\omega_{2n+1} &= m_{m+1,1}(X_1)dx_{n+1} + ... + m_{m+1,n}(X_1)dx_{2n} \\
&\quad + c_{m+1,1}(X)dt + ... + c_{m+1,n}(X)dt + e_{m+1}(X_1)dt \\
&\vdots \\
\omega_{3n-m} &= m_{n,1}(X_1)dx_{n+1} + ... + m_{n,n}(X_1)dx_{2n} \\
&\quad + c_{n,1}(X)dt + ... + c_{n,n}(X)dt + e_n(X_1)dt
\end{aligned}$$

We have developed a sequence of $3n - m$ differential forms in such a manner that $\omega_1, \omega_2, ..., \omega_{2n}$ forms a basis for the differential 1-forms (i.e., each different constraint of 1-form may then be written uniquely in the form $\sum_{i=1}^{2n} \alpha_i(X)$). Here, $\omega_{2n+1}, \omega_{2n+2}, ..., \omega_{3n-m}$ is a system of $n - m$ differential forms of degree 1. By the Frobenius Theorem, the above system is integrable if, and only if, the differential forms

$$d\omega_i \wedge \omega_{2n+1} \wedge \omega_{2n+2} \wedge ... \wedge \omega_{3n-m} = 0, \quad (2n + 1 \leq i \leq 3n - m) \tag{23.13}$$

where

$$\begin{aligned}
\omega_i &= m_{i-2n+m,1}(X_1)dx_{n+1} + ... + m_{i-2n+m,n}(X_1)dx_{2n} \\
&\quad + c_{i-2n+m,1}(X)dt + ... + c_{i-2n-m,n}(X)dt + e_{i-2n+m}(X_1)dt
\end{aligned}$$

We note that if the above system is integrable, the system (23.12) is only partially integrable, and hence we have the following conclusion.

The second order differential constraint (23.12) is partially integrable if, and only if, (23.13) holds.

The integrated constraints still contain the first order differential of variables in the generalized coordinate, such as $x_{n+1}, x_{n+2}, ..., x_{2n}$.

Furthermore, if the second order differential constraint (23.12) is partially integrable, and after it is integrated into new forms involving variables X_1 and X_2, then $\omega_{n+1}, \omega_{n+2}, ..., \omega_{2n}$ can be used to substitute X_2, and the first order constraint only involves X_1. The Frobenius Theorem can be used again to identify the integrability. Thus, the following completely integrable condition holds.

The second order differential constraint (23.12) is completely integrable if, and only if, it is partially integrable and the following condition holds

$$d\bar{\omega}_i \wedge \bar{\omega}_{2n+1} \wedge \bar{\omega}_{2n+2} \wedge ... \wedge \bar{\omega}_{3n-m} = 0, \quad (2n + 1 \leq i \leq 3n - m) \tag{23.14}$$

where $\bar{\omega}_i$ is the integrated 1-form corresponding to ω_i in (23.13) after substituting $\omega_{n+1}, \omega_{n+2}, ..., \omega_{2n}$.

Moving Coordinate Systems: Consider the case where a dynamic constraint is given in a moving coordinate as follows

$$M_u \dot{v} + C_u(v)v + e_u(q) = 0 \tag{23.15}$$

where $v \in \mathbb{R}^n$ denotes the linear and angular velocity vector with coordinates in the moving coordinate. M_u denotes the last $(n-m)$ rows of the inertia matrix, including added mass. $C_u(v) = c_u(v)v$ denotes the last $(n-m)$ rows of the matrix of Coriolis and centripetal terms, also including added mass and damping terms. Here, $\dot{q} = J(q)v$ is the kinematics equation and $J(q)$ is assumed to be invertible.

As in the previous discussion, let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, and

$$X_1 = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix} = q$$

$$X_2 = \begin{bmatrix} x_{n+1} \\ x_{n+2} \\ \vdots \\ x_{2n} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

Then

$$M_u dX_2 + C_u(X_2)X_2 dt + e_u(X_1)dt = 0$$
$$X_2 dt = J^{-1}(X_1)dX_1$$

where M_u and $J(q)$ are nonsingular matrices [4], and

$$M_u(X_1) = \begin{bmatrix} m_{m+1,1} & m_{m+1,2} & \dots & m_{m+1,n} \\ m_{m+2,1} & m_{m+2,2} & \dots & m_{m+2,n} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n,1} & m_{n,2} & \dots & m_{n,n} \end{bmatrix}$$

$$C_u(X_2) = \begin{bmatrix} c_{m+1,1} & c_{m+1,2} & \dots & c_{m+1,n} \\ c_{m+2,1} & c_{m+2,2} & \dots & c_{m+2,n} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n,1} & c_{n,2} & \dots & c_{n,n} \end{bmatrix}$$

$$e_u(X_2) = \begin{bmatrix} e_{m+1} \\ e_{m+2} \\ \vdots \\ e_n \end{bmatrix}$$

$$J^{-1}(X_1) = \begin{bmatrix} J_{11}^{-1} & J_{12}^{-1} & \dots & J_{1n}^{-1} \\ J_{21}^{-1} & J_{22}^{-1} & \dots & J_{2n}^{-1} \\ \vdots & \vdots & \vdots & \vdots \\ J_{n1}^{-1} & J_{n2}^{-1} & \dots & J_{nn}^{-1} \end{bmatrix}$$

After simple calculation, then

$$\begin{aligned}
\omega_1 &= dx_1 \\
&\vdots \\
\omega_n &= dx_n \\
\omega_{n+1} &= J_{11}^{-1}dx_1 + J_{12}^{-1}dx_2 + \dots + J_{1n}^{-1}dx_n - x_{n+1}dt \\
&\vdots \\
\omega_{2n} &= J_{n1}^{-1}dx_1 + J_{n2}^{-1}dx_2 + \dots + J_{nn}^{-1}dx_n - x_{2n}dt \\
\omega_{2n+1} &= m_{m+1,1}dx_{n+1} + \dots + m_{m+1,n}dx_{2n} \\
&\quad + c_{m+1,1}(X_2)x_{n+1}dt + \dots + c_{m+1,n}(X_2)x_{2n}dt \\
&\quad + e_{m+1}(X_1)dt \\
&\vdots \\
\omega_{3n-m} &= m_{n,1}dx_{n+1} + \dots + m_{n,n}dx_{2n} \\
&\quad + c_{n,1}(X_2)x_{n+1}dt + \dots + c_{n,n}(X_2)x_{2n}dt \\
&\quad + e_n(X_1)dt
\end{aligned}$$

Similar to the previous discussion, the following partial integrability condition is obtained.

The second order differential constraint 23.15 is partially integrable if, and only if, the following condition holds:

$$d\omega_i \wedge \omega_{2n+1} \wedge \omega_{2n+2} \wedge ... \wedge \omega_{3n-m} = 0, \quad (2n+1 \le i \le 3n-m)$$

where

$$\begin{aligned}\omega_i &= m_{i-2n+m,1}dx_{n+1} + ... + m_{i-2n+m,n}dx_{2n} \\ &\quad + c_{i-2n+m,1}(X_2)x_{n+1}dt + ... + c_{i-2n+m,n}(X_2)x_{2n}dt + e_{i-2n+m}(X_1)dt\end{aligned}$$

If the second order constraint (23.15) is partially integrable, and after it is integrated into new forms involving X_1 and X_2, we may use the kinematics transformation $X_2 = J^{-1}(X_1)\dot{X}_1$ to substitute X_2, and then a first order differential constraint involving only X_1 can be obtained. The Frobenius Theorem can be used again to identify the integrability of the new first order differential constraint. Thus, the following completely integrable condition holds.

The second order differential constraint (23.15) is completely integrable if, and only if, it is partially integrable and

$$d\bar{\omega}_i \wedge \bar{\omega}_{2n+1} \wedge \bar{\omega}_{2n+2} \wedge ... \wedge \bar{\omega}_{3n-m} = 0, \quad (2n+1 \le i \le 3n-m)$$

where $\bar{\omega}_i$ is the integrated 1-form corresponding to ω_i after substituting ω_{n+1}, ω_{n+2},..., ω_{2n}.

Simple investigation shows that the forms of the above integrability conditions in both a generalized coordinates or a moving coordinate are the same. The only difference is the coefficient terms in differential forms. Therefore, we have the following new integrability conditions for classifying holonomic and nonholonomic constraints.

Theorem 1: The second order differential constraints of under-actuated mechanical systems in the form of (23.12) or (23.15) is partially integrable if, and only if, the following condition holds

$$d\omega_i \wedge \omega_{2n+1} \wedge \omega_{2n+2} \wedge ... \wedge \omega_{3n-m} = 0, \quad (2n+1 \le i \le 3n-m) \tag{23.16}$$

where

$$\begin{aligned}\omega_i &= m_{i-2n+m,1}dx_{n+1} + ... + m_{i-2n+m,n}dx_{2n} + c_{i-2n+m,1}(X_2)dt \\ &\quad + ... + c_{i-2n+m,n}(X_2)dt + e_{i-2n+m}(X_1)dt \\ M_u &= [m_{i,j}]_{(n-m)\times n}, \quad m+1 \le i \le n, \quad 1 \le j \le n. \\ C_u &= [c_{i,j}]_{(n-m)\times n}, \quad m+1 \le i \le n, \quad 1 \le j \le n. \\ e_u &= [e_i]_{(n-m)\times 1}, \quad m+1 \le i \le n.\end{aligned}$$

Theorem 2: the second order differential constraint (23.12) or (23.15) is completely integrable if, and only if, it is partially integrable and

$$d\bar{\omega}_i \wedge \bar{\omega}_{2n+1} \wedge \bar{\omega}_{2n+2} \wedge ... \wedge \bar{\omega}_{3n-m} = 0, \quad (2n+1 \le i \le 3n-m)$$

where $\bar{\omega}_i$ is the integrated 1-form corresponding to ω_i after substituting ω_{n+1},ω_{n+2},...,ω_{2n}.

In summary, the constraint (23.12) or (23.15) is a holonomic constraint if it is completely integrable. Otherwise, it is a nonholonomic constraint.

23.4 Applications of the New Conditions

The three examples discussed earlier show that the integrability conditions for second order differential constraints are only sufficient. Let us see what happens if we apply the above new integrability conditions for the three examples.

Example One: The constraint $m_{u1}\ddot{q}_1 = 0$ is integrable in the generalized coordinate. Once transferred into a moving coordinate, then

$$m_{u1}\dot{v}_1 + C_1(v_2)v_1 = 0 \tag{23.17}$$

and we have

$$\begin{aligned} \omega &= m_{u1}dv_1 + C_1(v_2)J_1^T(q_2)dq_1 \\ d\omega &= dm_{u1} \wedge dv_1 + d\left[C_1(v_2)J_1^T(q_2)\right] \wedge dq_1 \end{aligned}$$

Then

$$\begin{aligned} &\omega \wedge d\omega = m_{u1}dv_1 \wedge d\left[C_1(v_2)J_1^T(q_2)\right] \wedge dq_1 \\ &\quad + C_1(v_2)J_1^T(q_2)dq_1 \wedge dm_{u1} \wedge dv_1 \\ &= d\left[C_1(v_2)J_1^T(q_2)\right] \wedge dq_1 \wedge m_{u1}dv_1 \\ &\quad + dm_{u1} \wedge dv_1 \wedge C_1(v_2)J_1^T(q_2)dq_1 \\ &= d\left[C_1(v_2)J_1^T(q_2)dq_1 \wedge m_{u1}dv_1\right] \\ &\quad + d\left[m_{u1}dv_1 \wedge C_1(v_2)J_1^T(q_2)dq_1\right] \\ &= d\left[C_1(v_2)J_1^T(q_2)dq_1 \wedge m_{u1}dv_1 + m_{u1}dv_1 \wedge C_1(v_2)J_1^T(q_2)dq_1\right] \\ &= d\left[C_1(v_2)J_1^T(q_2)dq_1 \wedge m_{u1}dv_1 - C_1(v_2)J_1^T(q_2)dq_1 \wedge m_{u1}dv_1\right] \\ &= 0 \end{aligned}$$

By the above **Theorem 1**, the constraint (23.17) is integrable, which is consistent with the result in the generalized coordinate.

Example Two: By equation (23.8) the exact transformation to a moving coordinate can be expressed as

$$m\dot{v}_1 + C_1(v_2)v_1 + \alpha v_1 = 0 \tag{23.18}$$

and we have

$$\begin{aligned} \omega &= mdv_1 + \left[C_1(v_2)J_1^T(q_2) + \alpha J_1^T(q_2)\right]dq_1 \\ d\omega &= dm \wedge dv_1 + d\left[C_1(v_2)J_1^T(q_2) + \alpha J_1^T(q_2)\right] \wedge dq_1 \end{aligned}$$

Then

$$\begin{aligned}
\omega \wedge d\omega &= m dv_1 \wedge d\left[C_1(v_2)J_1^T(q_2) + \alpha J_1^T(q_2)\right] \wedge dq_1 \\
&\quad + \left[C_1(v_2)J_1^T(q_2) + \alpha J_1^T(q_2)\right] dq_1 \wedge dm \wedge dv_1 \\
&= d\left[C_1(v_2)J_1^T(q_2) + \alpha J_1^T(q_2)\right] \wedge dq_1 \wedge m dv_1 \\
&\quad + dm \wedge dv_1 \wedge \left[C_1(v_2)J_1^T(q_2) + \alpha J_1^T(q_2)\right] dq_1 \\
&= d\left[\left(C_1(v_2)J_1^T(q_2) + \alpha J_1^T(q_2)\right) dq_1 \wedge m dv_1\right] \\
&\quad + d\left[m dv_1 \wedge \left[C_1(v_2)J_1^T(q_2) + \alpha J_1^T(q_2)\right] dq_1\right] \\
&= 0
\end{aligned}$$

By the above **Theorem 1**, the constraint (23.18) is integrable, which is consistent with the result in the generalized coordinate.

Example Three: Let $\dot{q}_1 = q_3$, and substitute it back into (23.10), then from (23.9)

$$J_1^{-1}(q_2)\dot{q}_3 + \left[\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right]\dot{q}_1 = 0 \tag{23.19}$$

and we have

$$\begin{aligned}
\omega &= J_1^{-1}(q_2)dq_3 + \left[\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right] dq_1 \\
d\omega &= d\left[J_1^{-1}(q_2)\right] \wedge dq_3 + d\left[\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right] \wedge dq_1
\end{aligned}$$

Then

$$\begin{aligned}
\omega \wedge d\omega &= J_1^{-1}(q_2)dq_3 \wedge d\left[\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right] \wedge dq_1 \\
&\quad + \left[\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right] dq_1 \wedge d\left[J_1^{-1}(q_2)\right] \wedge dq_3 \\
&= d\left[\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right] \wedge dq_1 \wedge J_1^{-1}(q_2)dq_3 \\
&\quad + d\left[J_1^{-1}(q_2)\right] \wedge dq_3 \wedge \left[\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right] dq_1 \\
&= d\left[\left(\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right) dq_1 \wedge J_1^{-1}(q_2)dq_3\right] \\
&\quad + d\left[\left(J_1^{-1}(q_2)\right) dq_3 \wedge \left(\dot{J}_1^{-1}(q_2) + J_1^{-1}(q_2)\right) dq_1\right] \\
&= 0
\end{aligned}$$

By the above **Theorem 1**, the constraint (23.19) is integrable, which is consistent with the result in the moving coordinate.

The above three examples show that the new sufficient and necessary integrability conditions can be applied to a dynamic system no matter which coordinate has been chosen to describe the physical system.

23.5 Conclusions

This paper considers whether a differential constraint for a dynamic system is a second order nonholonomic constraint, a first order nonholonomic constraint, or a holonomic constraint. The conditions are obtained by the Frobenius theorem in differential forms. The conditions are necessary and sufficient. Thus, it allows us

to identify the essential nature of a dynamic system no matter which coordinate system has been chosen to describe the system. If the second order differential constraints are not even partially integrable, we have a second order nonholonomic constraint. If the second order differential constraints are partially integrable, but not completely integrable, it is a first order nonholonomic constraint. Otherwise, it is a holonomic constraint. This definition of second order nonholonomic constraint is adopted from classical mechanics for the studies in systems and control area. In general, the conditions can be used for identifying integrability for any order differential constraints.

23.6 References

[1] X. P. Yun and N. Sarkar. Unified Formulation of Robotic Systems with Holonomic and Nonholonomic Constraints. *IEEE Trans. on Robotics and Automation*, 14:640–650, 1998.

[2] I. Kolmanovsky and N. H. McClamroch. Developments in Nonholonomic Control Problems. *IEEE Control System Magazine*, 15:20–36, 1995.

[3] G. Oriolo and Y. Nakamura. Control of Mechanical Systems with Second-Order Nonholonomic Constraints: Under-actuated Manipulators. *Proceedings of the 30th IEEE Conference on Decision and Control*,306-310,1991.

[4] K. Y. Wichlund, O. J. Sordalen and O. Egeland. Control Properties of Under-actuated Vehicles. *Proceedings of the IEEE International Conference on Robotics and Automation*, 2009–2014, 1995.

[5] R. M. Murray, Z. Li and S. S. Sastry. A Mathematical Introduction to Robotic Manipulation. *CRC Press*, 1994.

[6] X. P. Yang. Dynamic Modeling and Control of Underwater Vehicle with Multi-manipulator Systems. *D.Sc. dissertation, Washington University in St. Louis*, Sept. 1997.

[7] A. Isidori. Nonlinear control saystems. *Springer-Verlag, New York, 3rd edition*, 1995.

[8] H. Cartan. Differential Forms. *Hermann Publishers in Arts and Science,Paris, France*, 1970.

24

On Spectral Analysis Using Models with Pre-specified Zeros

Bo Wahlberg

Abstract

The fundamental theory of Lindquist and co-workers on the rational covariance extension problem provides a very elegant framework for ARMA spectral estimation. Here the choice of zeros is completely arbitrary, and can be used to tune the estimator. An alternative approach to ARMA model estimation with pre-specified zeros is to use a prediction error method based on generalizing autoregressive (AR) modeling using orthogonal rational filters. Here the motivation is to reduce the number of parameters needed to obtain useful approximate models of stochastic processes by suitable choice of zeros, without increasing the computational complexity.

The objective of this contribution is to discuss similarities and differences between these two approaches to spectral estimation.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 333-344, 2003.
© Springer-Verlag Berlin Heidelberg 2003

24.1 Introduction and Problem Formulation

The concept of representing complex systems by simple models is fundamental in science. The aim is to reduce a complicated process to a simpler one involving a smaller number of parameters. The quality of the approximation is determined by its usefulness, e.g. its predictive ability. Autoregressive (AR) and autoregressive moving-average (ARMA) models are the dominating parametric models in spectral analysis, since they give useful approximations of many stochastic processes of interest. The ARMA model leads to non-linear optimization problems to be solved for best approximation, while the special case of AR modeling only involves a quadratic least squares optimization problem. Hence, AR models are of great importance in applications where fast and reliable computations are necessary.

The fact that the true system is bound to be more complex than a fixed order AR model has motivated the analysis of high-order AR approximations, where the model order is allowed to tend to infinity as the number of observations tends to infinity. However, aspects such as the number of observations, computational limitations and numerical sensitivity set bounds on how high an AR order can be tolerated in practice. Herein, we shall study discrete orthogonal rational function model structures, which reduce the number of parameters to be estimated without increasing the numerical complexity of the estimation algorithm.

Suppose that $\{y(t),\ t = \ldots -1, 0, 1, \ldots\}$ is a stationary linear regular random process with Wold representation,

$$y(t) = \sum_{k=0}^{\infty} h_k^0 e(t-k), \quad h_k^0 \in \mathbb{R}, \quad h_0^0 = 1. \tag{24.1}$$

Here $\{e(t)\}$ is a sequence of random variables with the properties

$$\mathrm{E}\{e(t)|\mathcal{F}_{t-1}\} = 0, \quad \mathrm{E}\{e(t)^2|\mathcal{F}_{t-1}\} = \sigma_0^2, \quad \mathrm{E}\{e(t)^4\} < \infty, \tag{24.2}$$

The transfer function, often called the noise filter or the shaping filter,

$$H_0(q) = \sum_{k=0}^{\infty} h_k^0 q^{-k}, \quad H_0(\infty) = 1, \tag{24.3}$$

is a function of the shift operator q, $qe(t) = e(t+1)$. By q^{-1} we mean the corresponding delay operator $q^{-1}e(t) = e(t-1)$. The power spectral density of $\{y(t)\}$ equals

$$\Phi(e^{i\omega}) = \sigma_0^2 |H_0(e^{i\omega})|^2. \tag{24.4}$$

We shall assume that the complex function $[H_0(z)]^{-1}$, $z \in \mathbb{C}$, is analytic in $|z| > 1$ and continuous in $|z| \geq 1$. Then

$$[H_0(z)]^{-1} = \sum_{k=0}^{\infty} a_k^0 z^{-k}, \quad |z| \geq 1. \tag{24.5}$$

We shall impose a further smoothness condition on $[H_0(z)]^{-1}$, namely

$$\sum_{k=0}^{\infty} k|a_k^0| < \infty. \tag{24.6}$$

By truncating the expansion (24.5) at $k = n$, we obtain a n^{th} order autoregressive (AR) approximation of (24.1),

$$A_n^0(q)y(t) = e(t), \quad A_n^0(q) = 1 + \sum_{k=1}^{n} a_k^0 q^{-k}. \tag{24.7}$$

A crucial question is how large an order n must be chosen to obtain a useful AR approximation. From (24.5) and (24.6) we know that the process (24.1) can be arbitrarily well approximated by an AR model (in the mean square sense) by taking the order n large enough. However, nothing is said about the rate of convergence. Assume that $H_0(z)$ is a rational function with zeros $\{z_i\}$, $|z_i| < 1$. The error in the AR approximation (24.7) is then of order δ^n, where $\delta = \max_j |z_j|$. Hence, zeros close to the unit circle imply a slow rate of convergence and consequently a high model order n.

