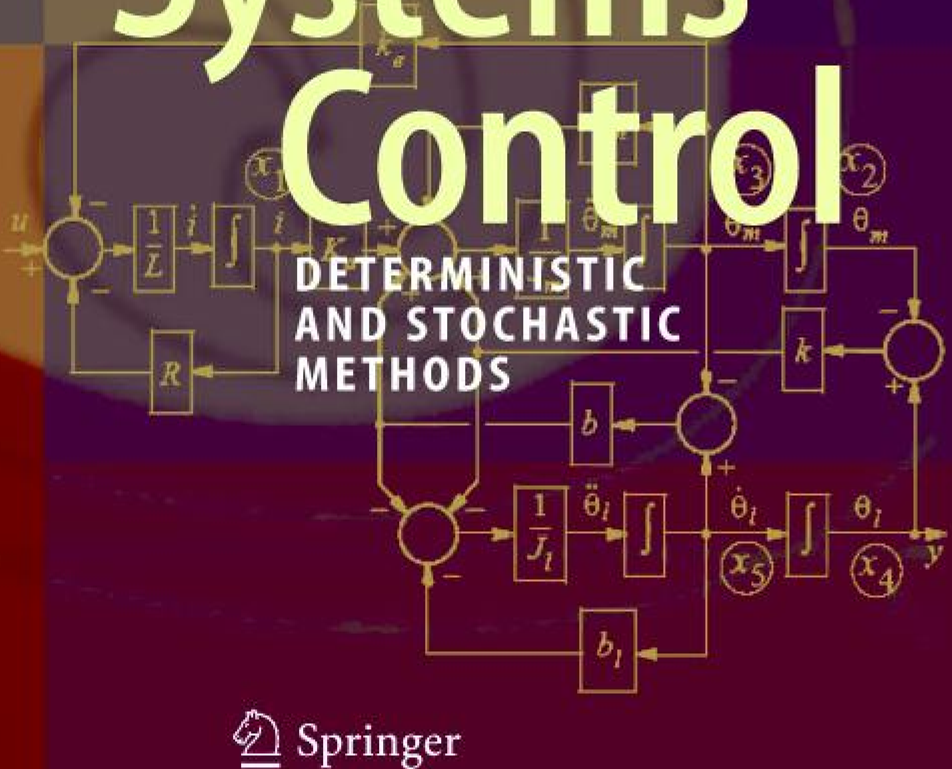
Elbert Hendricks
Ole Jannerup
Paul Haase Sørensen

# Linear Systems Control

## DETERMINISTIC AND STOCHASTIC METHODS



Springer

# Linear Systems Control

Elbert Hendricks • Ole Jannerup •
Paul Haase Sørensen

# Linear Systems Control

## Deterministic and Stochastic Methods

Springer

Elbert Hendricks
Department of Electrical Engineering
Automation
Technical University of Denmark (DTU)
Building 326
DK-2800 Lyngby
Denmark
eh@elektro.dtu.dk

Ole Jannerup
Department of Electrical Engineering
Automation
Technical University of Denmark (DTU)
Building 326
DK-2800 Lyngby
Denmark
oej@elektro.dtu.dk

Paul Haase Sørensen
Creare Inc.
16 Great Hollow Road
Hanover, NH-03755
USA
phs@creare.com

ISBN: 978-3-540-78485-2 e-ISBN: 978-3-540-78486-9

Library of Congress Control Number: 2008927517

© 2008 Springer-Verlag Berlin Heidelberg

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 any other way, 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. Violations are liable to prosecution under the German Copyright Law.

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.

*Cover design:* Kirchner, Erich

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

# Preface

Control engineering is a multidisciplinary subject which is critical to the operation of modern technology and which is nearly invisible in daily life. From atomic force microscopes, through PC disk drives and drive-by-wire engine controls to Boeing 747's and to space shuttles, multivariable controls underlie many of a the devices which are available in the modern world and without which, they would not function at all. This book is an introduction to the more advanced techniques which are used to design control systems for the devices above and which were developed for this purpose from about 1960.

Modern control theory and in particular state space or state variable methods can be adapted to the description of many systems because they depend strongly on physical modelling and physical intuition. The laws of physics are in the form of continuous differential equations and for this reason, this book concentrates on system descriptions in this form. This means coupled sets of linear or nonlinear differential equations. The physical approach is emphasized in this book because it is most natural for complex systems. It also makes what would ordinarily be a mathematical subject into a physical one which can straightforwardly be understood intuitively and which deals with concepts which engineering and science students are already familiar. In this way it is easier to apply the theory to the understanding and control of ordinary systems. Application engineers, working in industry and very pressed for time, should also find this approach useful for this reason. In particular for those in small and middle sized companies, like most of those in Denmark, this will be the case.

In line with the approach set forth above, the book first deals with the modelling of systems in state space form. Both transfer function and differential equation modelling are methods decribed with many examples. Linearization is treated and explained first for very simple nonlinear systems, then more complex systems. Because computer control is so fundamental to modern applications, discrete time modelling of systems as difference equations is introduced immediately after the more intuitive differential and transfer function models. The conversion of differential equation to difference equations is also discussed at length, including transfer function formulations.

An important adjunct to modelling is the analysis of state space models to understand their dynamics. One of the basic reasons for modelling systems is to better understand them: the basis of control is system first understanding and mathematical modelling second. The analysis of models encompasses both continuous and discrete time models and is based on methods of solution of state equations in both forms. This leads naturally to the important questions of first stability and then controllability and observability and finally canonical forms and realizability.

With the intuitive background established as above, the actual control problem for SISO (Single Input/Single Output) and MIMO (Multiple Input/Multiple Output) systems can be attacked in a simple way. This means using tools of analysis already introducted to establish overall stable systems with required time and frequency responses and necessary stability limits. Among the necessary characteristics of control systems may be requirements for zero steady state error in regulation and tracking systems and thus integral controllers are next introduced. As complex models can be effectively and inexpensively built into current control systems, deterministic observers are naturally a part of modern control systems and are introduced next, both for SISO and MIMO systems together with state feedback systems which use observers.

Once state feedback is used, it is natural to consider how it can be optimized and this is the next important subject which is treated. To make this important subject less mystical, the subject of the calculus of variations is described in a simple fashion and is further supported by an appendix on simple static optimization. This leads to the benchmark Linear Quadratic Regulators (LQR) which are the foundation of many modern controllers. Again both continuous time and discrete time regulators are derived, discussed and exemplified in the text.

A vital problem in classical and modern control is how to treat noise in control systems. Nevertheless this question is rarely treated in depth in many control system textbooks because it is considered to be too mathematical and too difficult for a second course on controls. In this textbook a simple physical approach is made to the description of noise and stochastic disturbances which is easy to understand and apply to common systems. This requires only a few fundamental statistical concepts which are given in a simple introduction. The basic noise paradigms, Wiener processes and white noise, are introduced and exemplified with simple physical models. These are then shown to give the Lyapunov equation for noise propagation.

With the Lyapunov equation available, it is a very small step to add the effects of state noise propagation and measurement noise to give the Riccati equation for optimal state estimators or Kalman filters. These important observers are derived and illustrated using simulations in terms that make them easy to understand and apply to real systems. The use of LQR regulators and Kalman filters give LQG (Linear Quadratic Gaussian) regulators which are introduced at the end of the book.

## Detailed Chapter Description

### *Chapter 1: Introduction*

Chapter 1 includes a history of automatic control of linear systems which emphasizes the continuity of the subject and points out the most important developments since ancient history, and briefly, the foundations for them. This includes sections on the most primitive, ancient developments, the pre-classical period, the classical period and finally the modern control period. The main aim of the chapter is to show how modern control theory grew out of specfic technical needs and to fix these developments approximately in time. Another important aim is to show that the understanding of control systems is heavily dependent on the mathematical descriptions which have been created for the study of dynamic systems in physics.

### *Chapter 2: State Space Modelling of Physical Systems*

In this chapter the reduction of physical systems to state space form is covered. This includes a careful description of the linearization process which gives the linear state space form most used in this book. Also treated are companion forms of the type useful in linear systems control and transfer function forms for state space systems. There are a number of examples of the use of the material presented including electrical systems, mechanical systems, as well as the use of the linearization method on common nonlinear control objects. This chapter is supported by a set of problems which can be solved analytically but also with the use of Matlab/Simulink.

### *Chapter 3: Analysis of State Space Models*

The concepts and tools necessary to analyze multivariable dynamic systems are detailed in this chapter. The main topics include solution of the linear state equation, including the transfer function method, natural system modes, modal decomposition, similarity transforms, stability definitions, stability of linear systems and external and internal stability. This consideration of stability leads naturally to the problems of controllability, observability, reachability, detectability and duality. Finally the chapter deals with modal decomposition, realizability and minimal forms. Again the chapter is provided with simple and easily followed examples which are easily remembered. The chapter is concluded with a summary of the most important concepts presented. The solution of a long problem set at the end of this chapter requires the use of Matlab/Simulink and the text in the chapter suggests some standard routines which might be useful for analysis.

## *Chapter 4: Linear Control System Design*

In Chapter 4 the basic theory of state space feedback and observer design is presented. This includes the general design rules for closed loop systems and specifically deals with pole placement and eigenstructure assignment design. The chapter has numerous examples of full state feedback control with both continuous and discrete time controllers. It also describes dead-beat controllers and introduces integral control of systems in multivariable, state space form. The chapter develops full and reduced order deterministic observers and pole placement and eigenstructure assignment for such estimators. The chapter concludes by combining observers, state feedback and integral control for the same control object in the same feedback loop. Both continuous and discrete time controllers and observers are considered. This chapter has examples simulated with Matlab/Simulink.

## *Chapter 5: Optimal Control*

This chapter starts by describing the optimal control and general optimal control problem in terms of the minimization of a performance index. With the help a simplified explanation of the calculus of variations the general optimal control problem is solved. The Linear Quadratic Regulator (LQR) optimal control problem is solved, step by step, both in the open and closed loops, both for continuous and discrete time systems. A careful discussion of the selection of the weighting matrices in the LQR index is given which makes the solution simple to apply to real problems. The steady state suboptimal solution is given and an eigenstructure approach is also detailed. The robustness of LQR control is discussed in a simple fashion which demarkates its limits without going into complex detail. Finally the discrete time LQR regulator is developed after a discussion of the conversion of the performance index to discrete time. In all cases attention is focussed on illustrating the design methods introduced using practical examples. These practical examples are often illustrated with Matlab/Simulink simulations and plots.

## *Chapter 6: Noise in Dynamic Systems*

A very clear and simple presentation is made of the problem of describing noise in continuous dynamic systems in this chapter. The treatment uses the continuous time treatment in order to make use of students' natural intuition about the physics of such systems. This chapter does not require any previous knowledge of random variables or stochastic processes. It introduces the necessary background in a simple, logical and intuitively appealing fashion. This makes it possible to introduce more complex noise descriptions as a natural result of basic statistical concepts. It contains a simple physical picture of noise which is useful for building an intuitive understanding of noise rather than a purely mathematical one. Simple guidelines are given which make it possible to quantify uniform or Gaussian distributed noise for simulation studies. Necessary concepts such as variance, covariance, stationarity, ergodicity and independent increment processes are illustrated with concrete examples and figures. The final results of the chapter are the Lyapunov equations for both continuous time and discrete time systems. These results give a natural introduction to the next chapter on Kalman filtering. As with earlier chapters, Matlab/Simulink support is used for the chapter text and a rather extentive set of problems.

## *Chapter 7: Optimal Observers: Kalman Filters*

This chapter uses the introduction to noise in the preceding chapter to develop Kalman filters as a natural consequence of applying the Lyapunov equation for process (or state) noise propagation in linear systems. When a measurement is added to the description of the process noise propagation, what results is the Riccati equation, the basis of Kalman filtering. While the emphasis in the chapter is on continuous systems for physical reasons, discrete time Kalman filters are shown to be derivable from the continuous filter in a simple way. After the discussion of Kalman filters, the separation theorem is introduced, making it possible to construct Linear Quadratic Gaussian (LQG) regulators for linear systems. Also given are the methods for calculating the noise performance of LQG regulators based on apriori estimates of the state and measurement noise in a control system. The unity of LQR regulators and Kalman filters is stressed (in an appendix), making the transition from the deterministic to the stochastic case simple and natural. In addition to simple problems, this chapter includes a set of longer exercises which illustrate some real industrical control/estimation problems.

## Appendices

Because some of the material presented in the book might require a larger mathematical background than some students may have at a level the book targets, appendices are provided which include some of the most important detail necessary to understand the text. These appendices, of which there are 4, deal with optimization basics, linear algebra, derivation of the Riccati equation and discrete time systems. In particular the last appendix should prove useful as many engineering curricula no longer include the complex function theory necessary to understand such systems. It is provided with extensive illustrations and examples.

## Class Room Experiences with the Book

At the Technical University of Denmark (DTU) this book has been tested in the second control course for 8 years and revised, extended and improved over this period. During this time the students have willowed out a number of errors and unclear formulations which existed in the original text. The authors are grateful for their patience and understanding during this process. They are also grateful for their real additions to the text through their questions and requests for more information.

At DTU the second course in controls for which this text was written is a one semester course. During this time the entire book is covered in lectures and group work. The lectures are given twice a week and are typically 2 × 45 minutes followed by a "laboratory exercise" of two hours. The laboratory exercise consists of longer problems formulated as Matlab/Simulink problems requiring problem formuation (model building from system descriptions), model analysis, control system design and control system testing. It must be admitted that this is perhaps too much of a work load but apparently the course and the problems are sufficiently interesting that the students always continue to work on the laboratory exercies long enough to get the content out of them. This has been revealed over a period of time by questions directed to the instructors between lecture days.

It is well known among the students that the course is not easy and is time consuming to attend. Nevertheless the course is often over subscribed and students appear after the course start to see if they can be allowed to attend the lectures and go to exam. Final examination results show that in spite of the difficulties involved, most of the students have obtained a good feel for the material. Thesis work after course attendence (done some years later) shows the material in the textbook has found application in the solution of problems on a somewhat higher level than the course itself.

Lyngby, Denmark *Elbert Hendricks*
Lyngby, Denmark *Ole Jannerup*
USA *Paul Haase Sørensen*

# Contents

# List of Examples

# Chapter 1
# Introduction

**Abstract** This book is primarily intended to be an introduction to modern linear control theory but it is useful to review the classical control background for this important area. This chapter gives a brief history of primitive and classical controls followed by a review of some of the most important points in the earlier history of modern control theory.

## 1.1 The Invisible Thread

Technological and scientific advances occur in steps with the creation of various devices. In very early history these devices were simple and included such advances as water clocks in 270 B. C., mechanical clocks in 1283 A. D., steam boilers in 1681 A. D., steam engines early in the 1700s, and flushing toilets in 1775 A. D. In more recent times the devices invented were somewhat more complex and included internal combustion engines in about 1886, automatic pilots in 1914, electronic amplifiers in 1912 and accurate bomb sights in 1934. What is not usually recognized about these advances is that they are all strongly dependent on the use of feedback and/or automatic control.

The water clock is older than 270 B. C. but was improved significantly in accuracy by the Greek Ktesibios who invented a float regulator for the clock about this time. Mechanical clocks using the verge and foliot escapement were much more accurate than water clocks and were first seen in Europe about 1283. Steam boilers were used for many purposes in the sixteen hundreds and became practical with the invention of a steam pressure safety valve in 1681 by Dennis Papin. Steam engines became the prime mover for the Industrial Revolution with the invention of the centrifugal governor speed control in 1788 by James Watt. Flush toilets were refined by Thomas Crapper using float regulators and he received a Knighthood from Queen Victoria for his troubles. Thus even in early times feedback control and regulators have played a critically important but often nearly invisible part in making even simple technical devices work in practice.