This motivates the investigation of alternative approximations which are less sensitive to the location of the zeros. As discussed above the AR approximation corresponds to a truncated series expansion of $[H_0(z)]^{-1}$ in the basis functions $\{z^{-k}\}$. An natural extension is to replace $\{z^{-k}\}$ by more general orthonormal rational functions $\{F_k(z)\}$ in order to get more efficient representations.

Over the last decades, there has been considerable interest in the systems, signal processing and control literature in representing linear time-invariant dynamic systems using an expansion in a rational orthonormal basis. The recent monograph [2] gives an excellent overview of the field of orthogonal rational functions, which can be viewed as generalizations of orthogonal polynomials. In parallel to the efforts in applied mathematics a very similar theory has been developed in the fields of signals, systems and control. See [1] for a thorough presentation of this theory. The paper [8] provides an overview of orthogonal rational functions using a transformation approach.

We will make use of the following mathematical notation. Let P^T denote the transpose of a matrix, and P^* the complex conjugate transpose. Let $\mathbb{E}$ denote the exterior of the unit disc: $\{z \in \mathbb{C} : \ |z| > 1\}$, and $\mathbb{T}$ the unit circle: $\{z \in \mathbb{C} : \ |z| = 1\}$. By $\mathcal{H}_2(\mathbb{E})$ we mean the Hardy space of square integrable functions on $\mathbb{T}$, analytic in the region $\mathbb{E}$. We denote the corresponding inner product for $X(z), Y(z) \in \mathcal{H}_2(\mathbb{E})$ by

$$\langle X, Y \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{i\omega})^* Y(e^{i\omega}) \omega = \frac{1}{2\pi i} \oint_{\mathbb{T}} X^T(1/z) Y(z) \frac{dz}{z}. \tag{24.8}$$

Two functions $F_1(z), F_2(z) \in \mathcal{H}_2(\mathbb{E})$ are called orthonormal if $\langle F_1, F_2 \rangle = 0$ and $\langle F_1, F_1 \rangle = \langle F_2, F_2 \rangle = 1$.

24.2 Orthonormal Rational Functions

First, we will review some basic facts for orthogonal all-pass transfer functions and corresponding state space realizations. Consider a real, single-input single-output, exponentially stable, all-pass transfer function $H_b(z)$, $H_b(z)H_b(1/z) = 1$, of order m, specified by its poles $\{\xi_j;\ |\xi_j| < 1,\ j = 1 \dots m\}$. Such a transfer function, often called an inner function, can be represented by a Blaschke product

$$H_b(z) = \prod_{j=1}^{m} \frac{1 - \xi_j^* z}{z - \xi_j}, \quad |\xi_j| < 1. \tag{24.9}$$

It can also be represented by an orthogonal state-space realization

$$\begin{bmatrix} x(t+1) \\ y(t) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}, \quad \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^T = I. \tag{24.10}$$

Such a realization can easily be obtained by change of state variables, and is by no means unique. The special cases of a first order system and a second order system are, however, of special importance:

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} a & \sqrt{1-a^2} \\ \sqrt{1-a^2} & -a \end{bmatrix}, \quad \Rightarrow \\ H_b(z) &= \frac{1 - az}{z - a}, \quad -1 < a < 1, \end{aligned} \tag{24.11}$$

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} b & \sqrt{1-b^2} & 0 \\ c\sqrt{1-b^2} & -bc & \sqrt{1-c^2} \\ \sqrt{1-b^2}\sqrt{1-c^2} & -b\sqrt{1-c^2} & -c \end{bmatrix} \quad \Rightarrow \\ H_b(z) &= \frac{-cz^2 + b(c-1)z + 1}{z^2 + b(c-1)z - c}, \quad -1 < b < 1,\ -1 < c < 1. \end{aligned} \tag{24.12}$$

An all-pass transfer function can be factorized as $H_b(z) = H_{b1}(z)H_{b2}(z)$, where $H_{b1}(z)$ and $H_{b2}(z)$ are lower order all-pass transfer functions with orthogonal state-space realization (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2), respectively. Denote the corresponding state vectors by $x_1(t)$ and $x_2(t)$. It turns out that an orthogonal state space realization of $H_b(z)$ can be directly obtained by using the lumped state vector $x(t) = [x_1^T(t)\ x_2^T(t)]^T$. This observation is due to Mullis and Roberts, see [7]. By recursively using the factorization results, an orthogonal state space realization (A, B, C, D) of $H_b(z)$ can be constructed by cascading orthogonal state space realization of its first order all-pass factors with a real pole (24.11), and its second order all-pass factors with two complex conjugated poles (24.12). The corresponding network is illustrated in Figure 24.1.

Beside the cascading all-pass systems, it is also possible to construct orthogonal filters by using feedback connections. This results in ladder representations.

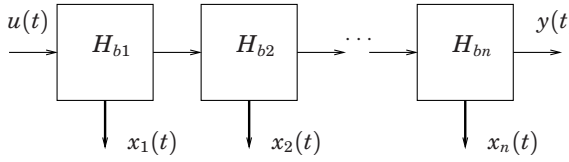


Figure 24.1 Network of all-pass filters.

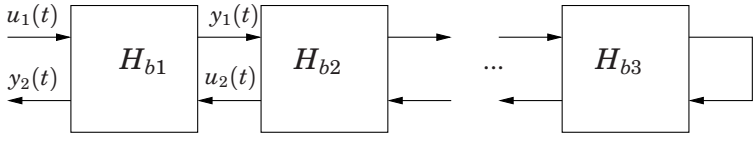


Figure 24.2 Two-port Feedback Network

The most famous being the Gray-Markel normalized ladder realization. Here $y(t) = [y_1(t)\ y_2(t)]^T$, and $u(t) = [u_1(t)\ u_2(t)]^T$, and

$$H_b(z) = \begin{bmatrix} \sqrt{1-\gamma^2} & \gamma \\ -\gamma & \sqrt{1-\gamma^2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix}, \quad -1 < \gamma < 1, \tag{24.13}$$

which is all-pass, $H_b(z)H_b^T(1/z) = I$. The corresponding two-port representation and feedback network are given in Figure 24.2. Notice that the feedback law $u_2(t) = y_1(t)$ gives

$$H_{bc}(z) = -\frac{1-az}{z-a}, \quad a = \sqrt{1-\gamma^2}.$$

An obvious generalization is to use filters of the form

$$H_b(z) = \begin{bmatrix} \sqrt{1-\gamma^2} & \gamma \\ -\gamma & \sqrt{1-\gamma^2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1-az}{z-a} \end{bmatrix}, \quad \begin{matrix} -1 < \gamma < 1 \\ -1 < a < 1 \end{matrix}. \tag{24.14}$$

Feedback lattice filters are closely related to forward lattice filters. The idea is to work with chain transfer functions, where the signal $u_1(t)$ now is considered as the first output and the signal $y_1(t)$ becomes the second input. The corresponding transfer function are now J-unitary instead of orthonormal. For example, transfer function (24.14) becomes

$$J(z) = \frac{1}{\sqrt{1-\gamma^2}} \begin{bmatrix} 1 & \gamma \\ \gamma & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix}, \quad -1 < \gamma < 1. \tag{24.15}$$

This is closely related to Mason's rule for network inversion, since one input becomes an output and *vice versa*. Cascading chain transfer functions of the form $J(z)$ leads to the well-known lattice filter, which often is used to decorrelate stochastic signals and estimation of AR models. Notice that the transfer functions

will have finite impulse response. A natural generalization to obtain infinite impulse filters is to use the chain transfer functions of the type

$$J(z) = \frac{1}{\sqrt{1-\gamma^2}} \begin{bmatrix} 1 & \gamma \\ \gamma & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1-az}{z-a} \end{bmatrix}, \quad \begin{matrix} -1 < \gamma < 1 \\ -1 < a < 1 \end{matrix}. \tag{24.16}$$

Study the input-to-state transfer function

$$V(z) = (zI - A)^{-1}B, \tag{24.17}$$

for an orthogonal state space realization of an all-pass transfer function. Denote the k^{th} canonical unit vector, $e_k = (0 \dots 1 \dots 0)^T$, where 1 is in position k. We then have the following fundamental result: Assume that (24.9) and (24.10) hold. The transfer functions $\{F_k(z) = e_k^T V(z),\ k = 1 \dots m\}$ are then orthonormal in $\mathcal{H}_2(\mathbb{E})$. The cascade realization gives

$$F_k(z) = \frac{\sqrt{1-|\xi_k|^2}}{z-\xi_k} \prod_{j=1}^{k-1} \frac{1-\xi_j^* z}{z-\xi_j}, \tag{24.18}$$

which forms a complete set in $\mathcal{H}_2(\mathbb{E})$ if

$$\sum_{j=1}^{\infty} 1 - |\xi_j| = \infty. \tag{24.19}$$

24.3 Least Squares Estimation

A natural extension of a AR model is to use the model structure

$$[H(z)]^{-1} = 1 + \sum_{k=1}^{n} f_k F_k(z), \tag{24.20}$$

where $\{F_k(z)\}$ is a set of orthonormal rational functions with pre-specified poles. This is nothing but a way to represent an ARMA model with fixed zeros, i.e.

$$H(z) = \frac{C_*(z)}{A(z)}, \tag{24.21}$$

where $C_*(z)$ is specified by the poles of $\{F_k(z)\}$, i.e. the zeros of $H(z)$. Consider the filtered process

$$y_c(t) = \frac{1}{C_*(q)} y(t). \tag{24.22}$$

The problem of estimating a rational orthogonal model is completely equivalent to estimating an AR model for the process $\{y_c(t)\}$. The obvious question is why one should study a more complex model structure then the simple AR. There are several answers. One is that this structure leads to much more robust

filter implementations, and numerically better scaled estimation problems. This structure also allows for simpler analysis as will be shown below.

Define the regression vector

$$\varphi(t) = [-F_1(q)y(t) \dots - F_n(q)y(t)]^T. \tag{24.23}$$

Given observations $\{y(1) \dots y(N)\}$, the least squares estimate of the parameter vector $\theta = (f_1 \dots f_n)^T$ is given by

$$\hat{\theta} = R_N^{-1} f_N, \quad R_N = \frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t), \quad f_N = \frac{1}{N}\sum_{t=1}^{N} \varphi(t)y(t). \tag{24.24}$$

Let $R = \mathrm{E}\{R_N\}$, i.e. the covariance matrix of the regression vector.

The numerical properties of the least squares problem depend on the condition number of R. The optimal case would be if R equals the identity matrix. This is the case for orthonormal rations functions models if $y(t)$ equals white noise, which of course is not of interest. For the AR case, $F_k(z) = z^{-k}$, the covariance matrix has a Toeplitz structure and a classical result is

$$\min_{\omega} \Phi(e^{i\omega}) \leq \mathrm{eig}\{R\} \leq \max_{\omega} \Phi(e^{i\omega}), \tag{24.25}$$

where $\Phi(e^{i\omega})$ denotes the power spectral density of $y(t)$. It can be shown that the same result holds for a general orthonormal rational function model, see [6]. The covariance matrix is also of interest in order to determine the statistical properties of the estimate since

$$\mathrm{Var}\{\hat{\theta}\} \approx \frac{\sigma_0^2}{N} R^{-1}, \tag{24.26}$$

where the expression is asymptotic in the number of data N, and the bias due to model errors is neglected.

The sensitivity of the parameter vector θ as such is often of secondary interest. A more invariant measure is the variance of the estimated frequency function

$$\mathrm{Var}\{[\hat{H}(e^{i\omega})]^{-1}\} \approx \frac{\sigma_0^2}{N}[F_1(e^{i\omega}) \dots F_n(e^{i\omega})]R^{-1}[F_1(e^{i\omega}) \dots F_n(e^{i\omega})]^*. \tag{24.27}$$

It is possible to derive a more explicit expression for the following cases:

For large model orders n we have

$$\mathrm{Var}\{[\hat{H}(e^{i\omega})]^{-1}\} \approx \frac{1}{N}\frac{\sum_{k=1}^{n} |F_k(e^{i\omega})|^2}{|H_0(e^{i\omega})|^2}. \tag{24.28}$$

From (24.18)

$$|F_k(e^{i\omega})|^2 = \frac{1 - |\xi_i|^2}{|e^{i\omega} - \xi_k|^2},$$

where $\{\xi_k\}$ are the zeros of $H(z)$. See e.g. [6].

If the zeros of $H_0(z)$ is known to belong to a given set, we have the following result. Assume that

$$H_0(z) = \frac{C_0(z)}{A_0(z)}, \quad H(z) = \frac{C_1(z)C_0(z)}{A(z)}, \tag{24.29}$$

where degree$[H(z)] = n \geq$ degree$[H_0(z)] = n_0$. Then

$$\mathrm{Var}\{[\hat{H}(e^{i\omega})]^{-1}\} \approx \frac{1}{N} \frac{\sum_{k=1}^{n_0} |F_k^0(e^{i\omega})|^2 + \sum_{j=1}^{n-n_0} |F_j^1(e^{i\omega})|^2}{|H_0(e^{i\omega})|^2}. \tag{24.30}$$

Here $\{F_k^0(z)\}$ are the orthogonal basis functions constructed using the poles of $H_0(z)$, and the $n - n_0$ basis functions $\{F_k^1(z)\}$ are constructed from the zeros $C_1(z)$. This result is a slight generalization of the AR variance expression given in [5]. The idea of proof is rather simple. Rewrite the prediction error problem as

$$y(t) = (1 - [H(q)]^{-1})y(t) + \varepsilon(t) \tag{24.31}$$

to obtain the artificial problem of estimating the transfer function of the system

$$y(t) = \frac{(C(q) - A(q))C_0(z)}{C(q)A^0(q)} e(t) + \varepsilon(t), \tag{24.32}$$

where $e(t)$ is a given white noise input signal with variance σ_0^2, $\varepsilon(t)$ is white measurement noise with variance σ_0^2, and $y(t)$ is the output signal. The transfer function $G(z) = (C(z) - A(z))/C_1(z)A^0(z)$ will have relative degree one with prespecified poles. The corresponding orthogonal basis functions can be chosen as $\{F_k(z) = F_k^0(z)\}$, $k = 1 \ldots n_0$ corresponding the the roots of $A_0(z) = 0$, and $F_{j+n_0}(z) = F_j^1(z)H_b(z)$, $j = 1 \ldots n - n_0$, where $H_b(z)$ is the all-pass transfer function with the same poles as $H_0(z)$ and $\{F_k^1(z)\}$ are constructed from $C_1(z)$. Since $e(t)$ is white noise with variance σ_0^2, the covariance matrix of the regression vector for this estimation problem equals $R = \sigma_0^2 I$ and

$$\mathrm{Var}\{\hat{G}(e^{i\omega})\} \approx \frac{\sum_{k=1}^{n_0} |F_k^0(e^{i\omega})|^2 + \sum_{j=1}^{n-n_0} |F_j^1(e^{i\omega})|^2}{N}. \tag{24.33}$$

Now $[H(e^{i\omega})]^{-1} = [1 - G(e^{i\omega})][H_0(e^{i\omega})]^{-1}$ and thus

$$\mathrm{Var}\{[\hat{H}(e^{i\omega})]^{-1}\} = \frac{\mathrm{Var}\{\hat{G}(e^{i\omega})\}}{|H_0(e^{i\omega})|^2}, \tag{24.34}$$

which gives (24.30). By setting $C_1(z) = z^{n-n_0}$ we obtain Theorem 5.1 in [5].

If the spectral density of $y(t)$ was known one could use basis functions that are orthonormal for the weighted scalar product

$$< F_1^w, F_2^w >_w = \frac{1}{2\pi i} \oint_{\mathbb{T}} F_1^w(1/z) F_2^w(z) \Phi(z) \frac{dz}{z}. \tag{24.35}$$

This is very closely related to using a feedforward lattice model. Since R then equals the unity matrix, we directly obtain

$$\mathrm{Var}\{[\hat{H}(e^{i\omega})]^{-1}\} \approx \frac{\sigma_0^2}{N}\sum_{k=1}^{n}|F_k^w(e^{i\omega})|^2. \tag{24.36}$$

Here we have neglected bias errors. Even if its easy to calculate the basis functions $F_k^w(z)$, it is more difficult to explicitly relate them to the poles and zeros of the model and the system.

To conclude: We shown how the choice of zeros will influence the variance of the estimator. The best choice is of course to take the true zeros of the process, i.e. $C(z) = C_0(z)$. The other extreme is to use a high order AR model with all zeros at $z = 0$. Orthogonal rational function models give a compromise between these extremes.

Next, we will list how some more results for AR estimation are generalized to spectral estimation using orthogonal rational functions models. The asymptotic, as the number of data tends to infinity, prediction error method estimate minimizes the cost function

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}|H(e^{i\omega})|^{-2}\Phi(e^{i\omega})d\omega. \tag{24.37}$$

Denote the corresponding minimum by σ^2. As we have shown the orthogonal rational function model approach is theoretically equivalent to AR estimation of the process

$$y_c(t) = \frac{1}{C_*(q)}y(t), \quad [H(z)]^{-1} = 1 + \sum_{k=1}^{n} f_k F_k(z) = \frac{A(z)}{C_*(z)}.$$

It is well known that the AR estimate will converge to an stable filter as the number of data tends to infinity. This means that $[\hat{H}(z)]^{-1} = 1 + \hat{a}_k F_k(z)$ will converge to a stable and minimum phase transfer function.

It is also well known that the covariance function of the limiting AR estimate perfectly fits the n first covariance values of the underlying process. Let $r_c(\tau) = \mathrm{E}\{y_c(t)y_c(t+\tau)\}$, $\tau = 0, \ldots n$. Hence, we solve the following covariance matching problem

$$\begin{aligned} r_c(\tau) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega\tau}\frac{\sigma^2}{|A(e^{i\omega})|^2}d\omega, \quad \tau = 0\ldots n \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{i\omega\tau}}{|C_*(e^{i\omega}|^2}\sigma^2|H(e^{i\omega})|^2 d\omega, \quad \tau = 0\ldots n. \end{aligned} \tag{24.38}$$

24.4 The Covariance Extension Problem

Consider the covariance lags

$$r(\tau) = \mathrm{E}\{y(t+\tau)y(t)\}, \tag{24.39}$$

and the corresponding spectral density

$$\Phi(e^{i\omega}) = \sum_{k=-\infty}^{\infty} r(\tau)e^{-i\omega\tau}, \quad r(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega\tau}\Phi(e^{i\omega})d\omega.$$

Assume an ARMA model $H(z) = C(z)/A(z)$ with innovation variance σ^2, and let

$$Q(z) = A(z)A(1/z)/\sigma^2 = q_0 + q_1(z + z^{-1}) \ldots + q_n(z^n + z^{-n}). \tag{24.40}$$

Notice that $Q(z) > 0$ on the unit circle, i.e. a positive function. An ARMA process with degree constraint and covariances $r(\tau)$, $\tau = 0 \ldots n$ must satisfy

$$r(\tau) - \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega\tau}\frac{|C(e^{i\omega})|^2}{Q(e^{i\omega})}d\omega = 0, \quad \tau = 0, \ldots n \tag{24.41}$$

But this is just the derivative of the cost-function:

$$V(q) = [r(0) \ldots r(n)]q - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log Q(e^{i\omega})|C(e^{i\omega})|^2 d\omega \tag{24.42}$$

with respect to $q = (q_0 \ldots q_n)^T$. The main result of Byrnes, Lindquist and co-workers is that $V(q)$ is a convex function. Hence we have only need to solve a convex optimization problem to find the solution of the covariance realization problem! The first term

$$[r(0) \ldots r(n)]q = \frac{1}{2\pi}\int_{-\pi}^{\pi} Q(e^{i\omega})\Phi(e^{i\omega})d\omega \tag{24.43}$$

is just the AR prediction error cost function (which is quadric in A). The second term

$$-\frac{1}{2\pi}\int_{-\pi}^{\pi} \log Q(e^{i\omega})|C(e^{i\omega})|^2 d\omega \tag{24.44}$$

can be viewed a barrier function to impose the positivity constraint. We refer to [4] for an excellent overview of the convex optimization approach to the rational covariance extension problem.