Closer to the present, diesel internal combustion engines became practically possible with the creation of a fuel pressure pump/regulator to inject fuel into

E. Hendricks et al., *Linear Systems Control*, DOI: 10.1007/978-3-540-78486-9_1,

© Springer-Verlag Berlin Heidelberg 2008

them by Otto Diesel in 1889. An automatic pilot or stability augmentation device based on the use of the gyroscope was first demonstrated at an air show in Paris in 1914 by Lawrence Sperry. Electronic amplifiers constructed by Edwin Armstrong in 1912, used positive feedback in radio receivers to increase their sensitivity, making possible practical radio reception. Later negative feedback amplifiers were made by Harold Black in 1927 which made it possible to reduce distortion in audio frequency amplifiers for voice telephone and later music reproduction. Accurate bombsights were made possible by the use of synchro-repeaters to send air speed, wind and altitude data to a simple computer used in the Norden bombsight. These were used extensively by America during the bombing attacks against Germany during the Second World War. Again feedback and automatic control played an obscure or invisible (but ever increasing) part in the operation of these very successful inventions.

Through history then, feedback control has been woven into many of the most important technological innovations which have been created from other technological bases. In many ways these innovations have overshadowed automatic control but feedback control has facilitated many developments. According to Berstein (2002) feedback control has made possible most of the great waves of technological and scientific development including "... the Scientific Revolution, the Industrial Revolution, the Age of Aviation, the Space Age and the Age of Electronics". Thus feedback control is "an invisible thread in the history of technology". Automatic control enables modern technology in the same way as do modern computers and software but is in general much less visible. The description of feedback control as an "invisible thread" is due to Berstein (2002).

## 1.2 Classical Control Systems and their Background

Classical control is rooted in the primitive period of control system design. This period is characterized by many intuitive technical inventions and by the lack of an underlying theoretical basis (except for a few notable exceptions). Often clever mechanical devices were used to implement the actual control strategies. This period extends from antiquity up to the beginning of the Second World War, about 1940. At this stage the requirements of the war together with the availability of electronic devices and of theoretical tools adapted from the communications industry made possible a transition to the classical period of control system development.

### *1.2.1 Primitive Period Developments*

This book deals with the theory of control systems so it is appropriate to start the review of the primitive control period at a time when theoretical tools came into use to analyze control systems. This ignores many of the earliest developments in control systems but is reasonable given the scope intended. For interesting reviews of the prehistory of controls the reader should consult Fuller (1976a,b), Bellman and Kalaba (1964), Åstrom (1968) and Berstein (2002).

General recognition of the value of control systems came during the Industrial Revolution when James Watt began to use fly-ball speed governors for his steam engines in about 1788. This invention was not new in itself: windmills and watermills had used similar devices earlier but the application to widely available practical engines brought the feedback control problem into the open.

Unfortunately in some cases use of the governor lead to speed instability and to solve this problem it was necessary to create a theory to explain why this could occur. The first analysis of speed governors was carried out by George Airy (Astronomer Royal at Greenwich from 1835). Clockwork mechanisms were used to control the movement of large telescopes to compensate for the rotation of the earth and these were fitted with speed governors in an attempt to improve their accuracy. Using energy and angular momentum considerations, Airy set up a simplified nonlinear differential equation for the system in about 1840. Considering the linearized form of this equation, it was possible to show that in some circumstances small oscillations could build up exponentially and thus account for the instability observed.

A direct attack on the problem of engine speed governors was made by James Clerk Maxwell in a paper from 1868, "On governors". Maxwell set up the differential equations for the overall system and linearized them. In this work it was pointed out that for stability the characteristic equation of the linearized system must have roots with negative real parts. Airy's and Maxwell's papers influenced indirectly the work of Edward Routh who published his well known stability criteria in an Adams Prize Essay in 1877 at Cambridge University.

Independently of Airy and Maxwell, I. Vyshnegradskii carried out an analysis of governors in Russia in 1876. In continental Europe there were theoretical developments inspired by Vyshnegradskii's paper (which was also published in French). Aurel Stodola, who was working on the control of hydroelectric turbines in Switzerland, noted Vyshnegradskii's paper and asked Adolf Hurwitz to consider the problem of finding a stability criteria for systems of any order. Hurwitz was not aware of the work of Routh and found his own form of the simple stability criteria for linear systems. This took the form of a set of determinant inequalities in Hurwitz's formulation.

In 1892 Alexandr Lyapunov considered the stability of a very general class of nonlinear systems based on earlier work by Jean-Louis Lagrange. On the basis of general energy considerations he was able to give a general stability criteria for nonlinear systems as well as the conditions under which the method of linearization yields a valid assessment of the stability of the underlying nonlinear system. Lyapunov's paper was not well known in the West until about 1960 when it was re-discovered in the literature. Lyapunov's analysis and methods are now in wide general use in the design of control systems. At approximately the same time (1892–1898) operational calculus (Laplace transform analysis) was invented by Oliver Heavyside in England to analyse the transient behavior of linear electrical systems. Both Lyapunov analysis and the operational calculus played an important part in the theoretical development of control theory in the next century.

### *1.2.2 Pre-Classical Period Developments*

Experimenting with lighting lamps in 1880, Thomas Edison discovered the Edison effect. This effect is that a current could flow through a vacuum to a metal conductor in the lamp envelope. No explanation for this effect could be given until the identification of the electron in 1897 by J. J. Thompson in England. The discovery of the Edison effect was followed in 1904 by the invention of the thermionic rectifier diode in 1904 by John Flemming in England and eventually by the invention of the thermionic triode amplifier by Lee de Forest in 1906 (though he did not understand how it worked). This was the beginning of the Age of Electronics. Very rapidly hereafter feedback was applied around the triode amplifier resulting first in regenerative or positive feedback radio frequency amplifiers in 1912 and finally in negative feedback audio amplifiers in 1927. These new inventions were the work of Edwin Armstrong and Harold Black respectively.

On the theoretical side the complex algebra for A. C. circuits and frequency analysis techniques were being developed by Charles Steinmetz at General Electric in the United States in around 1906. The stable oscillations produced by a triode amplifier when positive feedback was used were found to be due to its nonlinearity. A theoretical study of the circuit was made by Balthasar van der Pol in 1920 which is a classic paradigm of nonlinear time domain system analysis.

In the early 1930s Harold Hazen made important contributions to the solution of the problem of constructing simulators for control problems including the solution of differential equations. Hazen made use of the experience obtained to write two important papers which were published in 1934 which studied the effects of feedback on electro-mechanical devices. For these papers he coined the expression "servomechanisms". This is the first use of this now common word in the literature. Later the understanding of servomechanisms which Hazen had obtained was used in the design of fast fire control systems for ships, Bennett (1993).

The general use of amplifier circuits in radio frequency receivers and telephone communication lead to a need to understand more deeply the theoretical nature of amplifier circuits and the reasons for stable and unstable behavior. This resulted in the work of Harry Nyquist on amplifier stability based on transfer function analysis in 1932 and that of Hendrik Bode on magnitude and phase frequency plots of the open and closed loop transfer functions of a system in 1940. The work of Nyquist resulted in the derivation of the Nyquist stability criterion and that of Bode in the investigation of closed loop stability using the concepts of gain and phase margin. In contrast to earlier work these stability criteria gave an idea of the relative level of stability which could be obtained. All of the available theory thus came from the electronics and communication industries. Thus the stage was set for a period of intense control system development at the beginning of the Second World War.

The Second World War made it necessary to develop electronic controllers for many different devices including radar controlled gun laying, bomb aiming and automatic pilot systems. By this time the size and power consumption of triode and pentode amplifiers had been significantly reduced, while at the same time their amplification, frequency range, robustness and reliability had been increased. Thus these devices were available for application to the required control systems and their performance in practical applications was well understood. In fact automated devices of many types, including those mentioned above, were developed during the war. A great deal of the development of radar and general electronics was carried out at the Massachusetts Institute of Technology in the Radiation Laboratory. This work was done using frequency domain techniques and was of course classified during the war but became generally known immediately afterward, likewise equivalent work done in the aircraft industry. Many of the presently used control system design tools emerged from these efforts including the general use of transfer functions, block diagrams and frequency domain methods.

In 1947 Nathaniel Nichols published his Nichols chart for the design of feedback systems based on work done in the Radiation Laboratory. Walter Evans published in 1948 his root locus design technique which is suitable to handle the large number of different states which describe the motion of an aircraft. An important problem in all electronic systems and in particular radar is noise. Albert Hall and Norbert Wiener realized the importance of this problem and developed frequency domain methods to deal with it. Hall's contribution was a frequency domain formulation of the effect of noise on control systems in 1941. Wiener introduced in 1942 the use of stochastic models to treat the effect of noise in dynamic systems and created an optimal filter to improve the signal to noise ratio in communication systems, the Wiener filter. Hall and Wiener published their work immediately after the war in 1946 and 1949 respectively. More or less the entire record of the work done at the Radiation Laboratory was published in a 27 volume series edited by Louis Fidenour in 1946. There was in this collection of material a volume which dealt specifically with the design of feedback control systems called "The Theory of Servomechanisms".

Thus began the period of Classical Control Theory: just after the end of the Second World War. The main theoretical tools included the frequency domain methods of Laplace transforms, transfer functions and s-plane analysis. Textbooks treating these subjects became rapidly and widely available as well as standard design tools, mostly intented to be used for hand calculations and plots.

### 1.2.3 Classical Control Period

The Classical Control period was characterized by a concentration on single loop, Single Input, Single Output feedback (SISO) systems designed with the theoretical tools developed during and just after the Second World War. For the most part these could be applied only to linear time invariant systems. The main underlying concept is that closed loop characteristics of a system can be determined uniquely given the open loop properties of the system. This includes the important disturbance rejection and steady state error properties of the feedback system.

The theoretical tools were those developed by Nyquist, Bode, Evans and Nichols earlier and the connection between these methods was clarified and extended. Performance was assessed in terms of bandwidth, gain and phase margin or rise time, percentage overshoot, steady state error, resonances and damping. The connection between these performance goals was well understood. More refined methods of tuning and compensating (using simple lead/lag compensators) single loop systems were developed. However the single loops involved had to be closed one at a time using a trial and error process which could not guarantee good performance or even stability.

During the Classical Period feedback controllers became part of many industrial processes because the methods of applying them were well known and because of the commercial availability of electronic, pneumatic and hydraulic controllers. This made feedback controllers a part of even home appliances as well as more advanced technical systems. When transistors became available in the 1960s Classical controls became even more wide spread.

In about 1956 a great interest developed in the control of aeronautical and astronautical systems and in particular aircraft, missiles and space vehicles: the Age of Space was beginning. Such systems have naturally a Multiple Input, Multiple Output (MIMO) quality which confounds the use of Classical Control Theory. While single loop analysis can be performed on such systems, the interweaving of the responses in the different parts of the control object make it impossible to use SISO design methods. In addition uncertainties and modelling errors together with stochastic system disturbances and measurement noise increase the order of the system which must be controlled and this augments the design difficulties.

The problem in controlling ballistic (free flying) objects can be formulated physically in terms of a coupled set of physically derived nonlinear differential equations. Sensor dynamics and noise can easily be included in the description of the control object itself. Thus one is lead naturally back to the time domain problem formulation which was used by Airy and Maxwell in the 1800s. As an important help in doing this one has available the classic variational formulations of analytic mechanics as given by Lagrange and Hamilton on the basis of physical conservation principles. This differential equation approach to the problem formulation is called the state variable or state space approach in control technology and it the main subject of this book. While it is possible in some cases to handle the nonlinearities involved directly, this book will only deal with the linearization of the underlying system around some desired or nominal operating point. Use of the state space approach makes it possible to treat MIMO systems in a natural way and to use the compact and convenient vector and matrix description of the control object and its associated feedback system.

Thus the Classical Control Period lasted between approximately 1945 to 1956 and gave rise to the Modern Control Period from 1956 to the present. The transition from the earlier period was forced by a requirement that more complex control objects be analyzed and control systems be made for them. Handling the details of the more complex systems has been enabled by the availability of inexpensive digital computers for design purposes but also for the direct implementation of the control systems themselves.

### *1.2.4 Modern Control Theory*

The large number of states in MIMO state variable systems and the possibly large number of feedback loops which might exist in a closed loop system make it necessary to consider how decisions might be made about the feedback levels in the different loops. It has turned out that this is not a simple question and it is difficult to impossible in fact to make any reasonable, balanced statement about what might be required. To solve this problem it has been found to be appropriate to attempt to design feedback systems which are optimal in the sense of balancing input power in the different system inputs against the errors (powers) which can accepted in the system outputs and/or states. This must be done of course in such a way that the overall system is stable as well as the separate loops in it.

This would be a daunting problem if it were not for work done much earlier to solve physical optimization problems in the sixteen, seventeen and eighteen hundreds by Leonard Euler, the Bernoullis (Jacob, I, II, Johannes, I, II, III, IV, Daniel and Nicolaus, I, II, II), Joseph-Louis Lagrange, William Hamilton and others. These workers gave general solutions to a number of minimization and maximization problems which have been adapted to solve past and current MIMO control and estimation problems using the Calculus of Variations. The initial work in adapting these mathematical methods to control problems occurred in about 1957 and continues to the present.

In Modern Control Theory it is common to minimize a performance index which may be a generalized quadratic energy function, in many cases with some secondary constraints (or limitations on the range or character of the solution). This may be seen as a “natural” requirement as most systems in nature operate in such a way as to minimize energy consumption. This may be done either for continuous or discrete time systems considered in the time domain. There are however frequency domain formulations of the results found when desired.

In 1957 Richard Bellman applied dynamic programming to the optimal control of discrete time systems. This work demonstrated that it was possible to solve control problems based on a performance index resulting in closed loop feedback controls which could in general be nonlinear. The solution was also immediately adaptable to computer based discrete time control systems. Lev Pontryagin suggested a maximum principle in 1958 which solved the minimum time control problem for a class of control objects using relay control.

In 1960 and 1961 a significant set of breakthroughs were made generally available with the publication of four papers by Rudolf Kalman and co-workers. These papers dealt with (1) the optimal control of discrete time systems (with J. Bertram), (2) the design equations for the Linear Quadratic Regulator (LQR), (3) optimal filtering and estimation theory for discrete time systems and (4) the continuous time Kalman filter (with Richard Bucy). The first two control papers are based on minimizing a very general energy control performance index. The last two deal with applying the same techniques to the optimum filtering (or estimation) problem which is a generalization of the least squares fitting techniques due to Carl Fredrik Gauss. All of these solutions were immediately compatible with computer control and estimation at a time when computers first became generally available for on-line applications. Thus the four theoretical breakthroughs above found immediate practical application in the very visible aero-space programs of the time. LQR regulators and Kalman filters form a central part of the material which is presented in this book.

Computer control has become extremely important in control applications so that it should be mentioned that discrete time control emerged at the same time as modern control theory. This was due to the publications of a number of workers, among whom are: John Ragazzini, Gene Franklin, Lotfi Zedah, Eliahu Jury, Benjamin Kuo, Karl Åstrøm and Bjørn Wittenmark, in the period of 1952 to 1984. The sampling theory, on which the success of modern computer control is based, is due in part to the papers of Harry Nyquist from 1928, Claude Shannon, who worked with sampling theory at Bell Labs in 1949, and Vladimir Kotelnikov in Russia in 1933. Together with the state variable formulation of system dynamics this work on discrete time control forms the backbone of current technical and industrial control applications in many fields.