A natural question is why should one be interested in perfect fit of the n first covariances? It is well known that a fixed number of covariances is not a sufficient statistics for an ARMA model. As shown in Byrnes et. al. [3] it is possible to extend the covariance fitting problem to certain filter banks. Using a similar idea we will finally sketch on an extensions of this problem to the prediction error framework. As discussed in the previous section, the prediction error method approach gives perfect fit to the first $n + 1$ covariances of $y_c(t)$, and corresponds to what is called the central solution. Using the method of Lindquist et. al. one could as well determine an ARMA model $C(z)/A(z)$ of order n which fits $r_c(\tau)$, by solving

$$r_c(\tau) - \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega\tau}\frac{|C(e^{i\omega})|^2}{|C_*(e^{i\omega})|^2 Q(e^{i\omega})}d\omega = 0, \quad \tau = 0, \ldots n \tag{24.45}$$

The same trick applies here, i.e. integration w.r.t. q gives the convex cost function

$$V_c(q) = [r_c(0) \dots r_c(n)]q - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log Q(e^{i\omega}) \frac{|C(e^{i\omega})|^2}{|C_*(e^{i\omega})|^2} d\omega, \tag{24.46}$$

which can be used to calculate the model $H(z) = C(z)/A(z)$. The choice $C(z) = C_*(z)$ gives back the prediction error solution, while using other $C(z)$ allows for different extensions. The first term is nothing but the standard prediction error, and it correspond to

$$[r_c(0) \dots r_c(n)]q = \mathrm{E}\{([H(q)]^{-1}y(t))^2\}, \quad [H(z)]^{-1} = 1 + \sum_{k=1}^{n} f_k F_k(z) \tag{24.47}$$

with repect to $(f_1 \dots f_n)$. It seems more difficult to express the second term using $\{F_k(z)\}$. It is, however, possible to represent

$$\frac{C(z)}{C_*(z)} = 1 + \sum_{k=1}^{n} \bar{f}_k F_k(z). \tag{24.48}$$

It would be of interest to investigate if this representation leads to better numerical properties.

24.5 Conclusion and Future Work

Models are always approximations of true data generating process. The quality of a model depends heavily on its intended use. If the objective is prediction, the prediction error approach is optimal. If one is interested in spectral properties in certain frequency band the answer is more difficult. The covariance extension approach and its generalizations provide promising methods for this case. The objective of this paper has been to discuss some relation between these two approaches to spectral estimation, and to show that pre-specified zeros indeed are a valuable tool to tune a spectral estimators.

24.6 References

[1] J. Bokor, P. Heuberger, B. Ninness, T. Oliveira e Silva, P. Van den Hof, and B. Wahlberg. Modelling and identification with orthogonal basis functions. In *Workshop Notes, 14:th IFAC World Congress, Workshop nr 6*, Beijing, PRC, July 1999.

[2] A. Bultheel, P. González-Vera, and O. Njåstad. *Orthogonal Rational Functions*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 1999.

[3] C.I. Byrnes, T. T. Geeorgiou, and A. Lindquist. A new approach to spectral estimation: A tunable high-resolution spectral estimator. *IEEE Transactions on Signal Processing*, 48(11):3189–3205, November 2001.

[4] C.I. Byrnes, S.V. Gusev, and A. Lindquist. From finite windows to modeling filters: A convex optimization approach. *SIAM Review*, 43(4):645–675, 2001.

[5] L.-L. Lie and L. Ljung. Asymptotic variance expressions for estimated frequency functions. *IEEE Trans. Autom. Control*, 46(12):1887–1899, 2001.

[6] B.M. Ninness, H. Hjalmarsson, and F. Gustafsson. Generalised Fourier and Toeplitz results for rational orthonormal bases. *SIAM Journal on Control and Optimization*, 37(2):429–460, 1999.

[7] R.A. Roberts and C.T. Mullis. *Digital Signal Processing*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1987.

[8] B. Wahlberg. Orthonormal rational functions: A transformation analysis. *SIAM Review*, 2002. Accepted for publication.

25

Balanced State Representations with Polynomial Algebra

Jan C. Willems *Paolo Rapisarda*

Abstract

Algorithms are derived that pass directly from the differential equation describing the behavior of a finite-dimensional linear system to a balanced state representation.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 345-357, 2003.
© Springer-Verlag Berlin Heidelberg 2003

25.1 Introduction

The algorithms for model reduction are among the most useful achievements of linear system theory. A low order model that incorporates the important features of a high order one offers many advantages: it reduces the computational complexity, it filters out irrelevant details, it smooths the data, etc. Two main classes of algorithms for model reduction have been developed: (i) model reduction by *balancing*, and (ii) model reduction in the *Hankel norm* (usually called *AAK model reduction*). The implementation of these algorithms typically starts from the finite-dimensional state space system

$$\frac{d}{dt}x = Ax + Bu, y = Cx + Du,$$

commonly denoted as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}. \tag{25.1}$$

In the context of model reduction, it is usually assumed that this system is minimal (i.e., controllable and observable) and stable (i.e., the matrix A is assumed to be *Hurwitz*, meaning that all its eigenvalues have negative real part).

However, a state space system is seldom the end product of a first principles modeling procedure. Typically one obtains models involving a combination of (many) algebraic equations, (high order) differential equations, transfer functions, auxiliary variables, etc. Since model reduction procedures aim at systems of high dynamic complexity, it may not be an easy matter to transform the first principles model to state form. It is therefore important to develop algorithms that pass directly from model classes different from state space models to reduced models, without passing through a state representation.

There are, in fact, some interesting subtle algorithms that do exactly this for AAK model reduction in Fuhrmann's book [1]. These algorithms form the original motivation and inspiration for the present article. Its purpose is to present an algorithm for the construction of a balanced state representation directly from the differential equation (or the transfer function) that governs the system. For simplicity of exposition, we restrict attention in this paper to single-input/single-output systems.

A few words about the notation. We use the standard notation $\mathbb{R}$, $\mathbb{R}^{\mathtt{n}}$, $\mathbb{R}^{\mathtt{n}_1 \times \mathtt{n}_2}$, for the reals, the set of $\mathtt{n}$-dimensional real vectors, the set of $\mathtt{n}_1 \times \mathtt{n}_2$-dimensional real matrices. $M = [m(\mathtt{i},\mathtt{j})]_{\mathtt{i}=1,\ldots,\mathtt{n}_1}^{\mathtt{j}=1,\ldots,\mathtt{n}_2}$, denotes the $\mathtt{n}_1 \times \mathtt{n}_2$-matrix whose $(\mathtt{i},\mathtt{j})$—th element is $m(\mathtt{i},\mathtt{j})$, with an analogous notation $[m(\mathtt{j})]^{\mathtt{j}=1,\ldots,\mathtt{n}}$ for row and $[m(\mathtt{i})]_{\mathtt{i}=1,\ldots,\mathtt{n}}$ for column vectors. The ring of real one-variable polynomials in the indeterminate ξ is denoted by $\mathbb{R}[\xi]$, and the set of real two-variable polynomials in the indeterminates ζ, η is denoted by $\mathbb{R}[\zeta,\eta]$. $\mathbb{R}_{\mathtt{n}}[\xi]$ denotes the $(\mathtt{n}+1)$-dimensional real vector space consisting of the real polynomials of degree less that or equal to $\mathtt{n}$. $\mathfrak{L}_2^{\mathrm{loc}}(\mathbb{R},\mathbb{R})$ denotes the set of maps $f : \mathbb{R} \to \mathbb{R}$ that are locally square integrable, i.e., such that $\int_{t_1}^{t_2} |f(t)|^2\, dt < \infty$ for all $t_1, t_2 \in \mathbb{R}$; $\mathcal{L}_2(A,\mathbb{R})$ denotes the set of maps $f : A \to \mathbb{R}$ such that $||f||^2_{\mathcal{L}_2(A,\mathbb{R})} := \int_A |f(t)|^2\, dt < \infty$. $\mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ denotes the set of infinitely differentiable maps from $\mathbb{R}$ to $\mathbb{R}$, $\mathfrak{E}^+(\mathbb{R},\mathbb{R}) := \{w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}) \mid w|_{(-\infty,0]}$ has compact support$\}$, and $\mathfrak{D}(\mathbb{R},\mathbb{R})$ denotes the set of real distributions

on $\mathbb{R}$. Analogous notation is used for $\mathbb{R}$ replaced by the field of complex numbers $\mathbb{C}$. * denotes complex conjugation for elements of $\mathbb{C}$, Hermitian conjugate (conjugate transpose) for complex matrices, or, more generally, 'dual'.

25.2 The System Equations

Our starting point is the continuous-time single-input/single-output finite-dimensional linear time-invariant system described by the differential equation

$$p(\frac{d}{dt})y = q(\frac{d}{dt})u, \tag{25.2}$$

relating the input $u : \mathbb{R} \to \mathbb{R}$ to the output $y : \mathbb{R} \to \mathbb{R}$. The polynomials $p, q \in \mathbb{R}[\xi]$ parametrize the system behavior, formally defined as

$$\mathfrak{B}_{(p,q)} := \{(u, y) \in \mathfrak{L}_2^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}) \times \mathfrak{L}_2^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}) \mid (25.2) \text{ holds in the sense of distributions}\}.$$

In the sequel, we will identify the system (25.2) with its behavior $\mathfrak{B}_{(p,q)}$.

The system $\mathfrak{B}_{(p,q)}$ is said to be *controllable* if for all $(u_1, y_1), (u_2, y_2) \in \mathfrak{B}_{(p,q)}$ there exists $T > 0$ and $(u, y) \in \mathfrak{B}_{(p,q)}$ such that $(u_1, y_1)(t) = (u, y)(t)$ for $t \leq 0$ and that $(u_2, y_2)(t) = (u, y)(t + T)$ for $t > 0$. It is well-known (see [5]) that the system $\mathfrak{B}_{(p,q)}$ is controllable if and only if the polynomials p and q are co-prime (i.e., they have no common roots). It turns out that controllability is also equivalent to the existence of an *image representation* for $\mathfrak{B}_{(p,q)}$, meaning that the *manifest behavior* of the *latent variable system*

$$u = p(\frac{d}{dt})\ell, y = q(\frac{d}{dt})\ell, \tag{25.3}$$

formally defined as

$$\mathfrak{Im}_{(p,q)} := \{(u, y) \in \mathfrak{L}_2^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}) \times \mathfrak{L}_2^{\mathrm{loc}}(\mathbb{R}, \mathbb{R}) \mid \text{there exists } \ell \in \mathfrak{D}(\mathbb{R}, \mathbb{R}) \text{ such that (25.3) holds in the sense of distributions}\}$$

is *exactly* equal to $\mathfrak{B}_{(p,q)}$. In (25.3), we refer to ℓ as the *latent variable*.

Assume throughout that $p, q \in \mathbb{R}[\xi]$ are co-prime, with degree$(q) \leq$ degree(p) $=: \mathtt{n}$. Co-primeness of p and q ensures, in addition to controllability of $\mathfrak{B}_{(p,q)}$, *observability* of the image representation $\mathfrak{Im}_{(p,q)}$, meaning that, for every $(u, y) \in \mathfrak{Im}_{(p,q)} = \mathfrak{B}_{(p,q)}$, the $\ell \in \mathfrak{D}(\mathbb{R}, \mathbb{R})$ such that $u = p(\frac{d}{dt})\ell, y = q(\frac{d}{dt})\ell$, is unique.

In addition to expressing controllability, image representations are also useful for state construction (see [6] for an in-depth discussion). For the case at hand, it turns out that any set of polynomials $\{x_1, x_2, \ldots, x_{\mathtt{n}'}\}$ that span $\mathbb{R}_{\mathtt{n}-1}[\xi]$ defines a state representation of $\mathfrak{B}_{(p,q)}$ with state

$$x = (x_1(\frac{d}{dt})\ell, x_2(\frac{d}{dt})\ell, \ldots, x_{\mathtt{n}'-1}(\frac{d}{dt})\ell),$$

i.e., the manifest behavior of

$$u = p(\frac{d}{dt})\ell, y = q(\frac{d}{dt})\ell, x = \mathrm{col}(x_1(\frac{d}{dt})\ell, x_2(\frac{d}{dt})\ell, \ldots, x_{n'-1}(\frac{d}{dt})\ell) \tag{25.4}$$

satisfies the axiom of state (see [6] for a formal definition of the axiom of state). The associated system matrices (25.1) are then obtained as a solution matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ of the following system of linear equations in $\mathbb{R}_n[\xi]$:

$$\begin{bmatrix} \xi x_1(\xi) \\ \xi x_2(\xi) \\ \vdots \\ \xi x_{n'}(\xi) \\ q(\xi) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x_1(\xi) \\ x_2(\xi) \\ \vdots \\ x_{n'}(\xi) \\ p(\xi) \end{bmatrix}. \tag{25.5}$$

This state representation is minimal if and only if $n' = n$ and hence the polynomials $x_1, x_2, \ldots, x_n$ form a basis for $\mathbb{R}_{n-1}[\xi]$. Henceforth, we will concentrate on the minimal case, and put $n = n'$. Note that in this case the solution of (25.5) is unique.

The n-th order system (25.1), assumed minimal (i.e., controllable and observable) and stable, is called *balanced* if there exist real numbers

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0,$$

called the *Hankel singular values*, such that, with

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n),$$

there holds

$$\begin{aligned} A\Sigma + \Sigma A^\top + BB^\top &= 0 \\ A^\top\Sigma + \Sigma A + C^\top C &= 0. \end{aligned}$$

Of course, in the context of the state construction through an image representation as explained above, being balanced becomes a property of the polynomials $x_1, x_2, \ldots, x_n$. The central problem of this paper is:

> Choose the polynomials $x_1, x_2, \ldots, x_n$ so that (25.5) defines a balanced state space system.

25.3 The Controllability and Observability Gramians

In order to solve this problem, we need polynomial expressions for the controllability and observability gramians. These are actually quadratic differential forms (QDF's) (see [7] for an in-depth study of QDF's). The real two-variable polynomial

$$\Phi(\zeta, \eta) = \Sigma_{i,j} \Phi_{i,j} \zeta^i \eta^j$$

induces the map

$$w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}) \mapsto \Sigma_{\mathtt{i},\mathtt{j}} \frac{d^{\mathtt{i}}}{dt^{\mathtt{i}}} w \; \Phi_{\mathtt{i},\mathtt{j}} \; \frac{d^{\mathtt{j}}}{dt^{\mathtt{j}}} w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R}).$$

This map is called a *a quadratic differential form* (QDF). Denote it as Q_Φ. In view of the quadratic nature of this map, we will always assume that $\Phi_{\mathtt{i},\mathtt{j}} = \Phi_{\mathtt{j},\mathtt{i}}$, i.e., that Φ is *symmetric*, i.e., $\Phi = \Phi^*$, with $\Phi^*(\zeta,\eta) := \Phi(\eta,\zeta)$.

The derivative $\frac{d}{dt} Q_\Phi(w)$ of the QDF Q_Φ is again a QDF, $Q_\Psi(w)$, with $\Psi(\zeta,\eta) = (\zeta+\eta)\Phi(\zeta,\eta)$. It readily follows that a given QDF Q_Φ is the derivative of another QDF if and only if $\partial(\Phi) = 0$, where $\partial : \mathbb{R}[\zeta,\eta] \to \mathbb{R}[\xi]$ is defined by $\partial(\Phi)(\xi) := \Phi(\xi,-\xi)$, in which case the QDF Q_Ψ such that $Q_\Phi(w) = \frac{d}{dt} Q_\Phi(w)$ is induced by $\Psi(\zeta,\eta) = \dfrac{\Phi(\zeta,\eta)}{\zeta+\eta}$. Note that this is a polynomial since $\partial(\Phi) = 0$.

Every QDF Q_Φ can de written as the sum and difference of squares, i.e., there exist $f_1^+, f_2^+, \ldots, f_{\mathtt{n}_+}^+, f_1^-, f_2^-, \ldots, f_{\mathtt{n}_-}^- \in \mathbb{R}[\xi]$ such that

$$Q_\Phi(w) = \Sigma_{\mathtt{k}=1}^{\mathtt{n}_+} |f_{\mathtt{k}}^+(\frac{d}{dt})w|^2 - \Sigma_{\mathtt{k}=1}^{\mathtt{n}_-} |f_{\mathtt{k}}^-(\frac{d}{dt})w|^2.$$

Equivalently,

$$\Phi(\zeta,\eta) = \Sigma_{\mathtt{k}=1}^{\mathtt{n}_+} f_{\mathtt{k}}^+(\zeta) f_{\mathtt{k}}^+(\eta) - \Sigma_{\mathtt{k}=1}^{\mathtt{n}_-} f_{\mathtt{k}}^-(\zeta) f_{\mathtt{k}}^-(\eta).$$

If $f_1^+, f_2^+, \ldots, f_{\mathtt{n}_+}^+, f_1^-, f_2^-, \ldots, f_{\mathtt{n}_-}^- \in \mathbb{R}[\xi]$ are linearly independent over $\mathbb{R}$, then $\mathtt{n}_+ + \mathtt{n}_-$ is the *rank* and $\mathtt{n}_+ - \mathtt{n}_-$ the *signature* of Q_Φ (or Φ). The QDF Q_Φ is said to be *non-negative* if $Q_\Phi(w) \geq 0$ for all $w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$. Equivalently, if and only if the rank of Φ is equal to its signature.

While it would be natural to consider the controllability and observability gramians as QDF's on $\mathfrak{B}_{(p,q)}$, we will consider them as QDF's acting on the latent variable ℓ of the image representation (25.3). This entails no loss of generality, since there is a one-to-one relation between ℓ in (25.3) and $(u,y) \in \mathfrak{B}_{(p,q)}$.

The *controllability gramian* Q_K (equivalently, K) is defined as follows. Let $\ell \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ and define $Q_K(\ell)$ by

$$Q_K(\ell)(0) := \text{infimum} \int_{-\infty}^0 |p(\frac{d}{dt})\ell'(t)|^2 \, dt,$$

where the infimum is taken over all $\ell' \in \mathfrak{C}^+(\mathbb{R},\mathbb{R})$ that join ℓ at $t = 0$, i.e., such that $\ell(t) = \ell'(t)$ for $t \geq 0$.

The *observability gramian* Q_W (equivalently, W) is defined as follows. Let $\ell \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ and define $Q_W(\ell)$ by

$$Q_W(\ell)(0) := \int_0^\infty |q(\frac{d}{dt})\ell'(t)|^2 \, dt,$$

where $\ell' \in \mathfrak{D}(\mathbb{R},\mathbb{R})$ is such that

$$\begin{aligned}
&\text{(i)} \quad \ell|_{(-\infty,0)} = \ell'|_{-\infty,0)},\\
&\text{(ii)} \quad (p(\frac{d}{dt})\ell', q(\frac{d}{dt})\ell') \in \mathfrak{B}_{(p,q)},\\
&\text{(iii)} \quad p(\frac{d}{dt})\ell'(t)|_{(0,\infty)} = 0.
\end{aligned}$$

Thus ℓ' is a latent variable trajectory that continues ℓ at $t = 0$ with an ℓ' such that $u|_{(0,\infty)} = p(\frac{d}{dt})\ell'|_{(0,\infty)} = 0$. This continuation must be sufficiently smooth so that the resulting $(u, y) = (p(\frac{d}{dt})\ell', q(\frac{d}{dt})\ell')$ belongs to $\mathfrak{B}_{(p,q)}$, thus in particular to $\mathfrak{L}_2^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$. This actually means that the $(\mathtt{n} - 1)$-th derivative of ℓ' must be in $\mathfrak{L}_2^{\mathrm{loc}}(\mathbb{R}, \mathbb{R})$.

The computation of the two-variable polynomials K and W is one of the central results of this paper.

THEOREM 25.1
Consider the system $\mathfrak{B}_{(p,q)}$ with $p, q \in \mathbb{R}[\xi]$, degree$(q) \leq$ degree$(p) =: \mathtt{n}$, while p, q are co-prime and p Hurwitz (meaning that all its roots have negative real part). The controllability gramian and the observability gramian are QDF's. Denote them by Q_K and Q_W respectively, with $K \in \mathbb{R}[\zeta, \eta]$ and $W \in \mathbb{R}[\zeta, \eta]$. They can be computed as follows

$$\boxed{K(\zeta, \eta) = \frac{p(\zeta)p(\eta) - p(-\zeta)p(-\eta)}{\zeta + \eta},} \tag{25.6}$$

$$\boxed{W(\zeta, \eta) = \frac{p(\zeta)f(\eta) + f(\zeta)p(\eta) - q(\zeta)q(\eta)}{\zeta + \eta},} \tag{25.7}$$

with $f \in \mathbb{R}_{\mathtt{n}-1}[\xi]$ the (unique) solution of the Bezout-type equation

$$\boxed{p(\xi)f(-\xi) + f(\xi)p(-\xi) - q(\xi)q(-\xi) = 0.} \tag{25.8}$$

Moreover, both Q_K and Q_W are nonnegative and have rank $\mathtt{n}$. Finally, $Q_K(\ell)$ and $Q_W(\ell)$ only contain the derivatives $\ell, \frac{d}{dt}\ell, \ldots, \frac{d^{\mathtt{n}-1}}{dt^{\mathtt{n}-1}}\ell$. □

The proof of this theorem is given in the appendix (section 25.6).

Note that the equation for f has a unique solution in $\mathbb{R}_{\mathtt{n}-1}[\xi]$ since $p(\xi)$ and $p(-\xi)$ are co-prime, a consequence of the fact that p is Hurwitz.

What we call the controllability gramian measures the *difficulty* it takes to join the latent variable trajectory ℓ at $t = 0$ by a trajectory ℓ' that is zero in the far past, as measured by

$$\int_{-\infty}^{0} |u(t)|^2 \, dt = \int_{-\infty}^{0} |p(\frac{d}{dt})\ell'(t)|^2 \, dt.$$

The observability gramian on the other hand measures the *ease* with which it is possible to observe the effect of the latent variable trajectory ℓ as measured by

$$\int_{0}^{+\infty} |y(t)|^2 \, dt = \int_{0}^{+\infty} |q(\frac{d}{dt})\ell'(t)|^2 \, dt,$$

assuming that the input $u = p(\frac{d}{dt})\ell'(t)$ is zero for $t \geq 0$. Note the slight difference with the classical terminology where the controllability gramian corresponds to the 'inverse' of the QDF Q_K.

25.4 Balanced State Representation

The minimal state representation (25.4) with state polynomials $(x_1, x_2, \ldots, x_n)$ is *balanced* if

(i) for $\ell_i \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R})$ such that $x_j(\frac{d}{dt})\ell_i(0) = \delta_{ij}$ (δ_{ij} denotes the Kronecker delta), we have

$$Q_W(\ell_i)(0) = \frac{1}{Q_K(\ell_i)(0)},$$

i.e., the state components that are difficult to reach are also difficult to observe, and

(ii) the state components are ordered so that

$$0 < Q_K(\ell_1)(0) \leq Q_K(\ell_2)(0) \leq \cdots \leq Q_K(\ell_n)(0),$$

and hence

$$Q_W(\ell_1)(0) \geq Q_W(\ell_2)(0) \geq \cdots \geq Q_W(\ell_n)(0) > 0.$$

The general state construction (25.4) and a suitable factorization of the controllability and observability gramians readily lead to a balanced state representation.

It is a standard result from linear algebra (see [2], chapter 9) that theorem 25.1 implies that there exist polynomials

$$(x_1^{bal}, x_2^{bal}, \ldots, x_n^{bal})$$

that form a basis for $\mathbb{R}_{n-1}[\xi]$, and real numbers

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0 \tag{25.9}$$

(the σ_k's are uniquely defined by K and W) such that

$$K(\zeta, \eta) = \Sigma_{k=1}^n \sigma_k^{-1} x_k^{bal}(\zeta) x_k^{bal}(\eta), \tag{25.10}$$

$$W(\zeta, \eta) = \Sigma_{k=1}^n \sigma_k \; x_k^{bal}(\zeta) x_k^{bal}(\eta). \tag{25.11}$$

This leads to the main result of this paper.

THEOREM 25.2
Define the polynomials $(x_1^{bal}, x_2^{bal}, \ldots, x_n^{bal}) \in \mathbb{R}_{n-1}[\xi]$ and the real numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ by equations (25.9, 25.10, 25.11). Then the σ_k's are the Hankel singular values of the system $\mathfrak{B}_{(p,q)}$ and

$$u = p(\frac{d}{dt})\ell, y = q(\frac{d}{dt})\ell, x^{bal} = (x_1^{bal}(\frac{d}{dt})\ell, x_2^{bal}(\frac{d}{dt})\ell, \ldots, x_n^{bal}(\frac{d}{dt})\ell)$$

is a balanced state space representation of $\mathfrak{B}_{(p,q)}$. The associated balanced system matrices are obtained as the solution matrix $\begin{bmatrix} A^{bal} & B^{bal} \\ C^{bal} & D^{bal} \end{bmatrix}$ of the following system

of linear equations in $\mathbb{R}_{\mathtt{n}}[\xi]$:

$$\begin{bmatrix} \xi x_1^{\mathrm{bal}}(\xi) \\ \xi x_2^{\mathrm{bal}}(\xi) \\ \vdots \\ \xi x_{\mathtt{n}}^{\mathrm{bal}}(\xi) \\ q(\xi) \end{bmatrix} = \begin{bmatrix} A^{\mathrm{bal}} & B^{\mathrm{bal}} \\ C^{\mathrm{bal}} & D^{\mathrm{bal}} \end{bmatrix} \begin{bmatrix} x_1^{\mathrm{bal}}(\xi) \\ x_2^{\mathrm{bal}}(\xi) \\ \vdots \\ x_{\mathtt{n}}^{\mathrm{bal}}(\xi) \\ p(\xi) \end{bmatrix}. \tag{25.12}$$

□

The proof of this theorem is given in the appendix (section 25.6).

We summarize this algorithm:

DATA: $p, q \in \mathbb{R}[\xi]$, co-prime, degree$(q) \leq$ degree$(p) := \mathtt{n}$, p Hurwitz.

COMPUTE:

(i) $K \in \mathbb{R}[\zeta, \eta]$ by (25.6),

(ii) $f \in \mathbb{R}_{\mathtt{n}-1}[\xi]$ by (25.8) and $W \in \mathbb{R}[\zeta, \eta]$ by (25.7),

(iii) $(x_1^{\mathrm{bal}}, x_2^{\mathrm{bal}}, \ldots, x_{\mathtt{n}}^{\mathrm{bal}})$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\mathtt{n}} > 0$ by the expansions (25.9, 25.10, 25.11):

$$K(\zeta, \eta) = \Sigma_{k=1}^{\mathtt{n}} \sigma_{\mathtt{k}}^{-1} x_{\mathtt{k}}^{\mathrm{bal}}(\zeta) x_{\mathtt{k}}^{\mathrm{bal}}(\eta), W(\zeta, \eta) = \Sigma_{\mathtt{k}=1}^{\mathtt{n}} \sigma_{\mathtt{k}} \;\; x_{\mathtt{k}}^{\mathrm{bal}}(\zeta) x_{\mathtt{k}}^{\mathrm{bal}}(\eta),$$

(iv) the balanced system matrices $\begin{bmatrix} A^{\mathrm{bal}} & B^{\mathrm{bal}} \\ C^{\mathrm{bal}} & D^{\mathrm{bal}} \end{bmatrix}$ by solving (25.12).

The result is a balanced state representation of $\mathfrak{B}_{(p,q)}$.

The above algorithm shows how to obtain a balancing-reduced model. Assume that we wish to keep the significant states

$$x_1^{\mathrm{bal}}(\frac{d}{dt})\ell, x_2^{\mathrm{bal}}(\frac{d}{dt})\ell, \ldots, x_{\mathtt{n}_{\mathrm{red}}}^{\mathrm{bal}}(\frac{d}{dt})\ell,$$

and neglect the insignificant ones

$$x_{\mathtt{n}_{\mathrm{red}}+1}^{\mathrm{bal}}(\frac{d}{dt})\ell, x_{\mathtt{n}_{\mathrm{red}}+2}^{\mathrm{bal}}(\frac{d}{dt})\ell, \ldots, x_{\mathtt{n}}^{\mathrm{bal}}(\frac{d}{dt})\ell.$$

Now, solve the following linear equations in the components of the matrices

$$[\, A_{\mathtt{i},\mathtt{j}}^{\mathrm{balred}} \,]_{\mathtt{i}=1,\ldots,\mathtt{n}_{\mathrm{red}}}^{\mathtt{j}=1,\ldots,\mathtt{n}_{\mathrm{red}}}, [\, B_{\mathtt{i}}^{\mathrm{balred}} \,]_{\mathtt{i}=1,\ldots,\mathtt{n}_{\mathrm{red}}}, [\, C_j^{\mathrm{balred}} \,]^{\mathtt{j}=1,\ldots,\mathtt{n}_{\mathrm{red}}}, D^{\mathrm{balred}}$$

$$\xi x_{\mathtt{i}}^{\mathrm{bal}}(\xi) = \Sigma_{\mathtt{j}=1}^{\mathtt{n}_{\mathrm{red}}} A_{\mathtt{i},\mathtt{j}}^{\mathrm{balred}} x_{\mathtt{j}}^{\mathrm{bal}}(\xi) + B_{\mathtt{i}}^{\mathrm{balred}} p(\xi)$$
$$\text{modulo}(x_{\mathtt{n}_{\mathrm{red}}+1}^{\mathrm{bal}}(\xi), x_{\mathtt{n}_{\mathrm{red}}+2}^{\mathrm{bal}}(\xi), \ldots, x_{\mathtt{n}}^{\mathrm{bal}}(\xi))$$

$$q(\xi) = \Sigma_{j=1}^{n_{red}} C_j^{balred} x_j^{bal}(\xi) + D^{balred} p(\xi)$$
$$\text{modulo}(x_{n_{red}+1}^{bal}(\xi), x_{n_{red}+2}^{bal}(\xi), \ldots, x_n^{bal}(\xi)).$$

Then $\begin{bmatrix} A^{balred} & B^{balred} \\ C^{balred} & D^{balred} \end{bmatrix}$ is an n_{red}-th order balancing-reduced state space model for $\mathfrak{B}_{(p,q)}$.

25.5 Comments

Our algorithms for obtaining the controllability and observability gramians and balanced state representations, being polynomial based, offer a number of advantages over the classical matrix based algorithms. In particular, they open up the possibility to involve the know-how on Bezoutians, Bezout and Sylvester matrices and equations, and bring 'fast' polynomial computations to bear on the problem of model reduction.

Instead of computing the σ_k's and the x_k^{bal}'s by the factorization of K, W given by (25.9, 25.10, 25.11), we can also obtain the balanced state representation by evaluating K and W at n points of the complex plane.

Let $\lambda_1, \lambda_2, \cdots, \lambda_n \in \mathbb{C}$ be distinct points of the complex plane. Organize them into the diagonal matrix $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$, and define

$$K_\Lambda = [\, K(\lambda_i^*, \lambda_j)\,]_{i=1,\ldots,n}^{j=1,\ldots,n}$$
$$W_\Lambda = [\, W(\lambda_i^*, \lambda_j)\,]_{i=1,\ldots,n}^{j=1,\ldots,n}$$

Define further

$$X_\Lambda = [\, x_i^{bal}(\lambda_j)\,]_{i=1,\ldots,n}^{j=1,\ldots,n}$$
$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n).$$

There holds

$$K_\Lambda = X_\Lambda^* \Sigma^{-1} X_\Lambda, W_\Lambda = X_\Lambda^* \Sigma X_\Lambda.$$

It is easy to show that, since the λ_k's are distinct and the x_k^{bal}'s form a basis for $\mathbb{R}_{n-1}[\xi]$, X_Λ is non-singular. This implies that X_Λ and Σ can be computed by analyzing the regular pencil formed by the Hermitian matrices K_Λ, W_Λ.

Once X_Λ is known, the balanced state representation is readily computed. However, in order to do so, we need to evaluate K (or W) at one more set of points of the complex plane. Let $\lambda_{n+1} \in \mathbb{C}$ be distinct from the λ_k's. Define $x^{bal}(\lambda_{n+1}) = [\, x_i^{bal}(\lambda_{n+1})\,]_{i=1,\ldots,n}$. The vector $x^{bal}(\lambda_{n+1}) \in \mathbb{C}^n$ can be computed by solving the linear equation

$$K(\lambda_i^*, \lambda_{n+1})) = X_\Lambda^* \Sigma^{-1} x^{bal}(\lambda_{n+1}).$$

Define consecutively

$$\begin{aligned}\Lambda_{\text{ext}} &= \text{diag}(\Lambda, \lambda_{n+1}),\\ X_{\Lambda_{\text{ext}}} &= [\, X_\Lambda \quad x^{\text{bal}}(\lambda_{n+1})\,],\\ p_{\Lambda_{\text{ext}}} &= [\, p_\Lambda \quad p(\lambda_{n+1})\,],\\ q_{\Lambda_{\text{ext}}} &= [\, q_\Lambda \quad q(\lambda_{n+1})\,].\end{aligned}$$

Since the λ_k's are distinct, and $\{x_1^{\text{bal}}, x_2^{\text{bal}}, \cdots, x_n^{\text{bal}}, p\}$ forms a basis for $\mathbb{R}_n[\xi]$, $\begin{bmatrix} X_{\Lambda_{\text{ext}}} \\ p_{\Lambda_{\text{ext}}} \end{bmatrix}$ is also non-singular. The balanced state representation then follows by solving

$$\begin{bmatrix} X_{\Lambda_{\text{ext}}}\Lambda_{\text{ext}} \\ q_{\Lambda_{\text{ext}}} \end{bmatrix} = \begin{bmatrix} A^{\text{bal}} & B^{\text{bal}} \\ C^{\text{bal}} & D^{\text{bal}} \end{bmatrix} \begin{bmatrix} X_{\Lambda_{\text{ext}}} \\ p_{\Lambda_{\text{ext}}} \end{bmatrix}. \tag{25.13}$$

Note that the entries K_Λ follow immediately from (25.6). However, in order to compute the elements of W_Λ from (25.7) it seems unavoidable to have to solve (25.8) for f, at least, it is not clear if it is possible to evaluate the $f(\lambda_k)$'s directly from the $p(\lambda_k)$'s and $q(\lambda_k)$'s.

When we take for the λ_k's, the roots of p, assumed distinct, then f is not needed, and a very explicit expression for K and W is obtained. In this case

$$\begin{aligned}K_\Lambda &= -\left[\frac{p(-\lambda_i^*)p(-\lambda_j)}{\lambda_i^* + \lambda_j}\right]_{i=1,\ldots,n}^{j=1,\ldots,n}\\ W_\Lambda &= -\left[\frac{q(\lambda_i^*)q(\lambda_j)}{\lambda_i^* + \lambda_j}\right]_{i=1,\ldots,n}^{j=1,\ldots,n}\end{aligned}$$

Further, $x^{\text{bal}}(\lambda_{n+1})$ is then obtained from the linear equation

$$-\left[\frac{p(-\lambda_i^*)p(-\lambda_{n+1})}{\lambda_i^* + \lambda_{n+1}}\right]_{i=1,\ldots,n} = X_\Lambda^* \Sigma^{-1} x^{\text{bal}}(\lambda_{n+1}).$$

Equation (25.13) yields

$$\begin{bmatrix} A^{\text{bal}} = X_\Lambda \Lambda X_\Lambda^{-1} & B^{\text{bal}} = \dfrac{(\lambda_{n+1} I - A^{\text{bal}})x^{\text{bal}}(\lambda_{n+1})}{p(\lambda_{n+1})} \\ C^{\text{bal}} = q_\Lambda X_\Lambda^{-1} & D^{\text{bal}} = \dfrac{p_n}{q_n} \end{bmatrix}$$

with p_n and q_n the coefficients of ξ^n of p and q.

The balancing-reduced model is usually obtained by simply truncating the matrices of the balanced model. That is in fact what we also did in our discussion of the reduced model. However, in our algorithm, the system matrices of the balanced model are obtained by solving linear equations in $\mathbb{R}_n[\xi]$. This suggests other possibilities for obtaining the reduced system matrices. For example, rather that solving equations (25.12) modulo $(x_{n_{\text{red}}+1}^{\text{bal}}, x_{n\text{red}+2}^{\text{bal}}, \ldots, x_n^{\text{bal}})$, one could obtain the best least squares solution of these equations, perhaps subject to constraints, etc.

Further, by combining these least squares ideas with those of section 25.5, it may be possible to obtain balanced reductions that pay special attention to the fit of the reduced order transfer function with the original transfer function at certain privileged frequencies or selected points of the complex plane.

The algorithms discussed have obvious counterparts for discrete-time systems. It is interesting to compare our algorithm for obtaining a balanced state representation with the classical SVD-based algorithm of Kung [4]. Kung's algorithm starts from the Hankel matrix formed by the impulse response and requires the computation of the SVD of an *infinite* matrix. In contrast, our algorithm requires first finding (a least squares approximation of) the governing difference equation, followed by finite polynomial algebra.

25.6 Appendix

Proof of theorem 25.1:
Define K by (25.6). Note that K it is symmetric ($K = K^*$), and that the highest degree in ζ or η is $\mathtt{n}-1$. For every $w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$, there holds

$$\frac{d}{dt}Q_K(w) = |p(\frac{d}{dt})w|^2 - |p(-\frac{d}{dt})w|^2. \tag{25.14}$$

Let $\ell \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ be given. We first prove that

$$\text{minimum} \int_{-\infty}^{0} |p(\frac{d}{dt})\ell'|^2 \, dt = Q_K(\ell)(0),$$

where the minimum is taken over all $\ell' \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ such that

$$\ell'(t), \frac{d}{dt}\ell'(t), \ldots, \frac{d^{\mathtt{n}-1}}{dt^{\mathtt{n}-1}}\ell'(t) \to 0 \text{ as } t \to -\infty \tag{25.15}$$

and

$$\ell'(0) = \ell(0), \frac{d}{dt}\ell'(0) = \frac{d}{dt}\ell(0), \ldots, \frac{d^{\mathtt{n}-1}}{dt^{\mathtt{n}-1}}\ell'(0) = \frac{d^{\mathtt{n}-1}}{dt^{\mathtt{n}-1}}\ell(0). \tag{25.16}$$

Integrating (25.14), yields

$$\int_{-\infty}^{0} |p(\frac{d}{dt})\ell'|^2 \, dt = Q_K(\ell)(0) + \int_{-\infty}^{0} |p(-\frac{d}{dt})\ell'|^2 \, dt.$$

Therefore, the minimum is obtained by the solution of $p(-\frac{d}{dt})\ell' = 0$ that satisfies the initial conditions (25.16). Note that it follows that $Q_K(\ell)(0) \geq 0$. Moreover, $p(-\frac{d}{dt})\ell'(t) = 0, p(\frac{d}{dt})\ell'(t) = 0$ for $t < 0$ implies $\ell'(t) = 0$ for $t < 0$, since $p(\xi)$ and $p(-\xi)$ are co-prime. Therefore the rank of Q_K is $\mathtt{n}$.

Now use a smoothness argument to show that

$$\text{infimum} \int_{-\infty}^{0} |p(\frac{d}{dt})\ell'|^2 \, dt = Q_K(\ell)(0),$$

where the infimum is taken over all $\ell' \in \mathfrak{C}^+(\mathbb{R},\mathbb{R})$ (instead of just having the limit conditions (25.15)) such that $\ell'(t) = \ell(t)$ for $t \geq 0$ (hence ℓ and ℓ' must be glued at $t = 0$ in a $\mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ way, instead of just by the initial conditions (25.16)). This implies that Q_K is indeed the controllability gramian.

Next, consider W, defined by (25.7, 25.8). For every $w \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ here holds

$$\frac{d}{dt}Q_W(w) = -|q(\frac{d}{dt})w|^2 + 2p(\frac{d}{dt})w \quad f(\frac{d}{dt})w. \tag{25.17}$$

Note further that W is symmetric ($W = W^*$), and that the highest degree in ζ or η is $\mathtt{n} - 1$. Therefore, if $\ell' \in \mathfrak{D}(\mathbb{R},\mathbb{R})$ is such that $p(\frac{d}{dt})\ell'(t) = 0$ for $t \geq 0$ there holds, by integrating (25.17) and using the fact that p is Hurwitz,

$$\int_0^\infty |q(\frac{d}{dt})\ell'|^2 \, dt = Q_W(\ell')(0).$$

Let $\ell \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ be given, and assume that

$$\begin{aligned} &\text{(i) } \ell_{(-\infty,0)} = \ell'|_{(-\infty,0)}, \\ &\text{(ii) } (p(\frac{d}{dt})\ell', q(\frac{d}{dt})\ell') \in \mathfrak{B}_{(p,q)}. \end{aligned}$$

Then $\ell'|_{(-\infty,0]}$ is actually a function, with

$$\ell'(0) = \ell(0), \frac{d}{dt}\ell'(0) = \frac{d}{dt}\ell'(0), \ldots, \frac{d^{\mathtt{n}-1}}{dt^{\mathtt{n}-1}}\ell'(0) = \frac{d^{\mathtt{n}-1}}{dt^{\mathtt{n}-1}}\ell'(0).$$

Therefore, if, in addition to (i) and (ii), (iii) holds: $p(\frac{d}{dt})\ell'(t) = 0$ for $t \geq 0$, we obtain

$$\int_0^\infty |q(\frac{d}{dt})\ell'|^2 \, dt = Q_W(\ell)(0).$$

It follows that Q_W defines the observability gramian. That $Q_K \geq 0$ is immediate, and that its rank is $\mathtt{n}$ follows from the fact that p and q are co-prime.

Proof of theorem 25.2:

Using (25.10,25.11), we obtain

$$Q_K(\ell) = \Sigma_{k=1}^{\mathtt{n}} \sigma_{\mathtt{k}}^{-1} |x_{\mathtt{k}}^{\mathrm{bal}}(\frac{d}{dt})\ell|^2,$$

and

$$Q_W(\ell) = \Sigma_{\mathtt{k}=1}^{\mathtt{n}} \sigma_{\mathtt{k}} |x_{\mathtt{k}}^{\mathrm{bal}}(\frac{d}{dt})\ell|^2.$$

Hence, if $\ell_{\mathtt{i}} \in \mathfrak{C}^\infty(\mathbb{R},\mathbb{R})$ is such that $x_{\mathtt{i}}^{\mathrm{bal}}(\frac{d}{dt})\ell_{\mathtt{i}}(0) = \delta_{\mathtt{ij}}$, then

$$Q_K(\ell_{\mathtt{i}})(0) = \sigma_{\mathtt{i}}^{-1}, \text{ and } Q_W(\ell_{\mathtt{i}})(0) = \sigma_{\mathtt{i}}.$$

This shows that the $x_{\mathtt{k}}^{\mathrm{bal}}$'s define a balanced state representation, as claimed. That the $\sigma_{\mathtt{k}}$'s are the Hankel SV's of $\mathfrak{B}_{(p,q)}$ is a standard consequence of the theory of balanced state representations.