While automatic control systems and feedback have been an “invisible thread” during most of time they have existed they will become less so in the future. This is because the value of accurate control has come to be generally recognized in many applications. For example the newer fighter and bomber aircraft in the United States and Europe have been created to be unstable from the outset in order to give them high maneuverability or stealth characteristics and cannot be flown without their fly-by-wire control systems. Modern automotive engines have electronic throttles and fuel controllers and cannot operate without their digital Engine Control Units (ECUs). Such advanced applications will become much more numerous in the future as the price of computer control systems continues to decrease.

# Chapter 2
# State Space Modelling of Physical Systems

**Abstract** Modelling of state space models based on relevant physical laws is introduced. Linearization of nonlinear models is discussed and the connection between the transfer function model and the state space model is derived. Discrete time models are also introduced.

## 2.1 Modelling of Physical Systems

Design of control systems is in most cases based on a model of the system to be controlled. In ‘classical’ control engineering the model is usually presented as a transfer function. A transfer function can be formulated for a restricted class of systems: linear, time invariant systems. Such systems are very rare in the real world – if they exist at all – and the linear model is therefore almost always an approximate description of the real system. In spite of this, the transfer function is a very useful tool for analysis and design of control systems, and it is widely used by control engineers for a large variety of control problems.

It is quite obvious though that transfer functions have their limitations. While they are well suited for systems with one input and one output (also called Single-Input-Single-Output or *SISO-systems*), they can be awkward to use for systems with more than one input and one output (Multiple-Input-Multiple-Output or*MIMO-systems*).

In this textbook an alternative type of system model will primarily be used: the state space model. Such a model is, just like transfer function models, based on a suitable set of physical laws that are supposed to govern the system under consideration. In the state space modelling the model equation is not transformed into the frequency domain as is done in setting up transfer functions: one stays in the time domain. On the other hand, the set of governing differential equations (which may be of higher order) are translated to a set of *first order* coupled differential equations.

E. Hendricks et al., *Linear Systems Control*, DOI: 10.1007/978-3-540-78486-9_2,

© Springer-Verlag Berlin Heidelberg 2008

## 2.2 Linear System Models

For systems which are linear or for which one can formulate a linear model which is sufficiently accurate for control purposes, the modelling procedure is usually quite simple. The basic principles are illustrated in the examples below.

***Example 2.1.*** **RLC Circuit**

On Fig. 2.1 a simple passive electrical RLC-circuit is shown.

It is desired to formulate a model where the terminal voltage $v$ is the input and the voltage across the capacitor $v_C$ is the output. Ohm's and Kirchhoff's laws give the following relation for the voltages,

$$Ri(t) + L\frac{di(t)}{dt} + v_C(t) = v(t). \tag{2.1}$$

For the capacitor it is known that

$$Cv_c(t) = q(t) \Rightarrow C\frac{dv_C(t)}{dt} = \frac{dq(t)}{dt} = i(t). \tag{2.2}$$

From these equations it can be seen immediately that

$$\begin{aligned} \dot{i} &= -\frac{R}{L}i - \frac{1}{L}v_C + \frac{1}{L}v, \\ \dot{v}_c &= \frac{1}{C}i. \end{aligned} \tag{2.3}$$

Here the time-argument has been omitted.

The system can obviously be described by these two coupled first order ordinary differential equations. It is usually preferred to reformulate the equations into one *vector-matrix* equation,

$$\begin{bmatrix} \dot{i} \\ \dot{v}_C \end{bmatrix} = \begin{bmatrix} -\frac{R}{L} & -\frac{1}{L} \\ \frac{1}{C} & 0 \end{bmatrix} \begin{bmatrix} i \\ v_C \end{bmatrix} + \begin{bmatrix} \frac{1}{L} \\ 0 \end{bmatrix} v. \tag{2.4}$$

If the state variables are defined as

$$\mathbf{x} = \begin{bmatrix} i \\ v_C \end{bmatrix}, \ \dot{\mathbf{x}} = \begin{bmatrix} \dot{i} \\ \dot{v}_C \end{bmatrix}, \ \mathbf{A} = \begin{bmatrix} -\frac{R}{L} & -\frac{1}{L} \\ \frac{1}{C} & 0 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} \frac{1}{L} \\ 0 \end{bmatrix}, \ \ u = v,$$

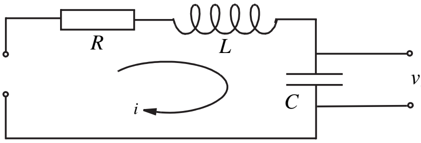


**Fig. 2.1** Dynamic electrical RLC-circuit

Equation (2.4) can be rewritten in a very compact way as

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u. \tag{2.5}$$

A block diagram of this system is shown on Fig. 2.2.

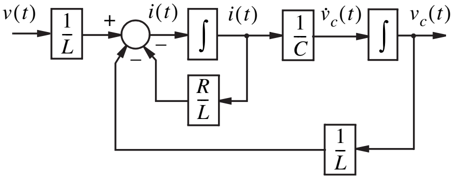


**Fig. 2.2** Block diagram of the RLC-circuit

Note that the entries $i(t)$ and $v_c(t)$ of the vector $\mathbf{x}$ are outputs from the two integrators in the block diagram. Equations (2.4) or (2.5) are called the state equations of the system on Fig. 2.1. Note that the differentiation of a vector with respect to the scalar argument (in this case the time $t$) is accomplished by simply differentiating each element of the vector.

Since the capacitor voltage is the output, a second equation should be added to the description of the problem, the output equation:

$$y = v_C = [0 \quad 1]\mathbf{x}. \tag{2.6}$$

The combination of Eqs. (2.5) and (2.6) is called the state space model of the system.

Laplace transformation of the Eq. (2.3) would lead to the transfer function model of the system,

$$G(s) = \frac{V_c(s)}{V(s)} = \frac{1}{LCs^2 + RCs + 1}. \tag{2.7}$$

❐

In Example 2.1 the two-dimensional state vector selected is $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} i \\ v_C \end{bmatrix}$. The reason for choosing precisely two variables in the vector is that one knows from electrical circuit analysis that such an RLC-circuit is a second order system. Even if one has no a priori knowledge of the system, it is usually not difficult to choose the correct number of elements in the $\mathbf{x}$-vector. This will be demonstrated later.

The model in the example is what is called a *linear* state space model since the right hand side of Eq. (2.4) is a linear function of $\mathbf{x}$ and $u$. The matrices $\mathbf{A}$ and $\mathbf{B}$ have constant elements and therefore the model is said to be *time invariant*.

The system above (or rather the model of the system) belongs to the class of models for which it is also possible to formulate a transfer function model. But the state space formulation is not limited to the description of such models. In fact, the general state space model can be nonlinear and can be expressed as:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t),\\ \mathbf{y}(t) &= \mathbf{g}(\mathbf{x}(t), \mathbf{u}(t), t).\end{aligned} \tag{2.8}$$

When state space models are formulated the following standard notation is used:

$\mathbf{x}(t)$: state vector of dimension $n$, $\mathbf{x}(t) \in \Re^n$,  
$\dot{\mathbf{x}}(t)$: time derivative of state vector, $\dot{\mathbf{x}}(t) \in \Re^n$,  
$\mathbf{u}(t)$: input vector, $\mathbf{u}(t) \in \Re^m$,  
$\mathbf{y}(t)$: output vector,. $\mathbf{y}(t) \in \Re^r$.

The functions **f** and **g** are vector functions:

$$\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) = \begin{bmatrix} f_1(\mathbf{x}(t), \mathbf{u}(t), t) \\ f_2(\mathbf{x}(t), \mathbf{u}(t), t) \\ \vdots \\ f_n(\mathbf{x}(t), \mathbf{u}(t), t) \end{bmatrix} \quad \text{and} \quad \mathbf{g}(\mathbf{x}(t), \mathbf{u}(t), t) \begin{bmatrix} g_1(\mathbf{x}(t), \mathbf{u}(t), t) \\ g_2(\mathbf{x}(t), \mathbf{u}(t), t) \\ \vdots \\ g_r(\mathbf{x}(t), \mathbf{u}(t), t) \end{bmatrix},$$

where the functions $f_i$ and $g_j$ are scalar functions of the vectors **x**(t) and **u**(t).

In the general linear case the system equations can be written:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t),\\ \mathbf{y}(t) &= \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t).\end{aligned} \tag{2.9}$$

Further, if all matrix elements are constant, these equations are reduced to:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t),\\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t),\end{aligned} \tag{2.10}$$

and the model is said to be Linear and Time Invariant (LTI).

The following names are generally used for the matrices in Eq. (2.10):

**A**: system or dynamic matrix, $\mathbf{A} \in \Re^{n \times n}$,  
**B**: input matrix, $\mathbf{B} \in \Re^{n \times m}$,  
**C**: output matrix, $\mathbf{C} \in \Re^{r \times n}$,  
**D**: direct transfer or feed forward matrix, $\mathbf{D} \in \Re^{r \times m}$.

The Eq. (2.10) can be drawn as the block diagram on Fig. 2.3. The double lines indicate that vector quantities (multiple variables) are passed between the blocks.

Sometimes it is convenient (or necessary) to divide the inputs of the system into two groups: The group of quantities which one can manipulate and the variables which have to be regarded as disturbances. The latter type of variables are determined by the world surrounding the system and they assume values beyond the designer's control. In such cases the state equation can be written,

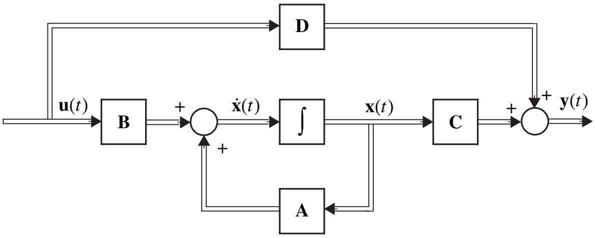


**Fig. 2.3** Block diagram of a general continuous linear state space model

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{v}(t), t), \tag{2.11}$$

or, in the LTI-case,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{B}_v\mathbf{v}(t), \tag{2.12}$$

where **v**(t) is a vector of *disturbance* variables,

$$\mathbf{v}(t) \in \Re^p, \mathbf{B}_v \in \Re^{n \times p}.$$

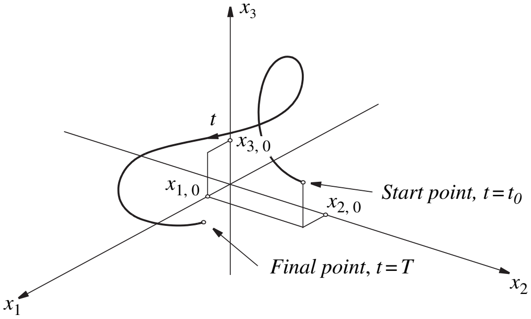
Although a model of a linear time invariant system can be a transfer function as well as a state space model, it is important to realize that these two models do not give the same insight into the system's properties. The transfer function is a very efficient description of the relationship between the input and the output. The state space model provides the same information, of course, but in addition, it also gives detailed information on the internal state variables. This means that one can keep track of what is going on internally in the system. For this reason the state space model is called an *internal model* whereas the transfer function is an *external model.*

There are other important differences between the two model types. When a transfer function is constructed one can be sure that it is a *unique* model. In contrast to this, the state space model is not unique. One has considerable freedom when selecting the state variables and the resulting model obviously depends on the specific choice of state variables. As will be seen later, this fact allows one to construct state space models with certain special and useful properties.

The state vector's components are functions of time. Over a period of time limited by the initial time and the final time *T* the state vector describes the system's behavior in an *n*-dimensional vector space. The path of the vector end point is called the *trajectory*. An example is shown on Fig. 2.4 for some third order system.

It is not easy to give a stringent definition of the term *state*. Its importance may be illustrated by considering the information it carries. If the state is known at a particular time, it will be possible from the model to calculate all other variables in the system at that time. Also, the model makes it possible to

**Fig. 2.4** System trajectory in a 3-dimensional vector space



calculate the state at any point in time, say $t$, if the initial state $\mathbf{x}(t_0)$ and the input in the time interval $[t_0, t]$ are known. This will be true for any proper choice of state variables.

How the state variables should be selected is also difficult to express in general terms. However, for particular system models, for instance presented in the form of block diagrams, it is quite straight forward to point out one proper choice of state variables. The procedure is illustrated in Example 2.2 below.

***Example 2.2.*** **DC Motor and Flexible Coupling**

Figure 2.5 shows an electric DC-motor with a flexible coupling in the shaft between the motor armature inertia and the load inertia. $R$ and $L$ are the resistance and the inductance of the armature windings, $k$ and $b$ are the spring constant and a viscous (linear) internal damping coefficient of the flexible coupling respectively and $J_m$ and $J_l$ are the moments of inertia of the motor armature and load respectively. The armature terminal voltage $u$ is the input to the system and the load angular position $\theta_l$ is the output.

If the two inertias are separated from each other and the appropriate torques $(T_k, T_b, T_a)$ added as shown on Fig. 2.6, one can write Newton's Second Law for both of them as

$$J_m\ddot{\theta}_m = K_a i + k(\theta_l - \theta_m) + b(\dot{\theta}_l - \dot{\theta}_m) - b_m\dot{\theta}_m, \tag{2.13}$$

$$J_l\ddot{\theta}_l = -k(\theta_l - \theta_m) - b(\dot{\theta}_l - \dot{\theta}_m) - b_l\dot{\theta}_l. \tag{2.14}$$

$K_a$ is the torque constant, $b_m$ and $b_l$ are viscous bearing friction factors of motor and load shaft respectively.

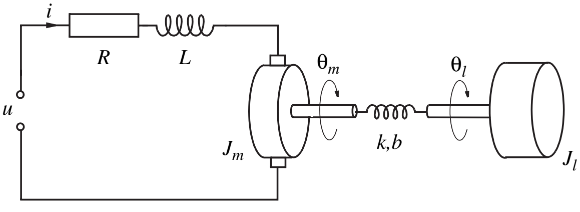


**Fig. 2.5** DC-motor with flexible coupling

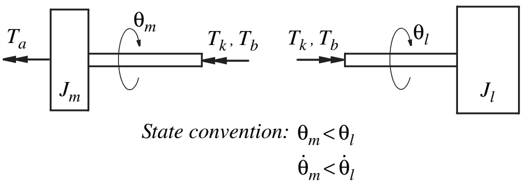

$T_a$
$J_m$
$\theta_m$
$T_k, T_b$
$T_k, T_b$
$\theta_l$
$J_l$
State convention: $\theta_m < \theta_l$
$\dot{\theta}_m < \dot{\theta}_l$

**Fig. 2.6** Sign and state convention for the system on Fig. 2.5

Ohm's and Kirchhoff's laws applied to the electrical circuit yield

$$u = Ri + L\frac{di}{dt} + k_e\dot{\theta}_m \tag{2.15}$$

or

$$\dot{i} = \frac{1}{L}(u - k_e\dot{\theta}_m - Ri), \tag{2.16}$$

where $k_e$ is the induction coefficient of the motor armature windings.

A block diagram drawn directly from Eq. (2.13), (2.14) and (2.16) is shown on Fig. 2.7. Since the state equations are first order differential equations, it is obvious that the integrator *outputs* would be natural candidates for state variables. With this choice, the first order time derivatives are simply the *inputs* to the integrators. If the 5th-order state vector is defined as

$$\mathbf{x} = [x_1\ x_2\ x_3\ x_4\ x_5]^T = \left[i\ \theta_m\ \dot{\theta}_m\ \theta_l\ \dot{\theta}_l\right]^T, \tag{2.17}$$

the time derivative of **x** will be

$$\dot{\mathbf{x}} = \left[\dot{i}\ \dot{\theta}_m\ \ddot{\theta}_m\ \dot{\theta}_l\ \ddot{\theta}_l\right]^T, \tag{2.18}$$

and the state equations can be written by inspection of the block diagram:

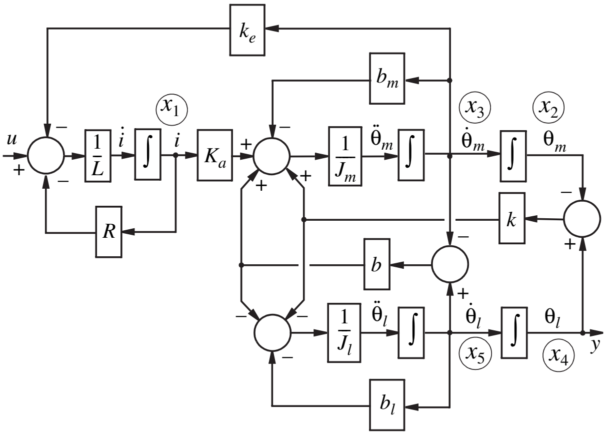


**Fig. 2.7** Block diagram for the system on Fig. 2.5

$$\begin{aligned}
\dot{x}_1 &= \frac{R}{L}x_1 - \frac{k_e}{L}x_3 + \frac{1}{L}u,\\
\dot{x} &= x_3,\\
\dot{x}_3 &= \frac{K_a}{J_m}x_1 - \frac{k}{J_m}x_2 - \frac{b+b_m}{J_m}x_3 + \frac{k}{J_m}x_4 + \frac{b}{J_m}x_5,\\
\dot{x}_4 &= x_5,\\
\dot{x}_5 &= \frac{k}{J_l}k_2 + \frac{b}{J_l}x_3 - \frac{k}{J_l}x_4 - \frac{b+b_l}{J_l}x_5.
\end{aligned} \tag{2.19}$$

In vector-matrix form this equation is written

$$\dot{\mathbf{x}} = \begin{bmatrix} -\frac{R}{L} & 0 & -\frac{k_e}{L} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{K_a}{J_m} & -\frac{k}{J_m} & -\frac{b+b_m}{J_m} & \frac{k}{J_m} & \frac{b}{J_m} \\ 0 & 0 & 0 & 0 & 1 \\ 0 & \frac{k}{J_l} & \frac{b}{J_l} & -\frac{k}{J_l} & -\frac{b+b_l}{J_l} \end{bmatrix}\mathbf{x} + \begin{bmatrix} \frac{1}{L} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} u. \tag{2.20}$$

The output equation is then

$$y = [0 \quad 0 \quad 0 \quad 1 \quad 0]\mathbf{x}. \tag{2.21}$$

One sees that the systems on Fig. 2.5 is a 5th order system: it has 5 states. ❐

The state variables selected in Example 2.2 were the outputs from the integrators in the block diagram. This particular set of state variables is called the *natural state variables*. It should be noted, that the block diagram is not unique and therefore the set of natural state variables is not unique either. The numbering order of the state variables can also be changed and this will result in other matrices than those shown in the example. The matrix elements will certainly be the same but they will appear in different places in the matrices.

***Example 2.3*. DC Motor Block Diagram**

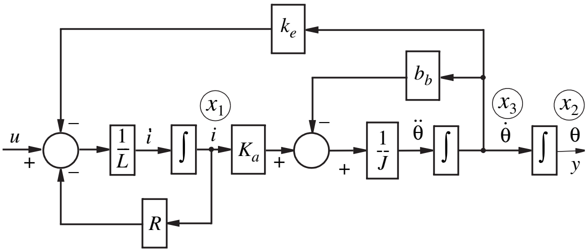
The electro-mechanical system in Example 2.2 had a flexible coupling between the two rotating inertias. If this flexibility is absent, which means that the coupling is completely stiff, the system equations will take on a quite different appearance.

A stiff coupling means that the spring coefficient is infinite: $k \cong \infty$. One cannot modify the elements of the matrices in Eq. (2.20) directly, since some of them would also be infinitely large and this is not possible. Instead one must revert to the set of Eqs. (2.13) and (2.14) and modify them. It is obvious that the two angle positions will now be equal: $\theta_m = \theta_l = \theta$ and the moment of inertia and the bearing friction factors will be the sum of the two separate ones: $J = J_m + J_l$ and $b_b = b_m + b_l$. The two equations are reduced to one second order differential equation (Newton's Second Law),

$$J\ddot{\theta} = K_a i - b_b\dot{\theta}. \tag{2.22}$$

The electrical equations will be the same as before and the block diagram of the new system is shown on Fig. 2.8.

**Fig. 2.8** Block diagram of the first reduced system



The number of states has been reduced to 3 and a proper choice is now

$$\mathbf{x} = [x_1\ x_2\ x_3]^T = \left[i\ \theta\ \dot{\theta}\right]^T \tag{2.23}$$

and the state and output equations are seen to be

$$\dot{\mathbf{x}} = \begin{bmatrix} -\frac{R}{L} & 0 & -\frac{k_e}{L} \\ 0 & 0 & 1 \\ \frac{K_a}{J} & 0 & -\frac{b_b}{J} \end{bmatrix}\mathbf{x} + \begin{bmatrix} \frac{1}{L} \\ 0 \\ 0 \end{bmatrix} u, \tag{2.24}$$

$$y = [0\ 1\ 0]\,\mathbf{x}. \tag{2.25}$$

Further reduction of the system can often be justified. For most small DC servo motors the armature inductance is very small and it is sensible to omit its influence on the model. Again, one cannot just set $L = 0$ in Eq. (2.24) but one must go back to the electric circuit Eq. (2.15) and make the change there,

$$u = Ri + k_e\,\dot{\theta}. \tag{2.26}$$

In the new block diagram (Fig. 2.9) the armature current $i$ is no longer a state and the only ones remaining are

$$\mathbf{x} = \left[\theta\ \dot{\theta}\right]^T, \tag{2.27}$$

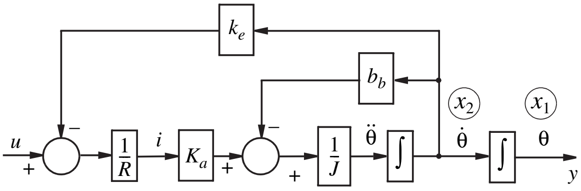


**Fig. 2.9** Block diagram of the system in Fig. 2.8 with L = 0

leading to the equations:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & -\dfrac{b_b R + K_a k_e}{JR} \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ \dfrac{K_a}{JR} \end{bmatrix} u, \tag{2.28}$$

$$y = [\,1\ 0\,]\mathbf{x}. \tag{2.29}$$

□

### *Example 2.4.* Double RC Low Pass Filter

Now it will be shown how to derive a state space model of the passive electrical network shown on Fig. 2.10. Using Ohm's and Kirchhoff's laws on the three loop currents yields

$$\begin{aligned} e_i &= R_1 i_1 + \frac{1}{C_1}\int (i_1 - i_2)dt, \\ 0 &= \frac{1}{C_1}\int (i_2 - i_1)dt + \frac{1}{C_2}\int i_2 dt + R_2 i_2, \\ -e_o &= -\frac{1}{C_2}\int i_2 dt. \end{aligned} \tag{2.30}$$

Rearranging these equations,

$$\begin{aligned} R_1 i_1 &= e_i - \frac{1}{C_1}\int (i_1 - i_2)dt, \\ R_2 i_2 &= \frac{1}{C_1}\int (i_1 - i_2)dt - \frac{1}{C_2}\int i_2 dt, \\ e_o &= \frac{1}{C_2}\int i_2 dt, \end{aligned} \tag{2.31}$$

leads to the block diagram on Fig. 2.11. Now it is simple to derive the state equations directly from the diagram:

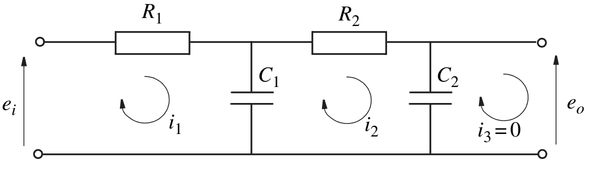
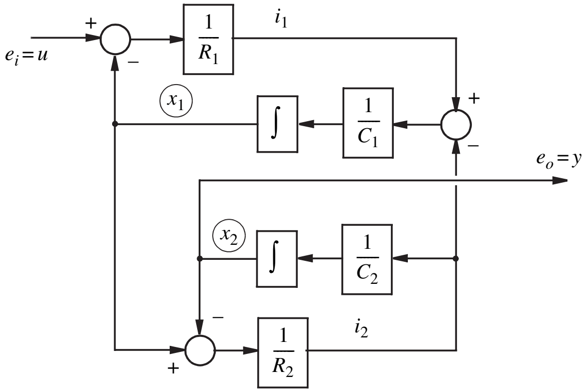


**Fig. 2.10** Simple electrical network

**Fig. 2.11** Block diagram of system on Fig. 2.10



$$\begin{aligned}\dot{x}_1 &= \frac{1}{C_1}\left(\frac{1}{R_1}(u - x_1) - \frac{1}{R_2}(x_1 - x_2)\right),\\ \dot{x}_2 &= \frac{1}{C_2}\left(\frac{1}{R_2}(x_1 - x_2)\right).\end{aligned} \tag{2.32}$$

Rearranging gives the state space model

$$\begin{aligned}\dot{\mathbf{x}} &= \begin{bmatrix} -\dfrac{R_1 + R_2}{C_1 R_1 R_2} & \dfrac{1}{C_1 R_2} \\ \dfrac{1}{C_2 R_2} & -\dfrac{1}{C_2 R_2}\end{bmatrix}\mathbf{x} + \begin{bmatrix}\dfrac{1}{C_1 R_1} \\ 0\end{bmatrix} u, y \\ y &= [0 \quad 1]\mathbf{x}.\end{aligned} \tag{2.33}$$

❐

### *Example 2.5.* Electrical Oven Process Plant

Figure 2.12 shows an electrically heated insulated oven which contains a product to be heated. The temperature of the air in the oven space is $T_s$, the

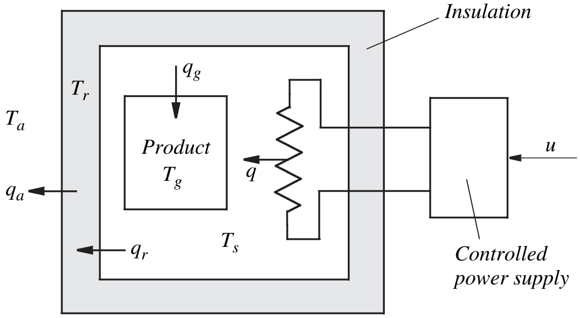


**Fig. 2.12** Electrically heated oven production system

temperatures of the product, the insulation material and the ambient air are $T_g$, $T_r$ and $T_a$ respectively. All temperatures are assumed to be uniform. The heating power from the electrical heating element is called $q$, and the powers entering the product and insulation are $q_g$ and $q_r$. The heat loss to the ambient air is $q_a$.

The controlled power supply is linear so that

$$q = ku, \tag{2.34}$$

where $k$ is a proportionality constant.

The exchange of heat energy is assumed to be by convection and therefore the powers and the temperatures are related as follows,

$$\begin{aligned} q_g &= k_g(T_s - T_g), \\ q_r &= k_r(T_s - T_r), \\ q_a &= k_a(T_r - T_a), \end{aligned} \tag{2.35}$$

where the $k$-coefficients are convection parameters depending on the area and the physical nature of the surfaces.

If the total heat capacity of the oven air space, the product and the insulation are denoted by $C_s$, $C_g$ and $C_r$, expressions for the time derivative of the temperature of the different parts of the system can be formulated. One finds

$$\begin{aligned} C_s \frac{dT_s}{dt} &= q - q_g - q_r, \\ C_g \frac{dT_g}{dt} &= q_g, \\ C_r \frac{dT_r}{dt} &= q_r - q_a. \end{aligned} \tag{2.36}$$

A block diagram of the model can be seen on Fig. 2.13.

It is quite obvious, directly from the differential Eq. (2.36), that a sensible choice of state variables could be

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} T_s \\ T_g \\ T_r \end{bmatrix}. \tag{2.37}$$

With input $u$, the disturbance $v = T_a$ and output $y = T_g$, it is now straightforward to obtain the set of state equations. It is only a matter of substituting the proper variables into the Eq. (2.36). The result is

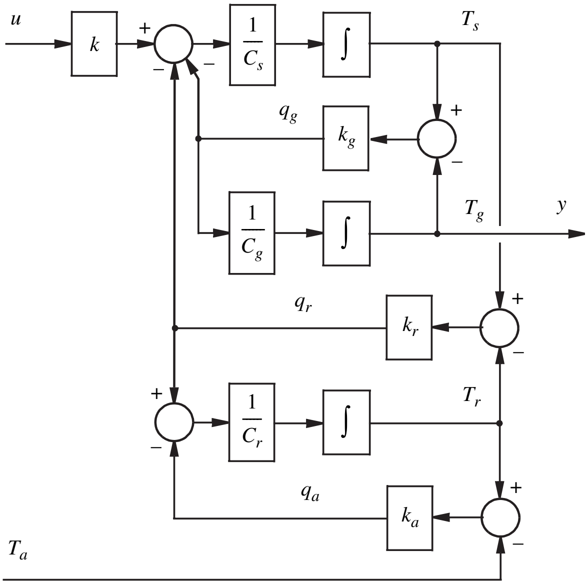


**Fig. 2.13** Block diagram of the oven production system

$$\begin{aligned}
\dot{x}_1 &= \frac{1}{C_s}(ku - k_g(x_1 - x_2) - k_r(x_1 - x_3)),\\
\dot{x}_2 &= \frac{1}{C_g}k_g(x_1 - x_2), \qquad (2.38)\\
\dot{x}_3 &= \frac{1}{C_r}(k_r(x_1 - x_3) - k_a(x_3 - v)),
\end{aligned}$$

or in vector-matrix form,

$$\dot{\mathbf{x}} = \begin{bmatrix} -\dfrac{k_g + k_r}{C_s} & \dfrac{k_g}{C_s} & \dfrac{k_r}{C_s} \\ \dfrac{k_g}{C_g} & -\dfrac{k_g}{C_g} & 0 \\ \dfrac{k_r}{C_r} & 0 & -\dfrac{k_r + k_a}{C_r} \end{bmatrix}\mathbf{x} + \begin{bmatrix} \dfrac{k}{C_s} \\ 0 \\ 0 \end{bmatrix}u + \begin{bmatrix} 0 \\ 0 \\ \dfrac{k_a}{C_r} \end{bmatrix}v, \qquad (2.39)$$

$$y = [0\ 1\ 0]\mathbf{x}.$$

❐

## 2.3 State Space Models from Transfer Functions

### *2.3.1 Companion Form 1*

For a SISO system with a known transfer function it is possible to formulate a state space model in a standard form.