Acknowledgment

The program of the SISTA research group of the University of Leuven is supported by grants from several funding agencies and sources: Research Council KUL: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering); Flemish Government: Fund for Scientific Research Flanders, project G.0256.97; Research communities ICCoS, ANMMM, IWT (Soft4s, softsensors), Eureka-Impact (control), Eureka-FLiTE (modeling); Belgian Federal Government: DWTC IUAP V-22 (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling).

25.7 References

[1] P.A. Fuhrmann, *A Polynomial Approach to Linear Algebra,* Springer Verlag, 1996.

[2] F.R. Gantmacher, *The Theory of Matrices, Volume 1,* Chelsea Publishing Co., 1959.

[3] K. Glover, All optimal Hankel-norm approximations of linear multivariable systems: Relations to approximation, *International Journal of Control,* volume 43, pages 1115-1193, 1984.

[4] S.Y. Kung, A new identification method and model reduction algorithm via singular value decomposition, *12-th Asilomar Conference on Circuits, Systems and Computation,* pages 705-714, 1978.

[5] J.W. Polderman and J.C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach,* Springer Verlag, 1998.

[6] P. Rapisarda and J.C. Willems, State maps for linear systems, *SIAM Journal on Control and Optimization,* volume 35, pages 1053-1091, 1997.

[7] J. C. Willems and H. L. Trentelman, On quadratic differential forms, *SIAM Journal of Control and Optimization*, volume 36, pages 1703-1749, 1998.

26

Nonconvex Global Optimization Problems: Constrained Infinite-Horizon Linear-Quadratic Control Problems for Discrete Systems

V.A. Yakubovich

Abstract

In recent works a method of solving some global optimization problems was proposed by the author. These problems may generally be nonconvex. In this paper we describe this method and apply it to solving linear-quadratic deterministic and stochastic infinity-horison optimization problems with integral quadratic constraints.

A. Rantzer, C.I. Byrnes (Eds.): Directions in Mathematical Systems Theory and Optimization, LNCIS 286, pp. 359-382, 2003.
© Springer-Verlag Berlin Heidelberg 2003

26.1 Introduction

A popular design method for linear discrete-time systems is linear-quadratic (LQ) optimal control, whose foundations were laid in the studies by R. Kalman [1], N.N. Krasovsky [2], A.M. Letov [3], A.I. Lur'e [4], (deterministic systems) and A.N. Kolmogorov [5], N. Wiener [6], R.S. Bucy [7] (stochastic systems). This paper is devoted to the similar problems obtained from infinite-horizon LQ-optimal control by inserting quadratic constraints (generally nonconvex) into their statements. The consideration of these constrained LQ-problems is very natural because many requirements, often faced by a controller designer can be expressed in the form of such quadratic constraints. At the same time the constraint LQ-problems may be nonconvex. It is well known that the nonconvexity is a reason of many difficulties for their solution. The Lagrange multiplier rule gives only necessary conditions of optimality and may produce a lot of "superfluous solutions". We use here the method proposed and justified by the author in the earlier papers [8]–[11] for the solution of a special global optimization problems. This class includes many constrained (generally, nonconvex) LQ-problems and, in particular, the problems under consideration. This paper may be considered therefore as an illustration of this method. It is worth noting, that the applicability of this method was extended by A. S. Matveev [12]–[15]. The Appendix gives the statement of a well-known algebraic lemma, essential for the proofs. This formulation taken from [16] is by author's knowlige somewhat more precise as compared with other works, for example [17].

The main part of this paper was written during the author's visit to the Royal Institute of Technology, Department of Mathematics, Optimization and Systems Theory, June 1997. It is supposed to connect the subject of this paper with our joint results with Anders Lindquist [18]–[21] concerning universal regulators design for the solving some LQ-optimization problems in the circumstances of uncertainties. With great pleasure I remember warmly the creative, open and friendly atmosphere of Anders Lindquist laboratory.

26.2 A Method for Solving Constrained Linear-Quadratic Problems (Abstract Theory)

Let $\mathbb{Z} = \{z\}$ be a real Hilbert space with the scalar product $(\cdot, \cdot)$ and

$$\Phi_j(z) = \Phi_j^0(z) + 2(g_j, z) + \gamma_j \qquad (j = 0, \ldots, \nu) \tag{26.1}$$

be the qiven continuous quadratic functionals. Here $g_j \in \mathbb{Z}$, $\Phi_j^0(z) = (G_j z, z)$, $G_j = G_j^*$ are bounded self-ajoint operators. Let $\mathcal{M}$ be subspace in $\mathbb{Z}$ and $\mathcal{L} = \mathcal{M} + z_*$, $z_* \in \mathbb{Z}$ be a fixed plane (affine subspace) in $\mathbb{Z}$. Let $\mathfrak{N}$ be defined by

$$\mathfrak{N} = \{z \in \mathbb{Z} : \quad \Phi_1(z) \geq 0, \ldots, \Phi_\nu(z) \geq 0\}. \tag{26.2}$$

We suppose that the constraints $\Phi_j(z) \geq 0$ are regular: there exists $z_* \in \mathbb{Z}$ such that

$$\Phi_1(z_*) > 0, \ldots, \Phi_\nu(z_*) > 0. \tag{26.3}$$

Consider the problem:

$$\text{Minimize} \quad \Phi_0(z) \quad \text{subject to} \quad z \in \mathcal{L} \cap \mathfrak{N}. \tag{26.4}$$

Note that the functional $\Phi_0(z)$ and the set $\mathfrak{N}$ may be nonconvex, so we consider in general case the problem of global optimization of nonconvex functional on nonconvex set. (Certainly the case of convex Φ_0 and $\mathfrak{N}$ is allowed also.)

Let us form the Lagrangian function for our problem:

$$S(\tau, z) = \Phi_0(z) - \tau_1\Phi_1(z) - \ldots - \tau_\nu\Phi_\nu(z). \tag{26.5}$$

Here $\tau_j \geq 0$. We will write these inequalities as $\tau \geq 0$.

To solve the constrained problem (26.4) using the method [8]–[10] we must take the following actions:

(I) Solve the problem:

$$\text{Minimize} \quad S(\tau, z) \quad \text{subject to} \quad z \in \mathcal{L}. \tag{26.6}$$

(This is the problem without "nonlinear" constraint $z \in \mathfrak{N}$.) More precisely at this step we must only find the value

$$S^0(\tau) = \inf_{z \in \mathcal{L}} S(\tau, z). \tag{26.7}$$

(II) Find any solution $\tau^0 = (\tau_1^0, \ldots, \tau_\nu^0)$ of "dual" problem

$$\text{Maximize} \quad S^0(\tau) \quad \text{subject to} \quad \tau_1 \geq 0, \ldots, \tau_\nu \geq 0. \tag{26.8}$$

(III) Find all the solutions $z(\tau^0)$ of the problem (26.6) for $\tau = \tau^0$.

(IV) Among all the solutions $z(\tau^0)$ find those values $z^0 = z(\tau^0)$ which satisfy the conditions

$$\Phi_j(z^0) \geq 0, \quad \tau_j^0\Phi_j(z^0) = 0, \quad j = 1, \ldots, \nu. \tag{26.9}$$

Let $\{z^0\}$ be the resulting set. (This set may be empty, particularly if τ^0 or $z(\tau^0)$ do not exist.)

We will say that *the rule* (I)–(IV) *is applicable to the considered constraint problem* (26.4) *if the resulting set* $\{z^0\}$ *coincides exactly with the set of all solutions of the problem* (26.4).

Before formulation of the applicability conditions of this rule describe why it is "good". The problem (26.6) is a "linear quadratic" problem: it contains only linear constraints. The theory of many such problems is worked out in LQ-control theory. So it is simple to solve. But here it is necessary to make a caveat: we need a complete solution of this problem, we need to know in which cases $S^0(\tau)$ is finite, in which cases infimum in (26.7) is achieved, and that is for the general case of functional $S(\tau, z)$. The corresponding quadratic form may be indefinite. But if we were to forget these difficulties and consider only the standart LQ-problems, the action (I) usually is reduced to the well studied problem.

The problem (26.8) is a finite dimensional *convex* problem. (The function $S^0(\tau)$ is concave as an infimum of linear functions.) The theory of such problems is also

well studied and the theory of convex programming has many efficient methods to solve them.

So the method considered here reduces difficult problem of global (noncovex in general case) optimization to two simpler problems. Now let us describe in which cases this rule is applicable.

We will assume that the constraints $\Phi_j(z) \geq 0$ are regular (see (26.3)).

THEOREM 26.1
For the case of one constraint ($\nu = 1$) the rule (I)–(IV) is applicable. □

This theorem follows simply from [8], although it is not formulated there. (It is proved also in [10].)

THEOREM 26.2—[9]
Let $\nu > 1$ and let the forms $\Phi_0^0(z)$, $[-\Phi_1^0(z)], \ldots, [-\Phi_\nu^0(z)]$ be convex for $z \in \mathcal{M}$. Then the rule (I)–(IV) is applicable. □

(This theorem is related to the convex programming and may considered as a well known.)

The following theorem is the most important for applications.

THEOREM 26.3—[9]–[11]
Assume that there exist linear bounded operators $T_k : \mathbb{Z} \to \mathbb{Z}$, $k = 1, 2, \ldots$, such that
(i) $(T_k z_1, z_2) \to 0$ as $k \to +\infty$ (for any $z_1 \in \mathbb{Z}$, $z_2 \in \mathbb{Z}$),
(ii) $T_k \mathcal{M} \subset \mathcal{M}$,
(iii) $\Phi_j^0(T_k z) \to \Phi_j^0(z)$ as $k \to +\infty$ (for any $z \in \mathcal{M}$, $j = 1, \ldots, \nu$).

Then the rule (I)–(IV) is applicable. □

If the assumptions of any of the theorem 26.1–26.3 hold then the following duality relation is true:

$$\inf_{z \in \mathcal{L} \cap \mathfrak{N}} \Phi_0(z) = \max_{\tau \geq 0} \inf_{z \in \mathcal{L}} S(\tau, z). \tag{26.10}$$

(If $\inf_{z \in \mathcal{L}} S(\tau, z) = -\infty$ for all τ then we put $\max(-\infty) = -\infty$.) If τ^0 exists then using the formula (26.7) we can rewrite (26.10) as

$$\inf_{z \in \mathcal{L} \cap \mathfrak{N}} \Phi_0(z) = S^0(\tau^0). \tag{26.11}$$

Let us now prove the following two simple prepositions which can sometimes be useful.

THEOREM 26.4
Suppose that there exists a solution $\tau^0 = (\tau_1^0, \ldots, \tau_\nu^0) \geq 0$ of the dual problem (26.8)[1]. Suppose also that for all $\tau \geq 0$ close to τ^0 the infimum

$$\inf_{z \in \mathcal{L}} S(\tau, z) = S(\tau, z(\tau)) \tag{26.12}$$

[1]As recently proved by A.S. Matveev τ^0 exists if the regularity condition (26.3) holds. So this assumption is met automatically and it may be dropped. We do not suppose in Theorem 26.4 and Theorem 26.5 that the regularity condition holds.

is achived in some point $z(\tau) \in \mathcal{L}$ and there exist the derivatives

$$\left(\frac{dz(\tau)}{d\tau_j}\right)_{\tau^0} \quad (j = 1, \ldots, \nu) \tag{26.13}$$

(one-sided if $\tau_j^0 = 0$). Then the rule (I)–(IV) is applicable and $z^0 = z^0(\tau^0)$, that is the conditions (26.9) are met. In addition, the duality relation (26.10) holds. □

THEOREM 26.5
Let

$$G_0 - \sum_{j=1}^{\nu} \tau_j^0 G_j =: G(\tau^0) \tag{26.14}$$

be the operator of the quadratic form part of $S(\tau^0, z)$. Suppose that for some $\delta > 0$

$$(G(\tau^0)z, z) \geq \delta \|z\|^2 \quad \text{for all} \quad z \in \mathcal{M}. \tag{26.15}$$

Then the rule (I)–(IV) is applicable and the solution of primary problem $z^0 = z(\tau^0)$ is uniquely defined. (So we do not need to verify the conditions (26.9).) In addition, the duality relation (26.10) holds. □

Proof of Theorem 26.4. If follows from (26.12) and (26.7)

$$S^0(\tau) = S[\tau, z(\tau)] \leq S(\tau, z) \qquad (\forall\, z \in \mathcal{L}) \tag{26.16}$$

for all $\tau \geq 0$ close to τ^0. Denote $z^0 = z(\tau^0)$. Any $z \in \mathcal{L}$ has the form $z = z^0 + y$, $y \in \mathcal{M}$. Therefore

$$S^0(\tau^0) = S(\tau^0, z^0) \leq S(\tau^0, z^0 + y) \qquad (\forall\, y \in \mathcal{M}). \tag{26.17}$$

The functional $S(\tau^0, z^0 + y)$ is quadratic in $y \in \mathcal{M}$, so the Frechet derivative $\partial S(\tau^0, z^0 + y)/\partial y$ exists and (26.17) implies

$$\frac{\partial S(\tau^0, z^0 + y)}{\partial y}\Big|_{y=0} = 0. \tag{26.18}$$

Since $z(\tau) \in \mathcal{L}$, $S(\tau, z)$ is linear in τ and quadratic in z and the derivative $\left(\frac{\partial z(\tau)}{\partial \tau_j}\right)_{\tau^0}$ exists then the derivative $\left(\frac{dS^0(\tau)}{d\tau_j}\right)_{\tau^0} = \left[\frac{dS(\tau, z(\tau))}{d\tau_j}\right]_{\tau^0}$ exists and

$$\kappa_j = \left(\frac{dS^0(\tau)}{d\tau_j}\right)_{\tau^0} = \left[\frac{dS(\tau, z(\tau))}{d\tau_j}\right]_{\tau^0} =$$

$$\left[\frac{\partial S(\tau, z(\tau))}{\partial \tau_j}\right]_{\tau^0} + \left[\frac{\partial S(\tau^0, z^0 + y)}{\partial y}\right]_{y=0} \cdot \left(\frac{dz(\tau)}{d\tau_j}\right)_{\tau^0}. \tag{26.19}$$

Since $S^0(\tau^0) \geq S^0(\tau)$ for $\tau \geq 0$, we obtain

$$\kappa_j = 0 \quad \text{if} \quad \tau_j^0 > 0, \qquad \kappa_j \leq 0 \quad \text{if} \quad \tau_j^0 = 0. \tag{26.20}$$

On the other hand, it follows from (26.18), (26.19) and (26.5) that

$$\kappa_j = -\Phi_j(z^0), \quad j = 1, \dots, \nu. \tag{26.21}$$

Therefore the conditions (26.9) hold and $z^0 \in \mathcal{L} \cap \mathfrak{N}$.

Let us show that z^0 is a solution of the problem (26.4). Let $z \in \mathcal{L} \cap \mathfrak{N}$ and $\tau \geq 0$. Then (26.5) implies $S(\tau, z) \leq \Phi_0(z)$ and

$$\inf_{z \in \mathcal{L}} S(\tau, z) \leq \inf_{z \in \mathcal{L}} \Phi_0(z) \leq \inf_{z \in \mathcal{L} \cap \mathfrak{N}} \Phi_0(z).$$

Therefore $S^0(\tau^0) = \max\limits_{\tau \geq 0} \inf_{z \in \mathcal{L}} S(\tau, z) \leq \inf_{\mathcal{L} \cap \mathfrak{N}} \Phi_0(z)$. But (26.20), (26.21) imply $\tau_j^0 \Phi_j(z^0) = 0$, $S(\tau^0) = \Phi_0(z^0)$. Since $z^0 \in \mathcal{L} \cap \mathfrak{N}$, we obtain the duality relation (26.10) and the needed relation

$$\Phi_0(z^0) = \inf_{z \in \mathcal{L} \cap \mathfrak{N}} \Phi_0(z).$$

Consider now an arbitrary solution $z^0 \in \mathcal{L} \cap \mathfrak{N}$ of the problem (26.8) and establish that it is obtained by our rule. It means that z^0 is a solution of the problem (26.6) for $\tau = \tau^0$. We have $\Phi_0(z^0) = \inf_{\mathcal{L} \cap \mathfrak{N}} \Phi(z)$. The duality relation (26.10) implies

$$\Phi_0(z^0) = \max_{\tau \geq 0} \inf_{z \in \mathcal{L}} S(\tau, z) = \inf_{z \in \mathcal{L}} S(\tau^0, z). \tag{26.22}$$

This shows that z^0 is a solution of the problem (26.6) for $\tau = \tau^0$, as required.

Moreover, the relations (26.9) hold. In fact, (26.22) implies

$$\Phi_0(z^0) \leq S(\tau^0, z^0) = \Phi_0(z^0) - \sum_{j=1}^{\nu} \tau_j^0 \Phi_j(z^0),$$

$$\sum_{j=1}^{\nu} \tau_j^0 \Phi_j(z^0) \leq 0.$$

But $\tau_j^0 \geq 0$, $\Phi_j(z^0) \geq 0$. Therefore $\tau_j^0 \Phi_j(z^0) = 0$.

Proof of Theorem 26.5. From (26.1) and (26.5)

$$S(\tau, z) = (G(\tau)z, z) + 2(g(\tau), z) + \gamma(\tau), \tag{26.23}$$

where $G(\tau) = G_0 - \sum_1^{\nu} \tau_j G_j$, $g(\tau) = g_0 - \sum_1^{\nu} \tau_j g_j$, $\gamma(\tau) = \gamma_0 - \sum_0^{\nu} \tau_j \gamma_j$. Let $z_* \in \mathcal{L}$, then $z \in \mathcal{L}$ iff $z = z_* + y$, $y \in \mathcal{M}$. For $z \in \mathcal{L}$ we obtain

$$S(\tau, z) = (Q(\tau)y, y) + 2(q(\tau), y) + \kappa(\tau), \tag{26.24}$$

where $Q(\tau) = PG(\tau)P$, $P = P^*$ is the orthoprojector on the subspace $\mathcal{M}$ and $Q(\tau), q(\tau) \in \mathcal{M}$ are linear functions of τ. By hypothesis of the Theorem 26.5

$$(G(\tau)y, y) = (Q(\tau)y, y) \geq \delta \|y\|^2 \quad \text{for all} \quad y \in \mathcal{M} \tag{26.25}$$

and for $\tau = \tau^0$. Since $Q(\tau)$ is continuous in τ then $Q(\tau)$ remains strictly positive and for all τ closed to τ^0. Therefore $Q(\tau)^{-1}$ exists. We have for $z \in \mathcal{L}$ and τ closed to τ^0

$$S(\tau, z) = (Q(y + Q^{-1}q), (y + Q^{-1}q)) + \kappa(\tau) - (Q^{-1}q, q), \tag{26.26}$$

where $Q = Q(\tau) > 0$, $q = q(\tau)$. The infimum $\inf_{z \in \mathcal{L}} S(\tau, z) = \inf_{y \in \mathcal{M}} S(\tau, z_* + y)$ is achived in the point $z = z(\tau) = z_* + y(\tau)$, $y(\tau) = -Q(\tau)^{-1}q(\tau)$, and this point is uniquely defined.

Since $Q(\tau), q(\tau)$ depend linearly in τ, the derivative $dz^0(\tau)/d\tau_j$ exists. So the assumptions of Theorem 26.4 are satisfied. That completes the proof of the Theorem 26.5.

Let us apply the general method described above to the solution of an LQ-optimization problem which differs from well known ones because of the presence of quadratic constraints.

26.3 Linear-Quadratic Deterministic Infinite-Horizon Constrained Optimization Problem

Problem statement. Let us consider the plant which described by usual equations

$$x_{t+1} = Ax_t + Bu_t, \quad x_0 = a \quad (t = 0, 1, 2, \ldots), \tag{26.27}$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$, A, B are the real constant matrices, the pair (A, B) is controllable. We consider all the controllers which ensure the "stability conditions"

$$\|x_t\|^2 = \sum_{t=0}^{\infty} |x_t|^2 < \infty, \quad \|u_t\|^2 = \sum_{t=0}^{\infty} |u_t|^2 < \infty \tag{26.28}$$

and the constraints

$$\Phi_j = \gamma_j - \sum_{t=0}^{\infty} \mathcal{G}_j(x_t, u_t) \geq 0, \quad j = 1, 2, \ldots, \nu. \tag{26.29}$$

Among these controllers, or (it is the same) among all processes $\left\{ \begin{bmatrix} x_t \\ u_t \end{bmatrix} \right\}_{t=0}^{\infty}$ satisfying (26.27)–(26.29) we wish to find the controller (the process) which minimizes the functional

$$\Phi_0 = \sum_{t=0}^{\infty} \mathcal{G}_0(x_t, u_t). \tag{26.30}$$

We will also consider the question if this optimal process may be yielded a realizable linear stabilizing regulator.