Assume that the transfer function is

$$G(s) = \frac{Y(s)}{U(s)} = \frac{b_n s^n + b'_{n-1}s^{n-1} + \ldots + b'_1 s + b'_0}{s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0}. \tag{2.40}$$

Note, that this transfer function is *proper*, i.e., the numerator and the denominator polynomials have the same order.

The procedure is started by performing the first step of a polynomial division

$$\frac{Y(s)}{U(s)} = b_n + \frac{b_{n-1}s^{n-1} + \ldots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0} = b_n + \frac{B(s)}{A(s)}. \tag{2.41.}$$

Introducing an auxiliary variable,

$$V(s) = \frac{1}{A(s)}U(s) \Rightarrow A(s)V(s) = U(s), \tag{2.42}$$

leads to

$$s^n V(s) = -a_{n-1}s^{n-1}V(s) - a_{n-2}s^{n-2}V(s) - \ldots - a_1 sV(s) - a_0 V(s) + U(s) \tag{2.43}$$

and to

$$Y(s) = b_n U(s) + B(s)V(s). \tag{2.44}$$

Now define the state variables as follows

$$\begin{aligned} X_1(s) &= V(s), \\ X_2(s) &= sV(s), \\ X_3(s) &= s^2\,V(s), \\ &\vdots \\ X_{n-2}(s) &= s^{n-3}V(s), \\ X_{n-1}(s) &= s^{n-2}V(s), \\ X_n(s) &= s^{n-1}V(s). \end{aligned} \tag{2.45}$$

Multiplying each expression by $s$ and substituting the state variables for the right hand sides of the equations gives

$$
\begin{aligned}
sX_1(s) &= X_2(s),\\
sX_2(s) &= X_3(s),\\
sX_3(s) &= X_4V(s),\\
&\vdots\\
sX_{n-2}(s) &= X_{n-1}V(s),\\
sX_{n-1}(s) &= X_n(s),\\
sX_n(s) &= s^nV(s).
\end{aligned}
$$

Inverse Laplace transformation of these equations and of (2.43) gives

$$
\begin{aligned}
&\dot{x}_1(t) = x_2(t),\\
&\dot{x}_2(t) = x_3(t),\\
&\dot{x}_3(t) = x_4(t),\\
&\vdots\\
&\dot{x}_{n-2}(t) = x_{n-1}(t),\\
&\dot{x}_{n-1}(t) = x_n(t),\\
&\dot{x}_n(t) = -a_0x_1(t) - a_1x_2(t) - \ldots - a_{n-2}x_{n-1}(t) - a_{n-1}x_n(t) + u(t).
\end{aligned}
$$

This set of first order differential equations constitutes the state equations for the System (2.40). The vector-matrix form is seen to be

$$
\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix}
= \begin{bmatrix}
0 & 1 & 0 & \ldots & 0 & 0\\
0 & 0 & 1 & \ldots & 0 & 0\\
\vdots & \vdots & \vdots & \ldots & \vdots & \vdots\\
0 & 0 & 0 & \ldots & 0 & 1\\
-a_0 & -a_1 & -a_2 & \ldots & -a_{n-2} & -a_{n-1}
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}
+ \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u
\tag{2.46}
$$

$$= \mathbf{A}\mathbf{x} + \mathbf{B}u.$$

The output equation can be found by inverse Laplace transformation of (2.44),

$$y = [b_0\, b_1 \ldots b_{n-2}\, b_{n-1}]\mathbf{x} + b_nu = \mathbf{C}\mathbf{x} + \mathbf{D}u. \tag{2.47}$$

Note that this state space model is very efficient in the sense of the number of matrix elements. The matrix elements are the same as the coefficients of the transfer function model (2.41) and apart from the zeros and ones these are the only parameters in the model. The coefficients of the denominator of (2.40)/(2.41) can be found in the last row of **A** with the opposite signs and in reverse order. The numerator coefficients of (2.41) are found in **C** in reverse order.

If the numerator polynomial is of lower order than the denominator (in which case the system is said to be *strictly proper*), $b_n$ and maybe some of the elements of **C** will be zero. The matrix **D** will be zero in this case. A block diagram of the system based on the state equations is shown on Fig. 2.14.

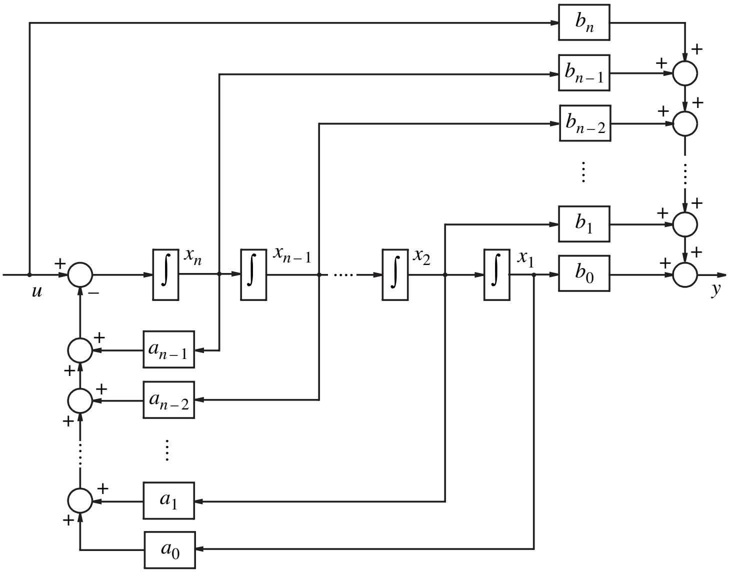


**Fig. 2.14** Block diagram of companion form 1 of a SISO system

The states selected here are often called the systems’*phase variables* and the model is said to be in *phase variable form*. A matrix with the structure of the system matrix in Eq. (2.46) is called a *companion matrix* and an alternative name for this model is the *companion form 1*.

### *Example 2.6.* DC Motor Position Control

Return now to the third order system on Fig. 2.8 of Example 2.3. If one Laplace transforms the operators in the block diagram, the transfer function for the system can be derived and the result is

$$\frac{\Theta(s)}{U(s)} = \frac{K_a}{s(LJs^2 + (RJ + Lb_b)s + Rb_b + K_aK_e)}.$$

With the component data

$$J = 0.2,\ L = 0.01,\ R = 3, b_b = 0.05,\ K_a = 0.35,\ K_e = 0.35,$$

the transfer function turns out to be

$$\frac{\Theta(s)}{U(s)} = \frac{0.5}{s(0.002s^2 + 0.6005s + 0.2725)} = \frac{175}{s^3 + 300.25s^2 + 136.25s}.$$

The matrices of the state space model can now be written down directly from Eq. (2.46) and (2.47),

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -136.25 & -300.25 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \mathbf{C} = [\,175 \quad 0 \quad 0\,], \mathbf{D} = 0.$$

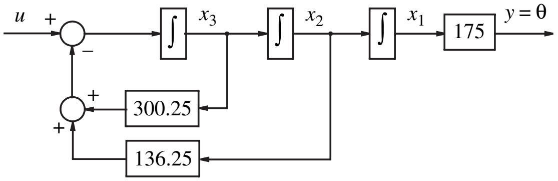


**Fig. 2.15** Block diagram of a third order system

A block diagram drawn along the same lines as Fig. 2.14 is shown on Fig. 2.15. If this the block diagram is compared with that on Fig. 2.8, it is noted that they are quite different although they are models of the same system. Figure 2.15 contains only 3 parameters, whereas Fig. 2.8 has 6. ❐

### 2.3.2 Companion Form 2

Returning to the transfer function (2.41), a state space model can be constructed by an alternative choice of state variables.

In this case an auxiliary variable is defined, but this time with the expression,

$$W(s) = \frac{B(s)}{S(s)}U(s), \tag{2.48}$$

so that

$$Y(s) = b_n U(s) + \frac{B(s)}{A(s)}U(s) = b_n U(s) + W(s) \tag{2.49}$$

and the following set of state variables is selected

$$
\begin{aligned}
&X_n(s) = W(s),\\
&X_{n-1}(s) = sW(s) + a_{n-1}W(s) - b_{n-1}U(s),\\
&X_{n-2}(s) = s^2\,W(s) + a_{n-1}sW(s) - b_{n-1}sU(s) + a_{n-2}W(s) - b_{n-2}U(s),\\
&\qquad\vdots\\
&X_1 = s^{n-1}W(s) + a_{n-1}s^{n-1}W(s) - b_{n-1}s^{n-2}U(s) + \ldots + a_1W(s) - b_1U(s).
\end{aligned}
$$

After inverse Laplace transformation these expressions can be written

$$
\begin{aligned}
&x_n(t) = w(t),\\
&x_{n-1}(t) = \dot{x}_n(t) + a_{n-1}x_n(t) - b_{n-1}u(t),\\
&x_{n-2}(t) = \dot{x}_{n-1}(t) + a_{n-2}x_n(t) - b_{n-2}u(t),\\
&\qquad\vdots\\
&x_1(t) = \dot{x}_2(t) + a_1x_n(t) - b_1u(t).
\end{aligned}
$$

Multiplying the above expression for $X_1(s)$ by $s$, using Eq. (2.48) and inverse Laplace transforming, it is seen that

$$\dot{x}_1(t) = -a_0x_n(t) + b_0u(t).$$

The set of first order differential equations can then be written in matrix-vector form as

$$
\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\\ \vdots\\ \dot{x}_{n-1}\\ \dot{x}_n\end{bmatrix}
=
\begin{bmatrix}
0 & 0 & 0 & \ldots & 0 & -a_0\\
1 & 0 & 0 & \ldots & 0 & -a_1\\
0 & 1 & 0 & \ldots & 0 & -a_2\\
\vdots & \vdots & \vdots & \ldots & \vdots & \vdots\\
0 & 0 & 0 & \ldots & 0 & -a_{n-2}\\
0 & 0 & 0 & \ldots & 1 & -a_{n-1}
\end{bmatrix}
\begin{bmatrix}x_1\\ x_2\\ x_3\\ \vdots\\ x_{n-1}\\ x_n\end{bmatrix}
+
\begin{bmatrix}b_0\\ b_1\\ b_2\\ \vdots\\ b_{n-2}\\ b_{n-1}\end{bmatrix}
u = \mathbf{A}\mathbf{x} + \mathbf{B}u. \tag{2.50}
$$

The output equation can be found from Eq. (2.49),

$$y = x_n + b_nu = [\,0\,0\,0\,\ldots\,0\,1\,]\mathbf{x} + b_nu = \mathbf{C}\mathbf{x} + \mathbf{D}u. \tag{2.51}$$

A block diagram of the companion form 2 is shown on Fig. 2.16.

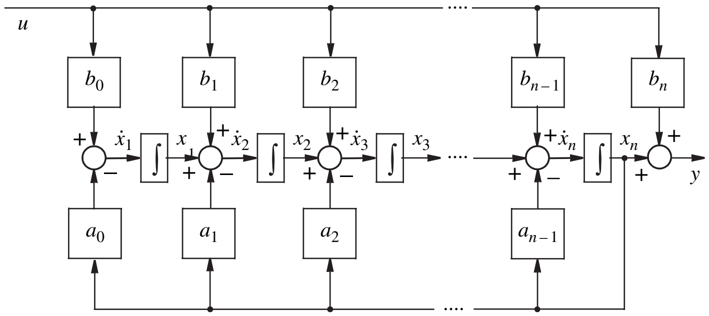


**Fig. 2.16** Block diagram of companion form 2 of a SISO system

## 2.4 Linearization

For many systems one does not have useful a priori knowledge which allows one to formulate a linear model immediately. In such cases it is necessary to start the modelling by constructing a nonlinear model using the physical laws and then - if possible - *linearizing* this model.

If one looks at the nonlinear model (2.11), it can be assumed that this equation has been solved for a nominal initial state $\tilde{\mathbf{x}}_0$, a nominal input function $\tilde{\mathbf{u}}(t)$ and a nominal disturbance function $\tilde{\mathbf{v}}(t)$. The resulting nominal state vector is supposed to be $\tilde{\mathbf{x}}(t)$. Now one would like to describe the system's behavior in the neighborhood of the nominal trajectory $\tilde{\mathbf{x}}(t)$. Assuming that a new initial state and new input and disturbance functions are defined which are 'close' to the nominal ones, one can define:

$$
\begin{aligned}
\mathbf{x}(t) &= \tilde{\mathbf{x}}(t) + \Delta\mathbf{x}(t),\\
\mathbf{u}(t) &= \tilde{\mathbf{u}}(t) + \Delta\mathbf{u}(t),\\
\mathbf{v}(t) &= \tilde{\mathbf{v}}(t) + \Delta\mathbf{v}(t),\\
\mathbf{x}_0 &= \tilde{\mathbf{x}}_0 + \Delta\mathbf{x}_0.
\end{aligned}
\tag{2.52}
$$

If it also is assumed that $\mathbf{x}(t)$ is close to $\tilde{\mathbf{x}}(t)$ and that the function $\mathbf{f}$ in (2.11) is differentiable to the first order in time, it is reasonable to expand $\mathbf{f}$ into a Taylor series about the nominal values. With the expressions in Eq. (2.52) one can write Eq. (2.11) as:

$$
\dot{\mathbf{x}}(t) = \dot{\tilde{\mathbf{x}}}(t) + \Delta\dot{\mathbf{x}}(t) = \mathbf{f}(\tilde{\mathbf{x}}(t) + \Delta\mathbf{x}(t),\ \tilde{\mathbf{u}}(t) + \Delta\mathbf{u}(t),\ \tilde{\mathbf{v}}(t) + \Delta\mathbf{v}(t),\ t). \tag{2.53}
$$

Since all terms of order higher than one are discarded, the series expansion will be:

$$\begin{aligned}\tilde{\mathbf{x}}(t) + \dot{\Delta\mathbf{x}}(t) \cong \mathbf{f}(\tilde{\mathbf{x}}(t),\, \tilde{\mathbf{u}}(t), \tilde{\mathbf{v}}(t),\, \tilde{\mathbf{v}},\, t) + \frac{\partial \mathbf{f}(\tilde{\mathbf{x}}(t),\, \tilde{\mathbf{u}}(t),\, \tilde{\mathbf{v}}(t), t)}{\partial \mathbf{x}} \Delta\mathbf{x}(t) \\ + \frac{\partial \mathbf{f}(\tilde{\mathbf{x}}(t),\, \tilde{\mathbf{u}}(t),\, \tilde{\mathbf{v}}(t), t)}{\partial \mathbf{u}} \Delta\mathbf{u}(t) + \frac{\partial \mathbf{f}(\tilde{\mathbf{x}}(t),\, \tilde{\mathbf{u}}(t),\, \tilde{\mathbf{v}}(t), t)}{\partial \mathbf{v}} \Delta\mathbf{v}(t)\end{aligned} \tag{2.54}$$

The partial derivatives which must be calculated for the nominal time functions, are in general matrices, the so called *Jacobians*. The i'th by j'th entry is the partial derivative of the i'th scalar function in **f** with respect to the j'th variable in **x**, **u** or **v**.

The Jacobian matrix $\partial \mathbf{f}/\partial \mathbf{x}$ is quadratic since **f** and **x** have the same number of elements:

$$\frac{\partial \mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial x_1} & \frac{\partial f_1(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial x_2} & \cdots & \frac{\partial f_1(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial x_n} \\ \frac{\partial f_2(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial x_1} & \frac{\partial f_2(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial x_2} & \cdots & \frac{\partial f_2(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial x_1} & \frac{\partial f_n(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial x_2} & \cdots & \frac{\partial f_n(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial x_n} \end{bmatrix}, \tag{2.55}$$

where the time argument has been omitted for simplicity. However, it is important to note that since $\tilde{\mathbf{x}}(t), \tilde{\mathbf{u}}(t)$ and $\tilde{\mathbf{x}}(t)$ are functions of time, the entries of the matrices are also in general functions of time (even if **f** is *not explicitly* a function of time). If **f** is not explicitly a function of time then the nonlinear system (2.11) is time invariant,

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{v}(t)). \tag{2.56}$$

From the assumptions above it is clear that

$$\dot{\tilde{\mathbf{x}}}(t) = \mathbf{f}(\tilde{\mathbf{x}}(t), \tilde{\mathbf{u}}(t), \tilde{\mathbf{v}}(t), t). \tag{2.57}$$

Inserting this into Eq. (2.54), one obtains:

$$\dot{\Delta\mathbf{x}}(t) = \mathbf{A}(t)\Delta\mathbf{x}(t) + \mathbf{B}(t)\Delta\mathbf{u}(t) + \mathbf{B}_v(t)\Delta\mathbf{v}(t) \tag{2.58}$$

where

$$\mathbf{A}(t) = \frac{\partial \mathbf{f}(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, \mathbf{t})}{\partial \mathbf{x}},\ \mathbf{B}(t) = \frac{\partial f(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial \mathbf{u}},\ \mathbf{B}_v(t) = \frac{\partial f(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}}, t)}{\partial \mathbf{v}}. \tag{2.59}$$

Equation (2.58) a *linear approximation* to the nonlinear system (2.11). The linear approximation expresses the behavior of the system in the close vicinity of the nominal trajectory $\tilde{\mathbf{x}}(t)$. The state variables here are the *incremental* state variables.

If the problem involves a nonlinear output equation (see Eq. (2.8)), the linearization of the output function follows exactly the same lines. The linear approximation to the output equation $\mathbf{y}(t) = \mathbf{g}(\mathbf{x}(t), \mathbf{u}(t), t)$ is (it is assumed that $\mathbf{y}(t)$ is not a function of the disturbance),

$$\Delta\mathbf{y}(t) = \mathbf{C}(t)\Delta\mathbf{x}(t) + \mathbf{D}(t)\Delta\mathbf{u}(t), \tag{2.60}$$

where the large signal output is defined as

$$\mathbf{y}(t) = \tilde{\mathbf{y}}(t) + \Delta\mathbf{y}(t). \tag{2.61}$$

The relevant matrices are found from:

$$\mathbf{C}(t) = \frac{\partial\mathbf{g}(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, t)}{\partial\mathbf{x}}, \quad \mathbf{D}(t) = \frac{\partial\mathbf{g}(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}, t)}{\partial\mathbf{u}}. \tag{2.62}$$

A very important special case arises if the nominal matrix elements are all *constants*. In this case the initial state and the nominal state can be considered the same and this state is called a *stationary state*. These are natural bases for the linearization. A stationary state is characterized by zero time derivatives and from Eq. (2.11) it is seen that that the stationary states must satisfy the nonlinear *algebraic* equation:

$$\mathbf{0} = \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0). \tag{2.63}$$

Note that here the treatment is confined to the time invariant case.

Having determined the stationary states from Eq. (2.63) (there may be more than one), one can define, as in (2.52) and (2.61):

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{x}_0 + \Delta\mathbf{x}(t), \\ \mathbf{u}(t) &= \mathbf{u}_0 + \Delta\mathbf{u}(t), \\ \mathbf{v}(t) &= \mathbf{v}_0 + \Delta\mathbf{v}(t), \\ \mathbf{y}(t) &= \mathbf{y}_0 + \Delta\mathbf{y}(t), \end{aligned} \tag{2.64}$$

The $\Delta$-variables above denote the (small) deviations from the stationary (constant) values and are called the *incremental states, inputs, disturbances* and *outputs*.

As before, the **f** vector function can be expanded about the stationary values to find:

$$\begin{aligned} \dot{\mathbf{x}}(t) = \dot{\mathbf{x}}_0 + \Delta\dot{\mathbf{x}}(t) \cong \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0) + \frac{\partial\mathbf{f}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0)}{\partial\mathbf{x}}\Delta\mathbf{x}(t) \\ + \frac{\partial\mathbf{f}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0)}{\partial\mathbf{u}}\Delta\mathbf{u}(t) + \frac{\partial\mathbf{f}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0)}{\partial\mathbf{v}}\Delta\mathbf{v}(t). \end{aligned} \tag{2.65}$$

The Jacobians are now constant matrices, e.g.:

$$\frac{\partial \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0)}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}_0 = \mathbf{A}. \tag{2.66}$$

The subscript 0 indicates that all entries in the matrix are calculated at the stationary (linearization) points.

$\mathbf{x}_0$ is constant and consequently $\dot{\mathbf{x}}_0 = \mathbf{0}$ and $\dot{\mathbf{x}}(t) = \dot{\Delta\mathbf{x}}(t)$. With the notation from Eq. (2.66), equation (2.65) becomes the linearized state equation:

$$\dot{\Delta\mathbf{x}}(t) = \mathbf{A}\Delta\mathbf{x}(t) + \mathbf{B}\Delta\mathbf{u}(t) + \mathbf{B}_v\Delta\mathbf{v}, \tag{2.67}$$

where

$$\mathbf{B} = \frac{\partial \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0)}{\partial \mathbf{u}} = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \frac{\partial f_1}{\partial u_2} & \dots & \frac{\partial f_1}{\partial u_m} \\ \frac{\partial f_2}{\partial u_1} & \frac{\partial f_2}{\partial u_2} & \dots & \frac{\partial f_2}{\partial u_m} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial u_1} & \frac{\partial f_n}{\partial u_2} & \dots & \frac{\partial f_n}{\partial u_m} \end{bmatrix}_0 \tag{2.68}$$

and

$$\mathbf{B}_v = \frac{\partial \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0)}{\partial \mathbf{v}} = \begin{bmatrix} \frac{\partial f_1}{\partial v_1} & \frac{\partial f_1}{\partial v_2} & \dots & \frac{\partial f_1}{\partial v_p} \\ \frac{\partial f_2}{\partial v_1} & \frac{\partial f_2}{\partial v_2} & \dots & \frac{\partial f_2}{\partial v_p} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial v_1} & \frac{\partial f_n}{\partial v_2} & \dots & \frac{\partial f_n}{\partial v_p} \end{bmatrix}_0 . \tag{2.69}$$

As before, Eq. (2.67) is called a linear approximation to Eq. (2.11). Since the matrices are constant, Eq. (2.67) is a LTI-model.

If the output equation is nonlinear as in Eq. (2.8), the problem is treated in the same way. The linearized output equation is:

$$\Delta\mathbf{y}(t) = \mathbf{C}\Delta\mathbf{x}(t) + \mathbf{D}\Delta\mathrm{u}(t), \tag{2.70}$$

where

$$\mathbf{C} = \frac{\partial \mathbf{g}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0)}{\partial \mathbf{x}} \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g_r}{\partial x_1} & \frac{\partial g_r}{\partial x_2} & \cdots & \frac{\partial g_r}{\partial x_n} \end{bmatrix}_0 \tag{2.71}$$

and

$$\mathbf{D} = \frac{\partial \mathbf{g}(\mathbf{x}_0, \mathbf{u}_0, \mathbf{v}_0)}{\partial \mathbf{u}} \begin{bmatrix} \frac{\partial g_1}{\partial u_1} & \frac{\partial g_1}{\partial u_2} & \cdots & \frac{\partial g_1}{\partial u_m} \\ \frac{\partial g_2}{\partial u_1} & \frac{\partial g_2}{\partial u_2} & \cdots & \frac{\partial g_2}{\partial u_m} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g_r}{\partial u_1} & \frac{\partial g_r}{\partial u_2} & \cdots & \frac{\partial g_r}{\partial u_m} \end{bmatrix}_0 . \tag{2.72}$$

***Example 2.7.* Rocket with Air Resistance**

This example deals with a rocket under the influence of gravity, an upward thrust and a nonlinear wind resistance force. It is assumed that the rocket moves in a vertical direction and that the long axis of the rocket is constantly vertical. A sketch of the rocket is shown in Fig. 2.17. The $x$ position axis points upwards and the velocity is called $v$.

Newton's second law says that

$$m\dot{v} = -F_a - F_g + T = \pm bv^2 - mg + T$$

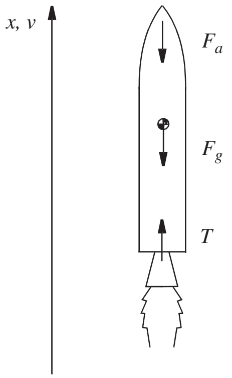


**Fig. 2.17** A rocket in vertical motion

or

$$\dot{v} = \pm \frac{b}{m} v^2 - g + \frac{1}{m} T = f(v, T), \tag{2.73}$$

where $m$ is the mass (which is assumed to be constant here), $b$ is the air resistance coefficient, $g$ is the acceleration due to gravity and $T$ is the thrust. When the rocket moves upwards, then the wind resistance is a downwards force. So the minus sign of the first right hand term is valid for upwards movement ($v > 0$) and the plus sign for downwards movement ($v < 0$). Equation (2.73) shows that the model is first order. The velocity is the only state variable, $T$ is the input variable ($u = T$) and the output is equal to the state, $y = v$.

The stationary states can be obtained by setting the time derivative of $v$ equal to zero. This leads to the following relationship between the stationary values of the variables,

$$v_0 = \begin{cases} \sqrt{\frac{1}{b}(T_0 - mg)} & \text{for} \quad v > 0 \\ -\sqrt{\frac{1}{b}(mg - T_0)} & \text{for} \quad v < 0. \end{cases}$$

A positive value for $v_0$ is only possible if $T_0 > mg$. A negative value requires that $T_0 < mg$. If the thrust is zero, the rocket will fall and reach the terminal velocity,

$$v_0 = v_{term} = -\sqrt{\frac{mg}{b}}.$$

As in (2.64) the deviations from the stationary values are defined as

$$v(t) = v_0 + \Delta v(t),$$
$$T(t) = T_0 + \Delta T(t),$$

and the linearized system model can be derived according to Eqs. (2.66), (2.67) and (2.68). The state equation of the system is thus

$$\dot{\Delta v} = A \Delta v + B \Delta T, \tag{2.74}$$

where

$$A = \left.\frac{d(f(v,T))}{dv}\right|_0 = \begin{cases} -\dfrac{2b}{m}v_0 & \text{for} \quad v > 0 \\ \dfrac{2b}{m}v_0 & \text{for} \quad v < 0, \end{cases}$$

$$B = \left.\frac{d(f(v,T))}{dT}\right|_0 = \frac{1}{m}.$$

The first order state Eq. (2.74) becomes

$$\dot{\Delta v} = -\left|\frac{2b}{m}v_0\right|\Delta v + \frac{1}{m}\Delta T.$$

This linear differential equation can be Laplace transformed and the following expression is found (for $v(0) = 0$):

$$s\Delta v(s) = -\left|\frac{2b}{m}v_0\right|\Delta v(s) + \frac{1}{m}\Delta T(s),$$

from which one obtains

$$\frac{\Delta v(s)}{T(s)} = \frac{1/m}{s + \left|\dfrac{2b}{m}v_0\right|},$$

which is the transfer function of a first order system. The time constant is

$$\tau = \left|\frac{m}{2bv_0}\right|$$

and therefore it is known from classical control system analysis that the system is stable for upwards as well as downwards movement. Note that the time constant is a function not only of $m$ and $b$ but also of $v_0$. A high velocity results in a small time constant and vice versa. The basic dynamic characteristics of the system are thus dependent on the stationary state in which the linearization is carried out. This is typical for nonlinear systems.

It is also typical for nonlinear systems that the differential equations cannot be solved analytically. This problem will be addressed in the next chapter. In the present simple case a solution can be found by separation of the variables in Eq. (2.73). Assuming that the thrust $T$ *is* equal to zero it is found that

$$\frac{dv}{\frac{b}{m}v^2 - g} = dt.$$

This equation can be integrated directly to yield,

$$v(t) = -\sqrt{\frac{mg}{b}}\tanh\left(\sqrt{\frac{bg}{m}}t\right) = \begin{cases} -gt,\ t < < \sqrt{m/(bg)} \\ -\sqrt{mg/b},\ t >> \sqrt{m/(bg)} \end{cases}, \tag{2.75}$$

if it is in addition assumed that the rocket is initially at rest, i.e., $v(0) = 0$. The solution shows that for small values of $t$ and therefore at low velocity, the rocket moves as if it were only subject to gravity and the constant acceleration $g$. At large $t$ and $v$, the velocity tends to the terminal velocity $v_{term}$.

Integrating Eq. (2.75) immediately gives the position of the rocket (assuming that $x(0) = x_0$),

$$\begin{aligned} x(t) - x_0 &= -\frac{m}{b}\ln\left(\cosh\left(\sqrt{\frac{bg}{m}}t\right)\right) \\ &= \begin{cases} (1/2)gt^2,\ t < < \sqrt{m/(bg)} \\ \left(\left(\frac{m}{b}\right)\ln(2) - \sqrt{(mg)/bt}\right),\ t >> \sqrt{m/(bg)}. \end{cases} \end{aligned}$$

❐

***Example 2.8*. Hydraulic Velocity Servo**

A simplified model for a hydraulic velocity servo can be expressed as follows

$$\begin{aligned} \frac{d}{dt}v(t) &= Ap(t) - cv(t), \\ \frac{d}{dt}p(t) &= aq(t) - Av(t), \\ q(t) &= k\sqrt{p(t)}u(t). \end{aligned}$$

$v$ is the piston velocity, $p$ is the differential pressure over the piston, $q$ is the volume flow from the servo valve and $u$ is the input voltage to the valve. $A$ is the piston area, $c$ is a viscous friction coefficient, $a$ is a constant related to the stiffness of the hydraulic fluid and $k$ is a valve coefficient.

The first step in the linearization procedure is to determine the stationary state. This is accomplished by setting the time derivatives to zero. With the initial elimination of $q$, the results are

$$0 = Ap_0 - cv_0,$$
$$0 = ak\sqrt{p_0}u_0 - Av_0.$$

The stationary values of the variables can be calculated for a given constant input voltage, $u_0$,

$$p_0 = \frac{a^2 k^2 c^2}{A^4} u_0^2,$$
$$v_0 = \frac{a^2 k^2 c}{A^3} u_0^2,$$
$$q_0 = \frac{ak^2 c}{A^2} u_0^2.$$

It is natural to choose the state and output variables,

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} v(t) \\ p(t) \end{bmatrix} \text{ and } y(t) = v(t).$$

The nonlinear state equation becomes

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} Ap(t) - cv(t) \\ ak\sqrt{p(t)}u(t) - Av(t) \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}(t), u(t)) \\ f_2(\mathbf{x}(t), u(t)) \end{bmatrix}.$$

After defining the variables and their deviations by the expressions

$$v(t) = v_0 + \Delta v(t),$$
$$p(t) = p_0 + \Delta p(t),$$
$$q(t) = q_0 + \Delta q(t),$$

the linear state model can be derived. One finds

$$\dot{\Delta\mathbf{x}}(t) = \mathbf{A}\Delta\mathbf{x}(t) + \mathbf{B}\Delta u(t),$$
$$\Delta y(t) = \mathbf{C}\Delta\mathbf{x}(t) + D\Delta u(t),$$

with the matrices,

$$A = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} \end{bmatrix}_0 = \begin{bmatrix} -c & A \\ -A & \dfrac{aku_0}{2\sqrt{p_0}} \end{bmatrix},$$

$$B = \begin{bmatrix} \dfrac{\partial f_1}{\partial u} \\ \dfrac{\partial f_2}{\partial u} \end{bmatrix}_0 = \begin{bmatrix} 0 \\ ak\sqrt{p_0} \end{bmatrix},$$

$$C = [\,1 \quad 0\,], D = 0$$

□

***Example 2.9.* Water Tank Process Plant**

Figure 2.18 shows water tank process for which a model is to be formulated.