In (26.29), (26.30) $\mathcal{G}_j(x, u)$ are given real quadratic forms in (x, u). The vector a and the numbers γ_j are given. The practical interpretation of this problem is evident and we omit it.

This problem differs from the one well known because of the presence the constraints (26.29).

Let us emphasize that the considered problem is a problem of global nonconvex minimization because the functionals Φ_j can be nonconvex.

Suppose first for simplicity that the constraints (26.29) are regular: the admissible process $\left\{\begin{bmatrix} x_t \\ u_t \end{bmatrix}\right\}_{t=0}^{\infty}$ exists such that $\Phi_j > 0$, $j = 1, \dots, \nu$.

Reduction to the abstract scheme. Let $z = \left\{\begin{bmatrix} x_t \\ u_t \end{bmatrix}\right\}_{t=0}^{\infty}$ be a process with only one property: $\|x_t\| < \infty, \|u_t\| < \infty$ and $\mathbb{Z} = \{z\}$ be a Hilbert space of all such processes with the usual scalar product:

$$(z', z'') = \sum_{t=0}^{\infty} [(x_t', x_t'') + (u_t', u_t'')].$$

Let $\mathcal{L}$ be the affine space of all $z \in \mathbb{Z}$ which satisfy the equations of the object dynamics: $x_{t+1} = Ax_t + Bu_t$, $x_0 = a$. Obviously the corresponding subspace $\mathcal{M} \subset \mathbb{Z}$ is defined by the same equations with $a = 0$:

$$x_{t+1} = Ax_t + Bu_t, \qquad x_0 = 0. \tag{26.31}$$

Let $\mathfrak{N}$ be the set of all processes $z \in \mathbb{Z}$ satisfying our constraints:

$$\mathfrak{N} = \{z : \quad \Phi_1(z) \geq 0, \dots, \Phi_\nu(z) \geq 0\}. \tag{26.32}$$

We obtain a special case of the problem considered in [9, 10]: To minimize $\mathcal{G}_0(z)$ subject to $z \in \mathcal{L} \cap \mathfrak{N}$.

Let us show that the conditions (i)–(iii) of Theorem 26.3 hold. This in turn implies that the procedure (I)–(IV) is applicable for solving our problem. Let us take T_k to be shift operators: if $z = \{z_t\}_0^\infty$, $z_t = \begin{bmatrix} x_t \\ u_t \end{bmatrix}$, then $z' = T_k z$ $(= \{z_t'\}_0^\infty)$ means $z_t' = 0$ for $t = 0, \dots, k-1$ and $z_t' = z_{t-k}$ for $t = k, k+1, \dots$ Obviously $T_k : \mathbb{Z} \to \mathbb{Z}$ and

(i) $(T_k z', z'') \to 0$ as $k \to \infty$ for any $z', z'' \in \mathbb{Z}$;

(ii) $T_k \mathcal{M} \subset \mathcal{M}$;

(iii) $\Phi_j^0(T_k z) = \Phi_j^0(z)$, where $\Phi_0^0 = \sum_{t=0}^{\infty} \mathcal{G}_0(x_t, u_t)$, $\Phi_j^0 = -\sum_{t=0}^{\infty} \mathcal{G}_j(x_t, u_t)$, $j = 1, \dots, \nu$.

(In fact $|(T_k z', z'')| = \left| \sum_{t=k}^{\infty} (z_{t-k}', z_t'') \right| \leq \sum_{t=k}^{\infty} |z_{t-k}'| \cdot |z_t''| \leq \left(\sum_{t=k}^{\infty} |z_{t-k}'|^2 \cdot \sum_{t=k}^{\infty} |z_t''|^2 \right) \to 0$ as $k \to \infty$. The properties (ii),(iii) are clearly met.)

Application the procedure (I)–(IV). By our rule we have to solve first the problem (26.6) without constraints. The Lagrangian is:

$$S(\tau, z) = \Phi_0(z) - \sum_{j=1}^{\nu} \tau_j \Phi_j(z) = \Phi(\tau, z) - \sum_{j=1}^{\nu} \tau_j \gamma_j, \tag{26.33}$$

where $\tau = \|\tau_j\|_{j=1}^{\nu}$, $\tau_j \geq 0$ are parameters to be find later and

$$\Phi(\tau, z) = \sum_{t=0}^{\infty} \mathcal{G}(\tau, x_t, u_t), \quad \mathcal{G}(\tau, x, u) = \mathcal{G}_0(x, u) + \sum_{j=1}^{\nu} \tau_j \mathcal{G}_j(x, u). \tag{26.34}$$

The step (I) of our procedure requires us to solve the following auxiliary problem: Find

$$S^0(\tau) = \inf_{z \in \mathcal{L}} S(\tau, z). \tag{26.35}$$

This is the usual LQ-problem without quadratic constraints: because of (26.33) we have to minimize the quadratic functional $\Phi(\tau, z)$ under constraints (26.27), (26.28). There are many results concerning various aspects of this problem. Contrary to the majority of these results, we need a complete solution: We have to know in which cases $\inf_{z \in \mathcal{L}} S(\tau, z)$ is attained and we have to know the value of $S_0(\tau)$ for all $\tau_j \geq 0$ and others. The complete solution of this problem (without quadratic constraints) follows simply from the results presented in Appendix. The results of this Appendix taken from [22] are now "almost known" and may be known. In each case they are closely related to the known ones and they are well known in their essence. However the author can not give the appropriate references in which this result and the complete solution are represented.

The step (II) of our procedure requires us to find

$$\tau^0 = \arg \max_{\tau \geq 0} S^0(\tau).$$

The step (III) requires us to find the solution $z = z(\tau^0)$ of the problem (26.35) for $\tau = \tau^0$. If the assumptions of Theorem 26.4 hold then $z(\tau^0)$ is a solution of our primary constraint problem. (In this case we do not need the regularity assumption.) Otherwise, according to the step (IV) we have to choose from all the solutions $z(\tau^0)$ those z^0 which satisfy the conditions $\Phi_j(z^0) \geq 0$, $\tau_j^0 \Phi_j(z^0) = 0$, $j = 1, \ldots, \nu$. By Theorem 26.3 these z^0 and only they are the solutions of our primary constraint problem. If such z^0 does not exist then infimum is not attained in our primary constraint problem. In this case we shall construct the optimizing sequences of the regulators. Let us proceed as described.

Solution of the auxiliary problem without quadratic constraints.

Recall that this is the problem of minimizing the functional $\Phi(\tau, z)$ in (26.34) (or equivalently the functional $S(\tau, z)$) subject to constraints (26.27), (26.28).

Let us introduce the notation:

$$A_\lambda = \lambda I_n - A, \quad \delta(\lambda) = \det A_\lambda, \quad Q_\lambda = \delta(\lambda) A_\lambda^{-1}. \tag{26.36}$$

Following Appendix consider the identity

$$\begin{aligned} &(Ax + Bu)^* P(\tau)(Ax + bu) - x^* P(\tau) x + \mathcal{G}(\tau, x, u) = \\ &|\kappa(\tau)(u - K(\tau)x)|^2 \qquad (\forall\, x \in \mathbb{R}^n, \quad \forall\, u \in \mathbb{R}^m). \end{aligned} \tag{26.37}$$

Here $P(\tau) = P(\tau)^*$, $\kappa(\tau) = \kappa(\tau)^*$, $K(\tau)$ are the real $n \times n$, $m \times m$ and $m \times n$ matrices to be find. The identity (26.37) may be transformed into the well-known

Lur'e-Riccati equation for the matrix $P(\tau)$. Let us define "the frequency matrix" $\Pi(\tau,\lambda)=\Pi(\tau,\lambda)^*$ (for $\lambda\in\mathbb{C}$, $|\lambda|=1$) by the identity

$$\mathcal{G}(\tau,\widetilde{x},\widetilde{u})=\widetilde{u}^*\Pi(\tau,\lambda)\widetilde{u}\qquad(\widetilde{x}=A_\lambda^{-1}B\widetilde{u},\ \det A_\lambda\neq 0,\ \forall\,\widetilde{u}\in\mathbb{C}^m). \tag{26.38}$$

We can use Theorem 26.14 from Appendix. Consider two domains in τ-space corresponding to WFC and SFC (see the terminology in Appendix):

$$D=\{\tau:\ \tau_j\geq 0,\ \Pi(\tau,\lambda)\geq 0,\ \text{for any }\lambda,\ |\lambda|=1,\ \delta(\lambda)\neq 0\} \tag{26.39}$$

$$\begin{aligned}D^0=\{\tau:\ \tau_j\geq 0,\ \Pi(\tau,\lambda)>0,\ \forall\,\lambda,\ |\lambda|=1,\ \delta(\lambda)\neq 0\quad\text{and}\\ \lim|\delta(\lambda)|^2\Pi(\tau,\lambda)>0\ \text{if}\ \lambda\to\lambda_j,\ \forall\,\lambda_j,\ |\lambda_j|=1,\ \delta(\lambda_j)=0\}.\end{aligned} \tag{26.40}$$

The last condition in (26.40) means that for any root λ_j of the polynomial $\delta(\lambda)=\det(\lambda I_n-A)$ such that $|\lambda_j|=1$ the inequality $\lim\limits_{\lambda\to\lambda_j}|\delta(\lambda)|\Pi(\tau,\lambda)>0$ holds. This condition is absent if $\delta(\lambda)$ has no roots λ_j with $|\lambda_j|=1$. Note that if $\delta(\lambda)$ has a root λ_j, $|\lambda_j|=1$, then λ_j is the singular point of $\Pi(\lambda)$. It is easy to show that the set D^0 can also be defined by the following conditions:

$$\begin{aligned}D^0=\{\tau:\ \tau_j\geq 0,\ \exists\,\varepsilon=\varepsilon(\tau)>0:\ \mathcal{G}(\tau,\widetilde{x},\widetilde{u})\geq\varepsilon(|\widetilde{x}|^2+|\widetilde{u}|^2),\\ \forall\,\widetilde{x}\in\mathbb{C}^n,\ \widetilde{u}\in\mathbb{C}^m,\ \lambda\in\mathbb{C},\ |\lambda|=1\ \text{such that}\ \lambda\widetilde{x}=A\widetilde{x}+B\widetilde{u}\}\end{aligned} \tag{26.41}$$

Obviously $D^0\subset D$. Using the terminology of Appendix we can say that D and D^0 are sets, where WFC and SFC hold correspondingly.

LEMMA 26.1
If $\tau\notin D$ then $S^0(\tau)=\inf_{z\in\mathcal{L}}S(\tau,z)=-\infty$. □

Proof. For $\tau\notin D$ the WFC fails, $z_0\in\mathcal{M}$ exists with $\Phi(\tau,z_0)<0$, $\inf_{z\in\mathcal{L}}\Phi(\tau,z)=-\infty$. It follows from (26.33) that $S^0(\tau)=\inf_{z\in\mathcal{L}}S(\tau,z)=-\infty$. The Lemma is proved.

Note that real solution $P(\tau)=P(\tau)^*$, $K(\tau)$, $\kappa(\tau)=\kappa(\tau)^*$ does not exist if $\tau\notin D$ (see Appendix). Therefore $S^0(\tau)=-\infty$ for all $\tau\geq 0$ if D is empty. By duality relation (26.10) we obtain for our primary constraint problem:

$$\inf_{\mathfrak{N}}\Phi^0(z)=-\infty\quad\text{if}\quad D=\varnothing. \tag{26.42}$$

Let $\tau\in D$, i.e. WFC holds. By discrete KYP-Lemma (see Appendix) there exist real matrices $P(\tau)=P(\tau)^*$, $K(\tau)$, $\kappa(\tau)=\kappa(\tau)^*>0$, such that the identity (26.37) holds and

$$\det[\lambda I_n-(A+BK)]\neq 0\quad\text{for}\quad|\lambda|>1. \tag{26.43}$$

Suppose that the given quadratic forms in (26.29) are:

$$\mathcal{G}_j(x,u)=\begin{bmatrix}x\\u\end{bmatrix}^*\begin{bmatrix}G_j & g_j\\ g_j^* & \Gamma_j\end{bmatrix}\begin{bmatrix}x\\u\end{bmatrix},\ j=0,\dots,\nu,\ (G_j=G_j^*,\ \Gamma_j=\Gamma_j^*). \tag{26.44}$$

It follows from (26.34), that

$$\mathcal{G}(\tau,x,u)=\begin{bmatrix}x\\u\end{bmatrix}^*\begin{bmatrix}G(\tau) & g(\tau)\\ g(\tau)^* & \Gamma(\tau)\end{bmatrix}\begin{bmatrix}x\\u\end{bmatrix}, \tag{26.45}$$

where

$$G(\tau)=G_0+\sum_{j=1}^{\nu}\tau_j G_j,\quad g(\tau)=g_0+\sum_{j=1}^{\nu}\tau_j g_j,\quad \Gamma(\tau)=\Gamma_0+\sum_{j=1}^{\nu}\tau_j\Gamma_j. \tag{26.46}$$

The identity (26.37) may be rewriten as the following Lur'e-Riccati equations

$$\left.\begin{aligned}&A^*PA-P+G=K^*\kappa^2K\\ &B^*PA+g^*=-\kappa^2K\\ &B^*PB+\Gamma=\kappa^2\end{aligned}\right\} \tag{26.47}$$

Using Theorem 26.14 (Appendix), we obtain $\inf_{z\in\mathcal{L}}\Phi(\tau,z)=a^*P(\tau)a$. Therefore, keeping in mind (26.33) and (26.35),

$$S^0(\tau)=a^*P(\tau)a-\sum_{j=1}^{\nu}\tau_j\gamma_j\quad\text{if}\quad\tau\in D. \tag{26.48}$$

But it is possible that in this case $\inf_{z\in\mathcal{L}}S(\tau,z)=S^0(\tau)$ is not attained (the optimal process does not exist).

Let $\tau\in D^0$, i.e. SFC holds. By Theorem 26.14 (Appendix), there exist real matrices $P(\tau)=P(\tau)^*$, $K(\tau)$, $\kappa(\tau)=\kappa(\tau)^*>0$ (and $P(\tau),K(\tau),\kappa(\tau)$ are uniquely defined), such that the identity (26.37) holds and

$$\det\left[\lambda I_n-(A+BK)\right]\neq 0\quad\text{for }|\lambda|\geq 1. \tag{26.49}$$

(We call these matrices a "strictly stabilizing" solution.) There are various methods for determining the matrices $P(\tau)$, $K(\tau)$, $\kappa(\tau)$; for the case $m=1$ one of them is represented in Appendix (it is convenient if the system contains the unknown parameters to be find), general case is considered in [22]. By Theorem 26.14 (Appendix) in this case the optimal process in the auxiliary problem exists and is defined by the regulator

$$u_t=K(\tau)x_t. \tag{26.50}$$

Consequently in this case the optimal process is defined uniquely. This finishes the action (I) of our rule.

The solution of the primary constraint problem. Continue to apply our rule. Accordingly to the action (II) we must find the value

$$\tau^0=\arg\max_{\tau\in D}S^0(\tau). \tag{26.51}$$

As we notes above the function $S^0(\tau)$ (defined by (26.48)) is a convex function, because we see from (26.35) that $S^0(\tau)$ is the infimum of the linear (in τ) functions.

So the problem to find τ^0 is a well known problem of finite dimension convex programming. There are many methods to solve it (see [22] and others).

Suppose that τ^0 is determined. According to the formulas (26.11) and (26.48) in our primary constraint problem, the infimum of $\Phi_0(z) = \mathcal{G}_0(z)$ is finite and

$$\inf_{\mathcal{L}\cap\mathfrak{N}}\Phi_0 = a^* P(\tau^0)a - \sum_{j=1}^{\nu}\tau_j^0\gamma_j. \tag{26.52}$$

To find the solution of our primary problem according to the step (III), we have to find all the solutions $z = z(\tau^0)$ of the auxiliary problem for $\tau = \tau^0$. Consider first the case $\tau^0 \in D^0$. As shown earlier, the optimal process $z(\tau^0) = \left\{\begin{bmatrix} x_t \\ u_t \end{bmatrix}\right\}_{t=0}^{\infty}$ exists, is defined uniquely and is determined by the regulator $u_t = K(\tau^0)x_t$. Instead of applying the step (IV), let us use the Theorem 26.5. As SFC holds for $\tau^0 \in D^0$ we conclude by Theorem 26.15 (Appendix) that the functional $\Phi(\tau^0, z)$ is strongly positive on the subspace

$$\mathcal{M} = \left\{ z = \left\{\begin{bmatrix} x_t \\ u_t \end{bmatrix}\right\}_{t=0}^{\infty} \in \mathbb{Z} :\ x_{t+1} = Ax_t + Bu_t, \quad x_0 = 0 \right\}.$$

By Theorem 26.5 we do not need in the step (IV) and the process $z(\tau^0)$ is the solution (unique) of our primary constraint problem. Recall that in this case we do not need the assumption that the constraints $\Phi_j \geq 0$ are regular.

Consider now the case of $\tau^0 \in D\backslash D^0$ and assume that the constraints $\Phi_j \geq 0$ are regular. Because WFC holds and SFC does not hold, the matrix $A + BK(\tau^0)$ has all the roots in the disk $|\lambda| \leq 1$ and (necessarily) part of them lie on the circle $|\lambda| = 1$. By Theorem 26.14 (Appendix) the optimal process $z(\tau^0) = \left\{\begin{bmatrix} x_t \\ u_t \end{bmatrix}\right\}_{t=0}^{\infty}$ exists (and is defined by the same regulator $u_t = K(\tau^0)x_t$) if the initial vector $x_0 = a$ belongs to the stable subspace of the matrix $A + BK(\tau^0)$ and does not exist in otherwise. Suppose that $x_0 = a$ belongs to this stable subspace. In the step (IV) we must consider the conditions $\Phi_j \geq 0$, $\tau_j^0\Phi_j = 0$, $j = 1, \ldots, \nu$ for this obtained process. If these conditions hold then this process is optimal for our primary constraint problem. If they do not hold then the infimum is not attained in our primary constraint problem.

If $x_0 = a$ does not belong to the stable subspace of $A + BK(\tau^0)$ then by Theorem 26.14 (Appendix) the auxiliary problem has no solutions (the set $\{z(\tau^0)\}$ is empty) and the infimum is not attained in our primary constraint problem.

So we obtain the solution of our constraint problem for all cases.

Note that if $\tau^0 \in D\backslash D^0$ and $x_0 = a$ belongs to the stable subspace and we get the solution $u = K(\tau^0)x$ then this solution is not satisfactory from practice point of view. In fact firstly the regulator $u = K(\tau^0)x$ is parametrically unstable (the matrix $A + BK(\tau^0)$ becomes unstable after small perturbations of the system coefficients and the stability condition (26.28) fails). Secondly, this regulator does not give the optimal process after suitable arbitrary small perturbations of the initial state $x_0 = a$.

Let us summarize the main results obtained.

THEOREM 26.6
If D is empty then in the primary constraint problem $\inf\Phi_0 = -\infty$. □

THEOREM 26.7
If D is not empty then for $\tau \in D$ there exist real matrices $P(\tau) = P(\tau)^*$, $K(\tau)$, $\kappa(\tau) = \kappa(\tau)^*$ satisfying to the identity (26.37) (or equivalently to the Lur'e-Riccati equations (26.47)), such that $\kappa > 0$, $\det[\lambda I_n - A - BK(\tau)] \neq 0$ as $|\lambda| > 1$ and in the primary constraint problem

$$\inf\Phi_0 = a^* P(\tau^0) a - \sum_{j=1}^{\nu} \tau_j^0 \gamma_j, \tag{26.53}$$

where

$$\tau^0 = \arg\max_{\tau \in D} \left[a^* P(\tau) a - \sum_{j=1}^{\nu} \tau_j \gamma_j \right]. \tag{26.54}$$

□

THEOREM 26.8
If D^0 is not empty and $\tau^0 \in D^0$, where τ^0 is defined by (26.54), then the optimal process $\left\{ \begin{bmatrix} x_t \\ u_t \end{bmatrix} \right\}_{t=0}^{\infty}$ exists for our primary constraint problem. It is defined uniquely and it is given by the stabilizing regulator $u_t = K(\tau^0) x_t$. Here $K(\tau^0)$ together with $P(\tau^0) = P(\tau^0)^*$, $\kappa(\tau^0) = \kappa(\tau^0)^*$ are defined by the identity (26.37) (or equivalently by Lur'e-Riccati equations (26.47)) and by the conditions $\kappa > 0$, $\det[\lambda I_n - A - BK(\tau^0)] \neq 0$ as $|\lambda| \geq 1$. □

Note that in this theorem we do not need the regularity assumption of the constraints $\mathcal{G}_j \leq \gamma_j$, $j = 1, \ldots, \nu$.

THEOREM 26.9
If D is not empty and $\tau^0 \in D \backslash D^0$, then for our primary constraint problem either the optimal process does not exist or it exists and it is given by the parametrical nonstable regulator $u_t = K(\tau^0) x_t$ and it can not exist after some arbitrary small perturbation of the initial state or of the system parameters. □

In the theorem 26.8, 26.9 τ^0 is defined by (26.53) and $K(\tau)$, $P(\tau)$, $\kappa(\tau)$ are matrices defined in Theorem 26.7.

Let us repeat shortly the procedure for solving the primary problem. We have the nonconvex (in general) global problem of minimizing the functional (26.30) subject to constraints (26.27), (26.28), (26.29).

Let $\tau = \|\tau_j\|_{j=1}^{\nu}$, $\tau_j \geq 0$ and

$$\mathcal{G}(\tau, x, u) = \mathcal{G}_0(x, u) + \sum_{j=1}^{\nu} \tau_j \mathcal{G}_j(x, u).$$

To solve this problem we must:
1^0. Determine $\Pi(\tau, \lambda)$ by the formulas (26.38).
2^0. Determine the domains D and D^0 by (26.39), (26.40).
3^0. Determine the "stabilizing" solution $P(\tau) = P(\tau)^*$, $\kappa(\tau) = \kappa(\tau)^*$, $K(\tau)$ of the identity (26.37) (or Lur'e-Riccati equations (26.47)), such that all eigenvalues of $A + BK(\tau)$ lie in $|\lambda| \leq 1$.
4^0. Put $S^0(\tau) = a^* P(\tau) a - \sum_{j=1}^{\nu} \tau_j \gamma_j$ and determine $\tau^0 = \arg\max_{\tau \in D} S^0(\tau)$. (If the constraints are regular τ^0 do exists.)
5^0. If $\tau^0 \in D^0$ then the chosen optimal process is determined by the stabilizing regulator $u_t = K(\tau^0) x_t$, the optimal process is unique and

$$\inf_{\Phi_j \geq 0} \Phi_0 = S^0(\tau^0). \tag{26.55}$$

6^0. If $\tau^0 \in D \backslash D^0$ then (26.55) remains true but either the optimal process does not exists or it exists but it is parametrical unstable in the above mentioned sense.

26.4 Linear-Quadratic Stochastic Infinite-Horizon Optimization Problem with White-Noise Disturbance

Problem statement. Let us consider the plant described by the equation

$$x_{t+1} = Ax_t + Bu_t + Cv_{t+1} \qquad (t = \ldots, -2, -1, 0, 1, 2, \ldots), \tag{26.56}$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$, $v_t \in \mathbb{R}^l$, A, B, C are the real constant matrices, the pair (A, B) is controllable, v_t is a normed white-noise disturbance

$$\mathbf{E}\, v_t = 0, \qquad \mathbf{E}\, v_t v_s^* = I_l \cdot \delta_{ts}. \tag{26.57}$$

Until otherwise specified, we assume that (x_t, u_t) is a stationary process. It must satisfy the quadratic constraints

$$\Phi_j = \gamma_j - \mathbf{E}\, \mathcal{G}_j(x_t, u_t) \geq 0, \quad j = 1, \ldots, \nu. \tag{26.58}$$

We wish to minimize the quadratic functional

$$\Phi_0 = \mathbf{E}\, \mathcal{G}_0(x_t, u_t) \tag{26.59}$$

subject to the constraints (26.56), (26.58). Here $\mathcal{G}_j(x, u)$ are the real quadratic forms

$$\mathcal{G}_j(x, u) = \begin{bmatrix} x \\ u \end{bmatrix}^* \begin{bmatrix} G_j & g_j \\ g_j^* & \Gamma_j \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}, \quad j = 0, 1, \ldots, \nu \tag{26.60}$$

and $G_j = G_j^*$, $\Gamma_j = \Gamma_j^*$.

The minimization must take place over all admissible regulators. As before it is convenient for us to consider instead of the class of admissible regulators

the class $\mathbb{Z}$ of admissible processes, i.e. of the processes determined by admissible regulators. Let $\mathbb{Z}$ be the set of all stationary processes $z = \left\{ \begin{bmatrix} x_t \\ u_t \end{bmatrix} \right\}_{-\infty}^{+\infty}$ which can be represented in the following form:

$$\begin{bmatrix} x_t \\ u_t \end{bmatrix} = \sum_{s=-\infty}^{t} \begin{bmatrix} X(t-s) \\ U(t-s) \end{bmatrix} v_s, \quad |X(t)| \in l_2(0,\infty), \quad |U(t)| \in l_2(0,\infty). \tag{26.61}$$

Here $X(t), U(t)$ are real deterministic $n \times n$ and $n \times m$ matrices and $|X(t)| \in l_2(0,\infty)$. As usually this means:

$$\|X(t)\|^2 = \sum_{t=0}^{\infty} |X(t)|^2 < \infty.$$

It is easy to verify that $\begin{bmatrix} x_t \\ u_t \end{bmatrix}$ is a stationary process. The definition (26.61) means that the value $\begin{bmatrix} x_t \\ u_t \end{bmatrix}$ depends on v_s linearly and that it can not depend on "future" values of the disturbance. We call the process $z = \left\{ \begin{bmatrix} x_t \\ u_t \end{bmatrix} \right\}_{-\infty}^{+\infty}$ **admissible** if $z \in \mathbb{Z}$ and it satisfies the equation (26.56). Every implementable linear stabilizing regulator $D(\sigma)u_t = N(\sigma)x_t$ (where $D(\lambda), N(\lambda)$ are the matrix polynomials and σ is forward-shift operator) together with (26.56) determines the admissible process. In this sense the class of admissible processes includes all processes determined by linear stabilizing regulators.

Thus our problem is to minimize the functional (26.59) subject to constraints (26.56), (26.58), (26.61). Certainly we also wish to find implementable stabilizing regulator which yields the optimal process. (Because of its stabilizability, any process in a closed-loop system will differ from this stationary process at the process decreasing to zero. Therefore this regulator will be optimal in the similar problem in which the sign of mathematical expectation $\mathbf{E}(\ldots)$ in the formulas (26.58), (26.59) is replaced by $\lim_{t\to\infty} \mathbf{E}(\ldots)$.)

It is evidently that many practically important problems may be reduced to the mentioned above mathematical formulation.

Note that the considered problem can be reduced to the deterministic problem of Section 26.3. But it is more interesting to solve this problem independently of Section 26.3 to clear the application of the general method.

Reduction to the abstract scheme.

According to our definition $\mathbb{Z}$ is the real Hilbert space of all stationary processes $z = \{z_t\}_{-\infty}^{+\infty}$, $z_t \in \mathbb{R}^{n+m}$, which have the representation

$$z_t = \sum_{s=-\infty}^{t} Z(t-s)v_s, \qquad |Z(t)| \in l_2(0,\infty), \tag{26.62}$$

where $Z(t)$ is $(n+m)\times l$ matrix function. (We will write (26.62) as $z \sim Z(t)$.) The scalar product in $\mathbb{Z}$ is defined in the usual way: if $z' = \{z'_t\}_{-\infty}^{+\infty}$, $z'' = \{z''_t\}_{-\infty}^{+\infty}$ then

$$(z', z'') = \mathbf{E}\,[(z''_t)^* z'_t] \qquad (= \text{const})$$

If according to (26.62) $z' \sim Z'(t)$, $z'' \sim Z''(t)$, then clearly

$$(z', z'') = \sum_{t=0}^{\infty} \operatorname{tr}[Z'(t)Z''(t)^*].$$

(It is necessary to use the formula $(z', z'') = \mathbf{E}\operatorname{tr}[z'_t(z''_t)^*]$.) The functionals Φ_j, $j = 0, 1, \ldots, \nu$ defined by (26.58), (26.59) are the quadratic in $z = \left\{ \begin{bmatrix} x_t \\ u_t \end{bmatrix} \right\}_{-\infty}^{+\infty} \in \mathbb{Z}$. Thus our problem is a special case of the general problem described in Section 26.2.

Note that this problem differs from the well known one, because of the presence the quadratic constraints (26.58). The functional Φ_j can be nonconvex, so we have a special nonconvex (in general) global minimization problem.

Let us apply the method described in Section 26.2 and show that the conditions (i)–(iii) of Theorem 26.3 are satisfied. This in turn imples that our method is applicable.

Let us define T_k as the shift operators: if $z' = \{z'_t\}_{-\infty}^{+\infty}$, $z = \{z_t\}_{-\infty}^{+\infty}$ then $z' = T_k z$ means that $z'_t = z_{t-k}$. Let $z' \sim Z'(t)$, $z'' \sim Z''(t)$. Then $T_k z' \sim Z_k(t)$, where $Z_k(0) = 0, \ldots, Z_k(k-1) = 0$, $Z_k(k) = Z'(0)$, $Z_k(k+1) = Z'(1), \ldots$ and by (26.63)

$$(T_k z', z'') = \sum_{t=0}^{\infty} \operatorname{tr}[Z_k(t)Z''(t)^*] = \sum_{t=k}^{\infty} \operatorname{tr}[Z'(t-k)Z''(t)^*].$$

Hence

$$|(T_k z', z'')|^2 \le \sum_{t=0}^{\infty} |Z'(t)|^2 \cdot \sum_{t=k}^{\infty} |Z''(t)|^2 \to 0 \quad \text{as } k \to \infty.$$

Consequently the condition (i) is satisfied. It is obvious that the equation $x_{t+1} = Ax_t + Bu_t$ is invariant under shifts, i.e. $T_k\mathcal{M} \subset \mathcal{M}$ and (ii) holds. The quadratic forms Φ_j^0 of the quadratic functionals Φ_j are $\Phi_0^0 = \mathbf{E}\,\mathcal{G}_0(x_t, u_t)$, $\Phi_j^0 = -\mathbf{E}\,\mathcal{G}_j(x_t, u_t)$, $j = 1, \ldots, \nu$ and they are invariant under shifts. So (iii) holds too. By Theorem 26.3 the procedure (I)–(IV) is applicable.

Application of the procedure (I)–(IV).

Let us form the Lagrangian of the problem:

$$S(\tau, z) = \Phi_0 - \sum_{j=1}^{k} \tau_j \Phi_j = \mathbf{E}\,\{z_t^* \widehat{G}(\tau) z_t - \sum_{j=1}^{k} \tau_j \gamma_j, \tag{26.63}$$

where $z = \{z_t\}$, $\tau_1 \ge 0, \ldots, \tau_k \ge 0$ and

$$\widehat{G}(\tau) = \widehat{G}_0 + \sum_{j=1}^{k} \tau_j \widehat{G}_j, \quad \widehat{G}_j = \begin{bmatrix} G_j & g_j \\ g_j^* & \Gamma_j \end{bmatrix}. \tag{26.64}$$

Step (I) of our procedure requires solving the problem (26.6), that is

$$\text{Minimize} \quad \mathbf{E}\{z_t^* \widehat{G}(\tau) z_t\} \quad \text{subject to} \quad \{z_t\} \in \mathcal{L}, \tag{26.65}$$

where $\mathcal{L}$ is defined by (26.56) and, more precisely, to find

$$S^0(\tau) = \inf_{\{z_t\} \in \mathcal{L}} \mathbf{E}\{z_t^* \widehat{G}(\tau) z_t\} - \sum_{j=1}^{k} \tau_j \gamma_j.$$

It is standart LQ-problem, if we ignore for now that we need **a complete** solution. (We have to find cases in which infimum is finite, in which it is attained or not and to find all the corresponding optimal regulators if its exist.) So now we have obtained the auxiliary problem without quadratic constraints.

Solution of the auxiliary problem without quadratic constraints.

The complete solution of this problem (the problem (26.56)–(26.60), (26.65)) is given by the similar to the solution of auxiliary problem in Section 26.3 and we omit the details. First we have to determine the function $\Pi(\lambda)$ for our case (see Appendix). The identity (26.87) (Appendix) in our case takes the form:

$$(Ax + Bu)^* P(Ax + Bu) - x^* P x + \begin{bmatrix} x \\ u \end{bmatrix}^* \widehat{G}(\tau) \begin{bmatrix} x \\ u \end{bmatrix} = |\kappa(u - K_0 x)|^2. \tag{26.66}$$

Let $P = P(\tau) = P(\tau)^*$, $K_0 = K_0(\tau)$, $\kappa = \kappa(\tau) = \kappa(\tau)^*$ be the "stabilizing" solution of this identity. (Recall, this means that $\widehat{A}(\tau) = A + BK(\tau)$ has all eigenvalues in $|\lambda| \le 1$ (not < 1 !))

The matrix $\Pi(\lambda) = \Pi(\lambda, \tau)$, defined in Appendix, takes the form:

$$\Pi(\lambda, \tau) = \begin{bmatrix} A_\lambda^{-1} B \\ I_m \end{bmatrix}^* \widehat{G}(\tau) \begin{bmatrix} A_\lambda^{-1} B \\ I_m \end{bmatrix} \tag{26.67}$$

From (26.64)

$$\Pi(\lambda, \tau) = \sum_1^k \tau_j \begin{bmatrix} A_\lambda^{-1} B \\ I_m \end{bmatrix}^* \begin{bmatrix} G_j & g_j \\ g_j^* & \Gamma_j \end{bmatrix} \begin{bmatrix} A_\lambda^{-1} B \\ I_m \end{bmatrix} = \sum_1^k \tau_j \Pi_j(\lambda).$$

Here $|\lambda| = 1$, $A_\lambda = \lambda I_n - A$, $\delta(\lambda) = \det(\lambda I_n - A) \neq 0$. Let us consider the weak and the strong frequency conditions (WFC and SFC):

(WFC) $$\Pi(\lambda, \tau) \ge 0 \qquad (\forall\, \lambda :\ |\lambda| = 1,\ \delta(\lambda) \neq 0),$$

(SFC) $$\begin{cases} \Pi(\lambda, \tau) > 0 \quad (\forall\, \lambda :\ |\lambda| = 1,\ \delta(\lambda) \neq 0), \\ |\delta(\lambda)|^2 \Pi(\lambda, \tau) > 0 \quad (\forall\, \lambda :\ |\lambda| = 1,\ \delta(\lambda) = 0). \end{cases}$$

(The second condition is understood as a limit and it is absent if $\delta(\lambda) \neq 0,\ \forall\, \lambda$, $|\lambda| = 1$.) Define the sets D and D^0 in the τ-space:

$$D = \{\tau :\ \tau \ge 0, \quad (WFC) \quad \text{holds}\} \tag{26.68}$$

$$D^0 = \{\tau : \ \tau \geq 0, \quad (SFC) \quad \text{holds}\} \tag{26.69}$$

As before (see Lemma 26.1, Section 26.3) we obtain

$$S^0(\tau) = -\infty \quad \text{if} \quad \tau \notin D. \tag{26.70}$$

Note that the solution P, K_0, κ of the identity (26.67) does not exist in this case. If $D = \varnothing$, then $S^0(\tau) \equiv -\infty$ for all $\tau \geq 0$. By duality relations (26.10), $\inf \Phi_0 = -\infty$. Let $D \neq \varnothing$. Like the Section 26.3 we obtain

$$S^0(\tau) = \mathrm{tr}\,[P(\tau)Q] - \sum_{j=1}^{k} \tau_j \gamma_j \quad \text{if} \quad \tau \in D. \tag{26.71}$$

So we have determined $S^0(\tau)$ and have completed the step (I) of our procedure. According to the step (II), we have to find $\tau^0 \geq 0$ which is a certain solution of finite-dimensional convex problem:

$$\text{Maximize} \quad S^0(\tau) = \mathrm{tr}\,[P(\tau)Q] - \sum_{j=1}^{k} \tau_j \gamma_j \quad \text{subject to} \quad \tau \in D. \tag{26.72}$$

The solution τ^0 exists. Two cases are possible: $\tau^0 \in D^0$ and $\tau^0 \in D \backslash D^0$.

Consider the case $\tau^0 \in D^0$. It will be easier now to use the Theorem 26.5. Because the SFC holds the functional $\Phi(z) = \mathbf{E}\,\{z_t^* \widehat{G}(\tau^0) z_t\}$ is strictly positive on $\mathcal{M}$. By Theorem 26.5 the procedure (I)–(IV) is applicable, the value $z^0 = z(\tau^0)$ is determined uniquely and we do not need the step (IV). The process $z^0 = z(\tau^0)$ is the optimal process in our constrained optimization problem and it is determined by the regulator

$$u_t = K_0(\tau^0) x_t. \tag{26.73}$$

This is answer in our problem for the case of $\tau^0 \in D^0$.

Consider the case of $\tau^0 \in D \backslash D^0$ and the step (III) of our procedure. In this case the matrix $\widehat{A}(\tau^0) = A + BK(\tau^0)$ has all eigenvalues in $|\lambda| \leq 1$ and part of them lies on the circle $|\lambda| = 1$. For any admissible process and for any $\tau \geq 0$ it follows from (26.66)

$$\mathbf{E}\,\{z_t^* \widehat{G}(\tau) z_t\} = \mathrm{tr}\,\{P(\tau)Q\} + \mathbf{E}\,|\kappa(\tau)(u_t - K_0(\tau)x_t)|^2. \tag{26.74}$$

Therefore (see (26.63))

$$S(\tau^0, z) = \mathrm{tr}\,\{P(\tau^0)Q\} + \mathbf{E}\,|\kappa(\tau^0)(u_t - K_0(\tau)x_t)|^2 - \sum_{j=1}^{k} \tau_j \gamma_j \tag{26.75}$$

and (see (26.71))

$$\inf_{z \in \mathcal{L}} S(\tau^0, z) = \mathrm{tr}\,\{P(\tau^0)Q\} - \sum_{j=1}^{k} \tau_j^0 \gamma_j.$$

Therefore $\inf_{z\in L} S^0(\tau, z)$ is attained if and only if the regulator $u_t = K(\tau^0)x_t$ performs an admissible process. Let us substitute (26.60) in (26.56) in the equation $u_t = K(\tau^0)x_t$ and use the orthogonality properties $\mathbf{E}\, v_t v_s^* = \delta_{ts}$. We obtain

$$X(t+1) = AX(t) + BU(t), \quad X(0) = C, \quad U(t) = K_0(\tau^0)X(t).$$

The condition $|X(t)| \in l_2$, $|U(t)| \in l_2$, is fulfiled if and only if the columns of C belong to stable subspace of the matrix $\widehat{A}(\tau^0)$. Suppose this condition holds. According to the action (IV) of our procedure we must verify the condition (26.9), that is

$$\mathbf{E}\, \mathcal{G}_0(x_t, u_t) \le \gamma_j, \quad \tau_j^0[\gamma_j - \mathbf{E}\, \mathcal{G}_0(x_t, u_t)] = 0, \quad j = 1, \dots, k. \tag{26.76}$$

If it holds then the controller

$$u_t = K_0(\tau^0)x_t \tag{26.77}$$

gives a solution of our problem. (But this solution is parametrically unstable.) In opposite case by Theorem 26.3 the problem under consideration has no solution.

Formulation of result.

Let $\Pi(\lambda, \tau)$ is defined by (26.67) and the sets D and D^0 are defined by (26.68), (26.69).

THEOREM 26.10
If $D = \varnothing$ then in the primary problem (26.61) $\inf\Phi_0 = -\infty$. □

Let $D \neq \varnothing$. For $\tau \in D$ there exists a "stabilizing solution" — the matrices $P(\tau), K_0(\tau), \kappa(\tau)$ which satisfy the identity (26.67):

$$(Ax + Bu)^* P(\tau)(Ax + Bu) - x^* P(\tau)x + \begin{bmatrix} x \\ u \end{bmatrix}^* G(\tau) \begin{bmatrix} x \\ u \end{bmatrix} =$$

$$|\kappa(\tau)(u - K_0(\tau)x)|^2 \tag{26.78}$$

(or equivalently the Lur'e-Riccati equations (26.47) with $K(\tau) = K_0(\tau)$), such that all eigenvalues of the matrix $\widehat{A}(\tau) = A + BK_0(\tau)$ lie in $|\lambda| \le 1$.

Define $S^0(\tau)$ by (26.71). Let $\tau^0 \ge 0$ be any solution of the following finite-dimensional convex optimization problem:

$$S^0(\tau^0) \ge S^0(\tau) \quad \text{for all} \quad \tau \ge 0,\ \tau \in D. \tag{26.79}$$

(The solution τ^0 exists.)

THEOREM 26.11
If $\tau^0 \in D^0$ then there exists (and it is unique) the optimal process in our constraint optimization problem and it is determined by the regulator

$$u_t = K_0(\tau^0)x_t. \tag{26.80}$$

□

THEOREM 26.12
If $\tau^0 \in D \backslash D^0$ then the optimal process in our constrained optimization problem exists if and only if the columns of the matrix C belong to a strictly stable subspace of the matrix $\widehat{A}(\tau^0) = A + BK_0(\tau^0)$ and the conditions (26.76) hold. In this case optimal process is determined by the regulator $u_t = K_0(\tau^0)x_t$. (This regulator is parametrically unstable.) □

THEOREM 26.13
For $\inf_{\Phi_j \geq 0}\Phi_0$ the following formula takes place:

$$\inf_{\Phi_j \geq 0}\Phi_0 = \mathrm{tr}\,[C^* P(\tau^0)C] - \sum_{j=1}^{k} \tau_j^0 \gamma_j.$$

□

26.5 Appendix. (Discrete KYP-Lemma.) The Frequency-Domain Method to Solve Discrete Lur'e-Riccati Equations

Consider the following inequality for an unknown real $n \times n$ matrix $P = P^*$:

$$(Ax + Bu)^* P(Ax + Bu) - x^* Px + \mathcal{G}(x,u) \geq 0 \quad (\forall\, x,\ \forall\, u). \tag{26.81}$$

Here $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, A, B are real $n \times n$ and $n \times m$ matrices, the pair (A, B) is a controllable, $\mathcal{G}(x,u)$ is a given real quadratic form in x, u, the inequality (26.81) has to be satisfied for every $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$. If $n = m = 1$, then (26.81) reduces to a quadratic inequality for real number P. The inequality (26.81) has a real solution $P = P^*$ only if its coefficients satisfy some relations to be found. The well known KYP-lemma (frequency-domain theorem) defines these conditions. We formulate this lemma below for the convenience of our readers, following [16]. It differs in details from other formulations.