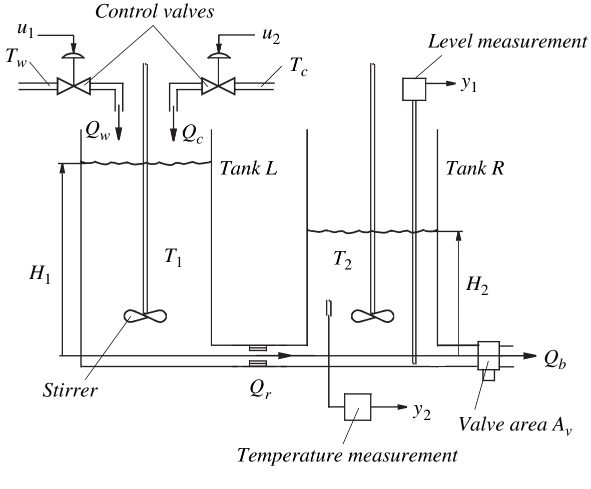


**Fig. 2.18** Water tank process plant

Two tanks denoted $L$ and $R$ are connected as shown and into the left tank can be pumped warm and cold water through control valves with the two input signals $U_1$ and $U_2$. The temperatures and the volume flows are $T_w$, $T_c$, $Q_w$ and $Q_c$. The water levels in the two tanks are $H_1$ and $H_2$ respectively and the tanks have the same cross sectional area $A$. The flow between the tanks is $Q_r$ and the flow out of the outlet valve of tank R is $Q_b$. This last valve has the variable opening area $A_v$. The water is stirred rapidly in both tanks and therefore the temperature is assumed to be constant over the entire volume of each of the tanks.

Two quantities can be measured on the plant: the level and the temperature of tank R. For the two measurement systems one has that:

$$y_1 = k_h H_2, \tag{2.76}$$

$$y_2 = k_t T_2, \tag{2.77}$$

where $k_h$ and $k_t$ are transducer gains.

The two control valves have the same flow characteristics and it is assumed that the following two relations are valid,

$$Q_w = k_a u_1, \tag{2.78}$$

$$Q_c = k_a u_2, \tag{2.79}$$

where $k_a$ is the flow coefficient.

The mathematical model of the plant involves volume and energy conservation laws and suitable relations describing flow through orifices. Conservation of fluid volume gives for the two tanks:

$$A\dot{H}_1 = Q_2 + Q_c - Q_r, \tag{2.80}$$

$$A\dot{H}_2 = Q_r - Q_b. \tag{2.81}$$

The energy content in the water volumes can be written:

$$E_L = AH_1\rho c(T_1 - T_0), \tag{2.82}$$

$$E_R = AH_2\rho c(T_2 - T_0). \tag{2.83}$$

$\rho$ and $c$ are the mass density and the specific heat capacity of water. $T_0$ is the reference temperature at which the energy is zero. It is easy to show that one can set $T_0 = 0$ and the energy equations are reduced to

$$E_L = AH_1\rho c T_1, \tag{2.84}$$

$$E_R = AH_2\rho c T_2. \tag{2.85}$$

For the flow through the orifices it is reasonable to assume that a square root relation is valid. The formula for this can be written,

$$Q = C_d A_o \sqrt{\frac{2}{\rho}\Delta P}, \tag{2.86}$$

where $\Delta P$ is the differential pressure over the orifice, $\rho$ is the mass density, $A_o$ is the area of the orifice and $C_d$ is a constant loss coefficient. The hydrostatic pressure in a liquid at a level $H$ below the surface is $P_h = \rho g H + P_a$ (where the atmospheric pressure is $P_a$), where $g$ is the acceleration due to gravity. The flow through the outlet valve can thus be written,

$$Q_b = D_v A_v \sqrt{H_2}, \tag{2.87}$$

where $D_v = C_d\sqrt{2g}$. The orifice between the tanks has a constant flow area and one can write,

$$Q_r = C_0 \sqrt{H_1 - H_2}, \tag{2.88}$$

where it is assumed that $H_1 > H_2$.

Now the fact can be utilized that the net power flux into the tanks equals the accumulated energy per time unit. Thus the time derivatives of the energy expressions (2.84) and (2.85) give the left hand sides of two new equations,

$$\frac{dE_L}{dt} = A\rho c\frac{d}{dt}(H_1 T_1) = Q_w \rho c T_w + Q_c \rho c T_c - Q_r \rho c T_1, \tag{2.89}$$

$$\frac{dE_R}{dt} = A\rho c\frac{d}{dt}(H_2 T_2) = Q_r \rho c T_1 - Q_b \rho c T_2. \tag{2.90}$$

Dividing all terms by $\rho c$ and differentiating the product yields:

$$A(\dot{H}_1 T_1 + H_1 \dot{T}_1) = Q_w T_w + Q_c T_c - Q_r T_1, \tag{2.91}$$

$$A(\dot{H}_2 T_2 + H_2 \dot{T}_2) = Q_r T_1 - Q_b T_2. \tag{2.92}$$

Inserting Eqs. (2.80) and (2.81) into Eqs. (2.91) and (2.92) respectively gives the final system equations:

$$AH_1\dot{T} = Q_wT_w + Q_cT_c - Q_cT_1 - Q_wT_1, \tag{2.93}$$

$$AH_2\dot{T}_2 = Q_rT_1 - Q_rT_2 \tag{2.94}$$

The complete model of the plant now consists of the Eqs. (2.76), (2.77), (2.78), (2.79), (2.80), (2.81), (2.87), (2.88), (2.93) and (2.94). A block diagram based on this set of equations can be seen on Fig. 2.19.

The natural choice of states is the output variables of the four integrators. The outlet valve area and the two inlet temperatures are disturbances and the two control valve voltages are the manipulable inputs. So, state, input and disturbance vectors will be:

$$\mathbf{x}(t) = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} H_1 \\ H_2 \\ T_1 \\ T_2 \end{bmatrix}, \ \mathbf{u}(t) = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \ \mathbf{v}(t) \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} A_v \\ T_w \\ T_c \end{bmatrix}. \tag{2.95}$$

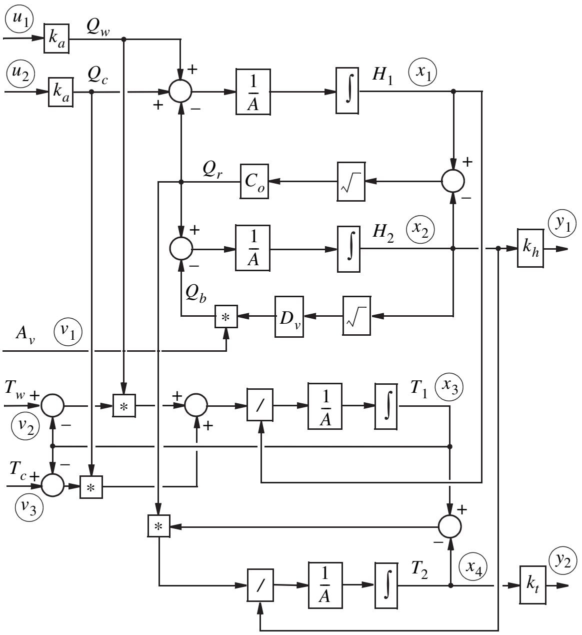


**Fig. 2.19** Block diagram of the process plant

Now there are four state equations and from the block diagram:

$$\begin{aligned}
\dot{x}_1 &= \frac{1}{A}[k_a(u_1 + u_2) - C_0\sqrt{x_1 - x_2}\,],\\
\dot{x}_2 &= \frac{1}{A}[C_0\sqrt{x_1 - x_2} - D_v\sqrt{x_2}v_1],\\
\dot{x}_3 &= \frac{1}{Ax_1}[(v_2 - x_3)k_a u_1 + (v_3 - x_3)k_a u_2],\\
\dot{x}_4 &= \frac{1}{Ax_2}(x_3 - x_4)C_0\sqrt{x_1 - x_2}.
\end{aligned} \tag{2.96}$$

The output equation is linear in the states,

$$\mathbf{y}(t) = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 & k_h & 0 & 0 \\ 0 & 0 & 0 & k_t \end{bmatrix}\mathbf{x}(t). \tag{2.97}$$

The state Equations (2.96) can be written in the form (2.56) and it is natural to linearize the model around a stationary operating point found using the Formula (2.63). The equations (2.96), with their left hand sides set to zero, lead to

$$\begin{aligned}
&k_a(u_{10} + u_{20}) = C_0\sqrt{x_{10} - x_{20}},\\
&C_0\sqrt{x_{10} - x_{20}} = D_v\sqrt{x_{20}}v_{10},\\
&(v_{20} - x_{30})u_{10} = -(v_{30} - x_{30})u_{20},\\
&(x_{30} - x_{40})\sqrt{x_{10} - x_{20}} = 0.
\end{aligned} \tag{2.98}$$

The extra subscript (zero) denotes that these are stationary values. The 4 equations in (2.98) contain 9 variables. If for example values for the 2 inputs and the 3 disturbance variables are selected, the 4 states can be determined from Eq. (2.98).

The following parameter values will be assumed:

$$
\begin{aligned}
A &= 0.785\,\mathrm{m}^2,\\
D_v &= 2.66\,\mathrm{m}^{1/2}/\mathrm{sec},\\
C_0 &= 0.056\,\mathrm{m}^{5/2}/\mathrm{sec},\\
k_a &= 0.004\,\mathrm{m}^3/\mathrm{volt}\cdot\mathrm{sec},\\
k_h &= 2\,\mathrm{volt/m},\\
k_t &= 0.1\,\mathrm{volt}/{}^\circ\mathrm{C},
\end{aligned}
$$

assuming that the input voltages can take on values in the interval 0–10 volts and choosing $u_{10} = u_{20} = 5$ volts. Further, setting $A_{v0} = 0.0122\,\mathrm{m}^2$, $T_{w0} = 60^\circ\mathrm{C}$ and $T_{c0} = 30^\circ\mathrm{C}$, the Eq. (2.98) give the values of the stationary states,

$$
\begin{aligned}
x_{10} &= 2.03\,\mathrm{m},\\
x_{20} &= 1.519\,\mathrm{m},\\
x_{30} &= x_{40} = 45^\circ\mathrm{C}.
\end{aligned}
$$

The matrices for the linearized system can be calculated according to Eqs. (2.66), (2.68), (2.69) and (2.71):

$$
\mathbf{A} = \begin{bmatrix}
\dfrac{-C_0}{2A\sqrt{x_{10}-x_{20}}} & \dfrac{C_0}{2A\sqrt{x_{10}-x_{20}}} & 0 & 0\\
\dfrac{C_0}{2A\sqrt{x_{10}-x_{20}}} & \dfrac{-C_0}{2A\sqrt{x_{10}-x_{20}}} - \dfrac{D_v v_{10}}{2A\sqrt{x_{20}}} & 0 & 0\\
0 & 0 & -\dfrac{k_a u_{10}+k_a u_{20}}{A x_1 0} & 0\\
0 & 0 & \dfrac{C_0\sqrt{x_{10}-x_{20}}}{A x_{20}} & -\dfrac{C_0\sqrt{x_{10}-x_{20}}}{A x_{20}}
\end{bmatrix}
\tag{2.99}
$$

$$
\mathbf{B} = \begin{bmatrix}
\dfrac{k_a}{A} & \dfrac{k_a}{A}\\
0 & 0\\
\dfrac{k_a(v_{20}-x_{30})}{A x_{10}} & \dfrac{k_a(v_{30}-x_{30})}{A x_{10}}\\
0 & 0
\end{bmatrix},
\tag{2.100}
$$

$$\mathbf{B}_v = \begin{bmatrix} 0 & 0 & 0 \\ -\dfrac{D_v\sqrt{x_{20}}}{A} & 0 & 0 \\ 0 & \dfrac{k_a u_{10}}{A x_{10}} & \dfrac{k_a u_{20}}{A x_{10}} \\ 0 & 0 & 0 \end{bmatrix}, \tag{2.101}$$

$$\mathrm{C} = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}. \tag{2.102}$$

With the parameter and stationary variable values above one can find:

$$\mathbf{A} = \begin{bmatrix} -0.0499 & 0.0499 & 0 & 0 \\ 0.0499 & -0.0667 & 0 & 0 \\ 0 & 0 & -0.0251 & 0 \\ 0 & 0 & 0.0335 & -0.0335 \end{bmatrix}, \tag{2.103}$$

$$\mathrm{B} = \begin{bmatrix} 0.00510 & 0.00510 \\ 0 & 0 \\ 0.0377 & -0.0377 \\ 0 & 0 \end{bmatrix}, \tag{2.104}$$

$$\mathbf{B}_v = \begin{bmatrix} 0 & 0 & 0 \\ -4.177 & 0 & 0 \\ 0 & 0.01255 & 0.01255 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.105}$$

The linearized model describes the behavior of deviations from the stationary values. If the incremental system vectors are defined as in Eq. (2.64):

$$\Delta\mathbf{x}(t) = \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \\ \Delta x_4 \end{bmatrix} = \begin{bmatrix} \Delta H_1 \\ \Delta H_2 \\ \Delta T_1 \\ \Delta T_2 \end{bmatrix}, \; \Delta\mathbf{u}(t) = \begin{bmatrix} \Delta u_1 \\ \Delta u_2 \end{bmatrix}, \tag{2.106}$$

$$\Delta(t) = \begin{bmatrix} \Delta v_1 \\ \Delta v_2 \\ \Delta v_3 \end{bmatrix} = \begin{bmatrix} \Delta A_v \\ \Delta T_w \\ \Delta T_c \end{bmatrix},$$

where

$$H_1(t) = H_{10} + \Delta H_1(t) \tag{2.107}$$

and likewise for the remainder of the variables one ends up with the state space model,

$$\dot{\Delta\mathbf{x}}(t) = \mathbf{A}\Delta\mathbf{x}(t) + \mathbf{B}\Delta\mathbf{u}(t) + \mathbf{B}_v\Delta\mathbf{v}(t), \tag{2.108}$$

$$\Delta\mathbf{y}(t) = \mathbf{C}\Delta\mathbf{x}(t).$$

***Example 2.10.*** **Two Link Robot Arm**

A two-link robot is shown in Fig. 2.20. The robot position is defined by the two angles, $\theta_1$ and $\theta_2$. The two robot links have the following characteristics:

Link 1: length $l_1$, total mass $m_1$, moment of inertia $J_1$.
Link 2: length $l_2$, total mass $m_2$, moment of inertia $J_2$.