Let $\widetilde{x} \in \mathbb{C}^n$, $\widetilde{u} \in \mathbb{C}^m$ be complex vectors and $\mathcal{G}(\widetilde{x}, \widetilde{u})$ be the Hermitian extention of $\mathcal{G}(x,u)$. If

$$\mathcal{G}(x,u) = x^* Gx + 2x^* gu + u^* \Gamma u, \tag{26.82}$$

where $G = G^*$, g, $\Gamma = \Gamma^*$ are real $n \times n$, $n \times m$ and $m \times m$ matrices and $(\ldots)^*$ means the transposition, then

$$\mathcal{G}(\widetilde{x}, \widetilde{u}) = \widetilde{x}^* G\widetilde{x} + 2\mathrm{Re}(\widetilde{x}^* g\widetilde{u}) + \widetilde{u}^* \Gamma \widetilde{u}, \tag{26.83}$$

where $(\ldots)^*$ means Hermitian conjugation.

Let us define the Hermitian matrix $\Pi(\lambda) = \Pi(\lambda)^*$ (for $\lambda \in \mathbb{C}$, $|\lambda| = 1$) by the relation

$$\mathcal{G}(\widetilde{x}, \widetilde{u}) = \widetilde{u}^* \Pi(\lambda)\widetilde{u} \quad \text{for } \lambda\widetilde{x} = A\widetilde{x} + B\widetilde{u},\ |\lambda| = 1,\ \det(\lambda I_n - A) \neq 0. \tag{26.84}$$

As before, let us use the notation:

$$A_\lambda = \lambda I_n - A, \quad (\text{for } \lambda \in \mathbb{C}) \quad \delta(\lambda) = \det A_\lambda, \quad Q_\lambda = \delta(\lambda) A_\lambda^{-1}. \tag{26.85}$$

Clearly,

$$\Pi(\lambda) = \begin{bmatrix} A_\lambda^{-1}B \\ I_m \end{bmatrix}^* \begin{bmatrix} G & g \\ g* & \Gamma \end{bmatrix} \begin{bmatrix} A_\lambda^{-1}B \\ I_m \end{bmatrix} \quad (\lambda \in \mathbb{C},\ |\lambda| = 1,\ \delta(\lambda) \neq 0) \tag{26.86}$$

Let us introduce the conditions which we will call respectively the weak and the strong frequency-domain conditions (WFC and SFC). WFC is defined by

(WFC) $\qquad \Pi(\lambda) \geq 0 \qquad$ for $\quad \lambda \in \mathbb{C}, \quad |\lambda| = 1, \quad \delta(\lambda) \neq 0,$

If $\delta(\lambda) \neq 0$ for $|\lambda| = 1$ (it means that $\Pi(\lambda)$ is defined for all λ, $|\lambda| = 1$) then SFC is the following condition:

(SFC) $\qquad \Pi(\lambda) > 0 \qquad$ for $\quad \lambda \in \mathbb{C}, \quad |\lambda| = 1.$

If $\delta(\lambda)$ has the roots λ_j, $|\lambda_j| = 1$, then SFC is given by

$$\text{(SFC)} \qquad \begin{cases} \Pi(\lambda) > 0 \quad \text{for} \quad \lambda \in \mathbb{C},\ |\lambda| = 1,\ \delta(\lambda) \neq 0 \quad \text{and} \\ \lim\limits_{\lambda \to \lambda_j} |\delta(\lambda)|^2 \Pi(\lambda) > 0 \quad \text{for } \lambda_j :\ |\lambda_j| = 1,\ \delta(\lambda_j) = 0. \end{cases}$$

Note that (in general case) (SFC) can also be defined as follows:

$$\text{(SFC)} \qquad \begin{cases} \exists\, \varepsilon > 0 : \mathcal{G}(\widetilde{x}, \widetilde{u}) \geq \varepsilon(|\widetilde{x}|^2 + \widetilde{u}|^2) \text{ for all } \widetilde{x} \in \mathbb{C}^n,\ \widetilde{u} \in \mathbb{C}^m,\ \lambda \in \mathbb{C}, \\ \text{such that} \quad \lambda\widetilde{x} = A\widetilde{x} + B\widetilde{u}, \quad |\lambda| = 1. \end{cases}$$

Very often $m = 1$ and $\delta(\lambda) \neq 0$ for $|\lambda| = 1$. In this "scalar" case $\Pi(\lambda)$ is a real function defined on the circle $|\lambda| = 1$ and the WFC and the SFC mean that $\Pi(e^{i\theta}) \geq 0$ or $\Pi(e^{i\theta}) > 0$ respectively for $0 \leq \theta \leq 2\pi$.

THEOREM 26.14—DISCRETE KYP-LEMMA
Let (A, B) be a controllable pair. The inequality (26.81) has a real solution $P = P^*$ if and only if the WFC holds. Furthermore, then there exist real matrices $P = P^*$, K, $\kappa = \kappa^*$ of the dimensions $n \times n$, $n \times m$, $m \times m$ respectively satisfying the following identity:

$$(Ax + Bu)^* P(Ax + Bu) - x^* P x + \mathcal{G}(x, u) = |\kappa(u - Kx)|^2 \ (\forall\, x \in \mathbb{R}^n,\ \forall\, u \in \mathbb{R}^m). \tag{26.87}$$

and such that

$$\kappa > 0, \quad \det[\lambda I_n - (A + BK)] \neq 0 \quad \text{for} \quad |\lambda| > 1. \tag{26.88}$$

If the SFC holds then there exist $P = P^*$, K, $\kappa = \kappa^*$, such that (26.87), (26.88) hold and the second inequality in (26.88) is true for $|\lambda| \geq 1$. Then matrices P, K, κ are defined uniquely. □

We shall call the solution P, K, κ the *stabilizing* solution. In the later case (if the SFC holds) we shall call P, K, κ the *strictly stabilizing* solution. Note that (26.87) can have other (nonstabilizing) solutions.

The proof can be found in [16] where the most general theorem is formulated and proved. Note that if the SFC holds then the last assertion of Theorem 26.14 remains true if the assumption of controllability of (A, B) is replaced by the weaker condition that (A, B) is stabilizable.

The identity (26.87) can be rewritten as well known Lur'e-Riccati equations

$$\left.\begin{aligned} &A^*PA - P + G = K^*\kappa^2 K \\ &B^*PA + g^* = -\kappa^2 K \\ &B^*PB + \Gamma = \kappa^2 \end{aligned}\right\} \tag{26.89}$$

There are many methods of solving the Lur'e-Riccati equations. Let us formulate for the "scalar" case $m = 1$ the frequency-domain method of determining P, K, κ. This method is close to the method of Kalman and Szegö [23] and follows from [16], where it is described for any $m \geq 1$.

THEOREM 26.15
Let $m = 1$, (A, B) be a controllable pair and let the WFC be true. The real matrices $P = P^*$, K and the real number κ can be determined in the following way.

(I) Determine $\Pi(\lambda)$ and $\varphi(\lambda) = |\det(\lambda I_n - A)|^2\Pi(\lambda)$ for $|\lambda| = 1$. The function $\varphi(\lambda)$ is the quasypolynomial

$$\varphi(\lambda) = \varphi_n\lambda^n + \ldots + \varphi_0 + \ldots + \varphi_{-n}\lambda^{-n} \tag{26.90}$$

with real coefficients $\varphi_j = \varphi_{-j}$, and therefore $\varphi(\lambda)$ can be extended for all $\lambda \in \mathbb{C}$, $\lambda \neq 0$ by this formula.

(II) Determine the polynomial $\psi(\lambda) = \psi_n\lambda^n + \ldots + \psi_0$ (with real coefficients ψ_j and $\psi_n > 0$) from the factorization relation $\varphi(\lambda) = \psi(\lambda)\psi(\lambda^{-1})$ $(\forall\, \lambda \in \mathbb{C},\ \lambda \neq 0)$. Put $\kappa = \psi_n$.

(III) Determine the $1 \times n$ matrix K from the identity

$$\det[\lambda I_n - (A + BK)] = \kappa^{-1}\psi(\lambda) \qquad (\forall\, \lambda \in \mathbb{C}). \tag{26.91}$$

(IV) Determine the real matrix $P = P^*$ from the identity (26.87). The solution exists and it is unique. □

The polynomial $\psi(\lambda|$ may be determined by a well known method. Note that $\varphi(\lambda)$ is a real symmetric quasypolynomial, so its roots are symmetric with respect to the real axes and the unit circle. We can separate them into two symmetric groups. Let us take the roots of one group as a roots of polynomial $\psi(\lambda)$. If $\varphi_n = 0$ then the number of roots in each group is $q < n$, in this case we complement them with $n - q$ roots equal to zero. So we obtain $\psi(\lambda)$ such that $\varphi(\lambda) = \psi(\lambda)\psi(\lambda^{-1})$ and $\psi_n \neq 0$.

If this group containes only roots in the disk $|\lambda| \leq 1$ (in the open disk $|\lambda| < 1$) we obtain the stabilising solutoin (the strictly stabilising solution). The equation (26.91) is transformed into a system oflinear equations for the coefficients of the matrix K. In fact it can be rewritten as

$$\delta(\lambda) - KQ_\lambda B \equiv \lambda^n + \frac{\psi_{n-1}}{\kappa}\lambda^{n-1} + \ldots + \frac{\psi_0}{\kappa}.$$

We have

$$Q_\lambda B = \lambda^{n-1} q_{n-1} + \ldots + q_0.$$

Here the vectors q_j are linearly independent because of the controllability of (A, B). So the $1 \times n$ matrix K is determined (uniquely) from the equations

$$\delta_j - K q_j = \frac{\psi_j}{\kappa}, \qquad j = 0, 1, \ldots, n-1.$$

26.6 References

[1] Kalman R.E., *Contributions to the theory of optimal control*, Bol. Soc. Math. Mexicana (2) **5** (1960), 102-119.

[2] Krasovskii N.N., *Stabilization problem of a control motion*, in Supplement 2 to I.G. Malkin, *Stability Theory of Motion* [in Russian], Nauka, Moscow (1976).

[3] Letov A.M., *Analitical regulator synthesis*, Avtomat. Telemekh., No. 4, (1960), 436-446; No. 5, (1960), 561-571.

[4] Lur'e A.I., *Minimal quadratic performance index for a control system*, Tekh. kibern., No. 4, (1963), 140-146 [in Russian].

[5] Kolmogorov A.N., *Interpolation and extrapolation of random sequences*, Izv. Akad. Nauk SSSR, Ser. Mat., No. 5, (1941), 3-14 [in Russian].

[6] Wiener N., *Extrapolation, interpolation and smoothing of stationary time series*. Cambridge, 1949.

[7] Bucy R.S. and Joseph P.D., *Filtering of stochastic processes with application to guidance*. N.Y., London, 1968.

[8] Yakubovich V.A., *Minimization of quadratic functionals under quadratic constraints and the necessity of the frequency condition for absolute stability of nonlinear control systems*, Dokl. Akad. Nauk SSSR, **209**, (1973), 1039-1042. (English transl. in Soviet Math. Dokl., **14**, No. 2, (1973), 593-597.)

[9] Yakubovich V.A., *A method for solution of special problems of global optimization*, Vestn. St. Petersburg Univ., Ser. 1, Vyp. 2 No. 8, (1992), 58-68. (Engl. transl. in Vestnik St. Petersburg Univ., Math., **25**, No. 2, (1992), 55-63.)

[10] Yakubovich V.A., *Nonconvex optimization problems: the infinite-horizon linear-quadratic problems with quadratic constraints*, Systems & Control Letters **16** (1992), 13-22.

[11] Yakubovich V.A., *Linear-quadratic optimization problems with quadratic constraints*, Proc. of the Second European Control Conf., the Netherlands **1** (1993), 346-359.

[12] Matveev A.S.,*Tests for convexity of images of quadratic mappings in optimal control theory of systems described by differential equations*, Dissertation, St. Petersburg, (1998) [in Russian].

[13] Matveev A.S., *Lagrange duality in nonconvex optimization theory and modifications of the Toeplitz-Hausdorf theorem*, Algebra and Analiz, **7**, No. 5, (1995), 126-159 [in Russian].

[14] Matveev A.S., *Lagrange duality in a specific nonconvex global optimization problem*, Vestn. St. Petersburg Univ.,Ser. Math., Mekh., Astron., No. 8, (1996), 37-43 [in Russian].

[15] Matveev A.S., *Spectral approach to duality in nonconvex global optimization*, SIAM J. Control and Optimization, vol.36 (1998), No. 1, 336-378.

[16] Yakubovich V.A. *A frequency theorem in control theory*, Sibirskii Mat. Z., **14**, vyp. 2, (1973), 384-419. (English transl. in Siberian Math. J., **14**, No. 2, (1973), 265-289.)

[17] Popov V.M. *Hiperstabilitea sistemlor automate* Editura Academiei Republicii Socialiste Romania.

[18] Lindquist A., Yakubovich V.A., *Universal Controllers for Optimal Damping of Forced Oscillations in Linear Discrete Systems*, Doklady Akad. Nauk, **352**, No. 3, (1997), 314-317. (English transl. in Doklady Mathematics, **55**, No. 1, (1997), 156-159.)

[19] Lindquist A., Yakubovich V.A., *Optimal Damping of Forced Oscillations in Discrete-Time Systems*, IEEE Transactions on Automatic Control, **42**, No. 6, (1997), 786-802.

[20] Lindquist A., Yakubovich V.A., *Universal Regulators for Optimal Signal Tracking in Linear Discrete Systems*, Doklady Akad. Nauk, **361**, No. 2, (1998) 177-180. (English transl. in Doklady Mathematics, **58**, No. 1, (1998), 165-168.)

[21] Lindquist A., Yakubovich V.A., *Universal Regulators for Optimal Tracking in Discrete-Time Systems Affected by Harmonic Disturbances*, IEEE Transactions on Automatic Control, AC-44, (1999), 1688-1704.

[22] Luenberger D.G., *Optimization by vector space methods*. John Wiley & Sons, Inc., New York – London.

[23] Szegö G., Kalman R.E., *Sur la stabilite absolute d'un sisteme d'equations aux differences finies*, C. R. Acad. Sci., **257**, (1963), 388-390.

Author List

Systems with Lebesgue Sampling

Karl Johan Åström
Department of Automatic Control
Lund Institute of Technology
Box 118
SE-221 00 Lund
Sweden

Bo Bernhardsson
Ericsson Mobile Platforms
Technology Strategies
Sölvegatan 53
SE-221 83 Lund
Sweden

Acoustic Attenuation Employing Variable Wall Admittance

H. T. Banks, K. Ito, N. S. Luke, C. J. Smith
Center for Research in Scientific Computation
Department of Mathematics
Box 8205
NCSU
Raleigh, NC 27695-8205
USA

K. M. Furati
Department of Mathematical Sciences
KFUPM
Dhahran 31261
Saudi Arabia

Some Remarks on Linear Filtering Theory for Infinite Dimensional Systems

Alain Bensoussan
Président du CNES
2 place Maurice Quentin
75039 PARIS cedex 01
France

A Note on Stochastic Dissipativeness

Vivek S. Borkar
School of Technology & Computer Science
Tata Institute of Fundamental Research
Mumbai
INDIA

Sanjoy K. Mitter
Department of Electrical Engineering and Computer Science
MIT
Cambridge, MA 02139
USA

Internal Model Based Design for the Suppression of Harmonic Disturbances

Christopher I. Byrnes, Alberto Isidori
School of Engineering and Applied Science
Washington University
One Brookings Drive
St. Louis, MO 63130
USA

David S. Gilliam
Department of Mathematics
Texas Tech University
Lubbock, TX 79409
USA

Yutaka Ikeda
The Boeing Company

Lorenzo Marconi
Università di Bologna
Viale Risorgimento 2
40136 Bologna
Italy

Conditional Orthogonality and Conditional Stochastic Realization

P. E. Caines
McGill University
Department of Electrical and Computer Engineering
McConnell Engineering Building, Rm.512
3480 University Street
Montreal, Québec
Canada H3A-2A7

R. Deardon and H. P. Wynn
Department of Statistics
University of Warwick
Coventry CV4 7AL
United Kingdom

Geometry of Oblique Splitting Subspaces, Minimality and Hankel Operators

Alessandro Chiuso and Giorgio Picci
Department of Information Engineering
University of Padova
via Gradenigo 6A
35131 Padova
Italy

Linear Fractional Transformations

Harry Dym
Department of Mathematics
The Weizmann Institute
Rehovot 76100
Israel

Structured covariances and related approximation questions

Tryphon T. Georgiou
Dept. of Electrical and Computer Engineering
University of Minnesota
200 Union street SE
Minneapolis, MN 55455
USA

Risk Sensitive Identification of ARMA Processes

László Gerencsér
Computer and Automation Institute of the Hungarian Academy of Sciences
H-1111 Kende 13-17
Budapest
Hungary

György Michaletzky
Department of Probability Theory and Statistics
Eötvös Loránd University
H-1117 Pázmány Péter sétány 1/C
Budapest
Hungary

Input Tracking and Output Fusion for Linear Systems

Xiaoming Hu, Ulf T. Jönsson
Division of Optimization and Systems Theory
Royal Institute of Technology
10044 Stockholm
Sweden

Clyde F. Martin
Department of Mathematics and Statistics
Texas Tech University
Lubbock, TX 79409
USA

The Convergence of the Extended Kalman Filter

Arthur J. Krener
Department of Mathematics
University of California
Davis, CA 95616-8633
USA

On the Separation of Two Degree of Freedom Controller and its Application to H_∞ Control for Systems with Time Delay

Yohei Kuroiwa and Hidenori Kimura
The University of Tokyo
Graduate School of Frontier Sciences
Department of Complexity Science and Engineering
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033
JAPAN

The Principle of Optimality in Measurement Feedback Control for Linear Systems

Alexander B. Kurzhanski
Moscow State (Lomonosov) University
Faculty of Computational Mathematics and Cybernetics
119992 Moscow
Russia

Linear System Identification as Curve Fitting

Lennart Ljung
Division of Automatic Control
Linköping University
SE-581 83 Linköping
Sweden

Optimal Model Order Reduction for Maximal Real Part Norms

A. Megretski
35-418 EECS MIT
Cambridge, MA 02139
USA

Quantum Schrödinger Bridges

Michele Pavon
Dipartimento di Matematica Pura e Applicata
Università di Padova
via Belzoni 7
and LADSEB-CNR
35131 Padova
Italy

Segmentation of Diffusion Tensor Imagery

Eric Pichon, Allen Tannenbaum
Department of Electrical and Computer Engineering
777 Atlantic Drive
Georgia Institute of Technology
Atlanta, GA 30332-0250
USA

Guillermo Sapiro
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455
USA

Robust Linear Algebra and Robust Aperiodicity

Boris T. Polyak
Institute for Control Science
Profsojuznaja 65
Moscow 117806
Russia

On Homogeneous Density Functions

Stephen Prajna
Control and Dynamical Systems
California Institute of Technology
1200 E. California Blvd
Pasadena, CA 91125
USA

Anders Rantzer
Department of Automatic Control
Lund Institute of Technology
Box 118
SE-221 00 Lund
Sweden

Stabilization by Collocated Feedback

Olof J. Staffans
Åbo Akademi University
Department of Mathematics
FIN-20500 Åbo
Finland

High-order Open Mapping Theorems

Héctor J. Sussmann
Department of Mathematics
Rutgers, the State University of New Jersey
Hill Center—Busch Campus
110 Frelinghuysen Road
Piscataway, NJ 08854-8019
USA

New Integrability Conditions for Classifying Holonomic and Nonholonomic Systems

Tzyh-Jong Tarn, Mingjun Zhang
Campus Box 1040
Washington University
St.Louis, MO 63130-4899
USA

Andrea Serrani
412 Dreese Laboratory
Department of Electrical Engineering
The Ohio State University
2015 Neil Avenue
Columbus, OH 43210
USA

On Spectral Analysis using Models with Pre-Specified Zeros

Bo Wahlberg
S3 - Automatic Control, KTH
SE-100 44 Stockholm
Sweden

Balanced State Representations with Polynomial Algebra

Jan Willems
ESAT-SISTA
K.U. Leuven
Kasteelpark Arenberg 10
B-3001 Leuven-Heverlee
Belgium

Paolo Rapisarda
Maastricht University
Dept. Mathematics
PO.Box 616
6200 MD Maastricht
The Netherlands

Nonconvex Global Optimization Problems: Constrained Infinite-horizon Linear-quadratic Control Problems for Discrete Systems

V.A. Yakubovich
Department of Mathematics and Mechanics
St. Petersburg University
Bibliotechnaya pl. 2, Petrodvoretz
St. Petersburg 198904
Russia