It is assumed that the two links have a symmetric mass distribution so that the centre of gravity is in the middle of the link. In order to formulate the equations of motion the so-called Lagrange method can be used. This method defines a set of generalized coordinates that specify the position of the system uniquely. In this case the generalized coordinates are the two angles $\theta_1$ and $\theta_2$. The Lagrangian is defined as the difference between the kinetic and potential energies,

$$L(\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2) = K - P. \tag{2.109}$$

Here, $K$ and $P$ are the total kinetic and the total potential energy of the system expressed in terms of the generalized coordinates and their derivatives, the generalized velocities.

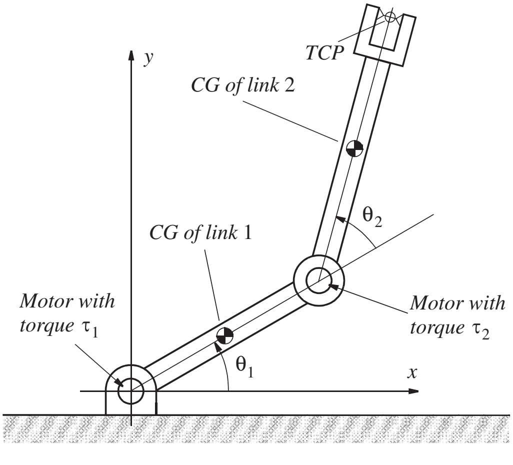


**Fig. 2.20** Two-link robot arm

The equations of motion can now be derived as the following Lagrange equations,

$$\begin{aligned}\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_1}\right) - \frac{\partial L}{\partial \theta_1} &= \tau_1,\\ \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_2}\right) - \frac{\partial L}{\partial \theta_1} &= \tau_2.\end{aligned} \tag{2.110}$$

These two equations are fully equivalent to Newton's equations of motion.

The kinetic energy of the total system is the sum of the kinetic energy of each link and this in turn is the sum of the kinetic energy due to linear motion of the centre of gravity and the kinetic energy of the rotation of the link around the centre of gravity, i.e.,

$$K = \frac{1}{2}m_1\mathbf{v}_{G_1}^2 + \frac{1}{2}J_1\dot{\theta}_1^2 + \frac{1}{2}m_2\mathbf{v}_{G_2}^2 + \frac{1}{2}J_2(\dot{\theta}_1 + \dot{\theta}_2)^2. \tag{2.111}$$

If it is assumed that the center of gravity is positioned in the middle of each of the links, the cartesian coordinates of the two centers are

$$\begin{aligned}(x_{G_1}, y_{G_1}) &= \left(\frac{l_1}{2}\cos\theta_1, \frac{l_1}{2}\sin\theta_1\right),\\ (x_{G_2}, y_{G_2}) &= \left(l_1\cos\theta_1 + \frac{l_2}{2}\cos(\theta_1+\theta_2), l_1\sin\theta_1 + \frac{l_2}{2}\sin(\theta_1+\theta_2)\right).\end{aligned} \tag{2.112}$$

The two velocity vectors will be

$$\begin{aligned}\mathbf{v}_{G_1} &= \frac{d}{dt}(x_{G_1}, y_{G_1}),\\ \mathbf{v}_{G_2} &= \frac{d}{dt}(x_{G_2}, y_{G_2}),\end{aligned} \tag{2.113}$$

Evaluating these derivatives and evaluating their squares in Eq. (2.111) gives for the kinetic energy, $K$,

$$\begin{aligned}K = &\left(\frac{1}{2}m_1\frac{l_1^2}{4} + \frac{1}{2}J_1 + \frac{1}{2}m_2 l_1^2\right)\dot{\theta}_1^2\\ &+ \left(\frac{1}{2}m_2\frac{l_2^2}{4} + \frac{1}{2}J_2\right)(\dot{\theta}_1 + \dot{\theta}_2)^2\\ &+ \frac{1}{2}m_2 l_1 l_2\cos(\theta_2)\dot{\theta}_1(\dot{\theta}_1 + \dot{\theta}_2).\end{aligned} \tag{2.114}$$

The other term in the Lagrangian is the potential energy of the robot links due to gravity,

$$P = m_1 g \frac{l_1}{2} \sin \theta_1 + m_2 g \left( l_1 \sin \theta_1 + \frac{l_2}{2} \sin(\theta_1 + \theta_2) \right). \tag{2.115}$$

The two expressions in Eqs. (2.114) and (2.115) respectively are inserted into the Lagrange equation and after some straightforward calculation one arrives at the following equations of motion:

$$\begin{aligned} M_{11}\ddot{\theta}_1 + M_{12}\ddot{\theta}_2 + K_1(\theta_2, \dot{\theta}_1, \dot{\theta}_2) + G_1(\theta_1, \theta_2) &= \tau_1, \\ M_{21}\ddot{\theta}_1 + M_{22}\ddot{\theta}_2 + K_2(\theta_2, \dot{\theta}_1, \dot{\theta}_2) + G_2(\theta_1, \theta_2) &= \tau_2, \end{aligned} \tag{2.116}$$

where the inertial components are,

$$\begin{aligned} M_{11} &= \left(\frac{1}{4}m_1 + m_2\right) l_1^2 + \frac{1}{4} m_2 l_2^2 + J_1 + J_2 + m_2 l_1 l_2 \cos \theta_2, \\ M_{12} &= M_{21} = \frac{1}{4} m_2\, l_2^2 + J_2 + \frac{1}{2} m_2 l_1 l_2 \cos \theta_2, \\ M_{22} &= \frac{1}{4} m_2 l_2^2 + J_2, \\ \mathbf{M} &= \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}. \end{aligned} \tag{2.117}$$

The centrifugal and coriolis force components are,

$$\begin{aligned} K_1(\theta_2, \dot{\theta}_1, \theta_2) &= -\frac{1}{2} m_2 l_1 l_2 \sin(\theta_2) \cdot \dot{\theta}_2 (2\dot{\theta}_1 + \dot{\theta}_2), \\ K_2(\theta_2, \dot{\theta}_1, \dot{\theta}_2) &= \frac{1}{2} m_2 l_1 l_2 \sin(\theta_2) \cdot \dot{\theta}_1^2. \end{aligned} \tag{2.118}$$

Finally, the terms due to gravity are,

$$\begin{aligned} G_1(\theta_1, \theta_2) &= m_1 g \frac{l_1}{2} \cos \theta_1 + m_2 g \left( l_1 \cos \theta_1 + \frac{l_2}{2} \cos(\theta_1 + \theta_2) \right), \\ G_2(\theta_1, \theta_2) &= m_2 g \frac{l_2}{2} \cos(\theta_1 + \theta_2). \end{aligned} \tag{2.119}$$

The complete robot system is a control object with two inputs. Normally the torques are delivered by DC or AC motors with their own specific dynamics but in the present case it will be assumed that the control inputs are the torques themselves,

$$\mathbf{u}(t) = \begin{bmatrix} \tau_1(t) \\ \tau_2(t) \end{bmatrix}. \tag{2.120}$$

The state vector is defined as

$$\mathbf{x}(t) = \begin{bmatrix} \theta_1(t) \\ \dot{\theta}_1(t) \\ \theta_2(t) \\ \dot{\theta}_2(t) \end{bmatrix} \tag{2.121}$$

With $\mathbf{H} = \mathbf{M}^{-1}$ the non-linear state equation for the two-link robot takes the following form:

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} x_2(t) \\ -H_{11}(K_1 + G_1) - H_{12}(K_2 + G_2) \\ x_4(t) \\ -H_{21}(K_1 + G_1) - H_{22}(k_2 + G_2) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ H_{11} & H_{12} \\ 0 & 0 \\ H_{21} & H_{22} \end{bmatrix} \mathbf{u}(t). \tag{2.122}$$

In order to linearize this model, it is first noted that the coriolis and centrifugal forces are all quadratic in the angular velocity, so that for any linearization around a stationary point, i.e., one with $\dot{\theta}_{10} = \dot{\theta}_{20} = 0$, these terms disappear. Therefore for linearization around $\mathbf{x}_0$ the state variables are

$$\mathbf{x}(t) = \mathbf{x}_0 + \Delta\mathbf{x}(t) = \begin{bmatrix} \theta_{10} \\ 0 \\ \theta_{20} \\ 0 \end{bmatrix} + \Delta\mathbf{x}(t),$$

$$\mathbf{u}(t) = \mathbf{u}_0 + \Delta\mathbf{u}(t) = \begin{bmatrix} \tau_{10} \\ \tau_{20} \end{bmatrix} + \begin{bmatrix} \Delta\tau_1(t) \\ \Delta\tau_2(t) \end{bmatrix}.$$

From these equations the incremental linear state space model is obtained,

$$\dot{\Delta\mathbf{x}}(t) = \mathbf{A}\Delta\mathbf{x}(t) + \mathbf{B}\Delta\mathbf{u}(t), \tag{2.123}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\left.\dfrac{\partial}{\partial\theta_1}(H_{11}G_1 + H_{12}G_2)\right|_0 & 0 & -\left.\dfrac{\partial}{\partial\theta_2}(H_{11}G_1 + H_{12}G_2)\right|_0 & 0 \\ 0 & 0 & 0 & 1 \\ -\left.\dfrac{\partial}{\partial\theta_1}(H_{21}G_1 + H_{22}G_2)\right|_0 & 0 & -\left.\dfrac{\partial}{\partial\theta_2}(H_{21}G_1 + H_{22}G_2)\right|_0 & 0 \end{bmatrix} \tag{2.124}$$

and

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ H_{11}|_0 & H_{12}|_0 \\ 0 & 0 \\ H_{21}|_0 & H_{22}|_0 \end{bmatrix}. \tag{2.125}$$

In control applications the measured variables are often the joint angles so that the output matrix will be

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t)\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\mathbf{x}(t). \tag{2.126}$$

If the joint velocities are also measured, the **C** matrix will be the identity matrix.

In some real applications one is interested in the cartesian coordinates of the tool-centre-point (TCP), i.e., the end point of link 2, where a tool is attached. In this case the output expression will be nonlinear,

$$\mathbf{y}(t) = \begin{bmatrix} l_1 \cos\theta_1 + l_2 \cos(\theta_1 + \theta_2) \\ l_1 \sin\theta_1 + l_2 \sin(\theta_1 + \theta_2) \end{bmatrix}. \tag{2.127}$$

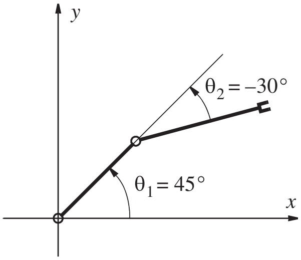
**Numerical Example**

The matrices for a small robot with the following data will now be calculated using the parameter values,

$$m_1 = 3\,\text{kg},\ l_1 = 0.6\,\text{m},\ J_1 = 0.12\,\text{kgm}^2,$$
$$m_2 = 2.5\,\text{kg},\ l_2 = 0.8\,\text{m},\ J_2 = 0.15\,\text{kgm}^2.$$

The first step of the linearization is to determine a stationary point. Inspecting Eq. (2.116), it is seen that with all derivatives set to zero, a set of simple expressions relating the remaining quantities at the stationary point can be found,

**Fig. 2.21** Robot configuration at the desired linearization point



$$\begin{aligned} G_1(\theta_{10}, \theta_{20}) &= \tau_{10}, \\ G_2(\theta_{10}, \theta_{20}) &= \tau_{20}, \end{aligned} \tag{2.128}$$

This shows that one can choose any set of desired angles $(\theta_{10}, \theta_{20})$ and calculate the appropriate torques from (2.128). If the angles shown on Fig. 2.21 are selected, the stationary torques are found from (2.128) and (2.119),

$$\begin{aligned} \tau_{10} &= G_{10} = 26.124\,\text{N}, \\ \tau_{20} &= G_{20} = 9.476\,\text{N}. \end{aligned}$$

For the matrix **M** one has

$$\mathbf{M} = \begin{bmatrix} 1.84 + 1.20\cos\theta_2 & 0.55 + 0.6\cos\theta_2 \\ 0.55 + 0.6\cos\theta_2 & 0.55 \end{bmatrix},$$

and for **H,**

$$\mathbf{H} = \frac{1}{det(\mathbf{M})} \begin{bmatrix} 0.55 & -(0.55 + 0.6\cos\theta_2) \\ -(0.55 + 0.6\cos\theta_2) & 1.84 + 1.20\cos\theta_2 \end{bmatrix},$$

where

$$det(\mathbf{M}) = -0.36\cos^2\theta_2 + 0.7095.$$

Furthermore from Eqs. (2.118) and (2.119) one obtains

$$\begin{aligned} K_1(\theta_2, \dot\theta_1, \dot\theta_2) &= -0.6\sin(\theta_2) \cdot \dot\theta_2(2\dot\theta_1 + \dot\theta_2) \\ K_2(\theta_2, \dot\theta_1, \dot\theta_2) &= 0.6\sin(\theta_2) \cdot \dot\theta_1^2 \end{aligned}$$

and

$$G_1(\theta_1, \theta_2) = 23.544\cos\theta_1 + 9.81\cos(\theta_1 + \theta_2),$$
$$G_2(\theta_1, \theta_2) = 9.81\cos(\theta_1 + \theta_2).$$

The stationary values for **M** and **H** are

$$\mathbf{M}_0 = \begin{bmatrix} 2.8792 & 1.0695 \\ 1.0695 & 0.55 \end{bmatrix}, \quad \mathbf{H}_0 = \begin{bmatrix} 1.2514 & -2.4337 \\ -2.4337 & 6.5512 \end{bmatrix},$$

which means that the entries of the matrix **B** of the linearized system have also been found (see Eq. (2.125)),

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ 1.2514 & -2.4337 \\ 0 & 0 \\ -2.4337 & 6.5512 \end{bmatrix}.$$

The remaining step in the linearization, i.e., determining **A** from (2.124), is not difficult but quite laborious. The result is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 17.832 & 0 & -3.0024 & 0 \\ 0 & 0 & 0 & 1 \\ -30.063 & 0 & 10.456 & 0 \end{bmatrix}.$$

The eigenvalues of the matrix **A** can be found to be

$$\lambda_{\mathbf{A}} = \begin{cases} \pm 4.933 \\ \pm 1.988 \end{cases}.$$

As will also be seen in Chap. 3, the eigenvalues contain important information about stability of the system. ❒

## 2.5 Discrete Time Models

In the previous sections models have been dealt with in the case where all the variables were functions of the continuous time $t$. This is quite natural since the physical phenomena which it is desired to model usually are of continuous nature. However, it is often practical to reformulate the continuous time models to discrete time models, i.e. models in which the variables are functions of discrete time.

Discrete time models are usually the prerequisite for designing discrete time controllers. Such controllers are of increasing importance because the digital computer has become more and more common as the hardware basis for the implementation of controllers. And, in contrast to natural physical phenomena, a computer works in discrete time. In this section the discrete time state space model will only be presented. The further treatment and the way discrete models emerge during the analysis, will be left to later chapters and appendix D.

In principle the discrete time model (DTM) looks very much like the continuous time counterpart (CTM) and the time argument $t$ is replaced with the discrete argument, $kT$, where $T$ is the time period between the instants for which the model is valid and $k$ is the current instaneous sample number, counted from some initial time, $k_0$. $T$ is usually called the *sampling period or interval.*

Whereas the CTM is a vector differential equation, the DTM is a vector *difference equation*. The equations can be written in different ways. The most correct notation for the system description is

$$\begin{aligned}\mathbf{x}((k+1)T) &= \mathbf{F}\mathbf{x}(kT) + \mathbf{G}\mathbf{u}(kT),\\ \mathbf{y}(kT) &= \mathbf{C}\mathbf{x}(kT) + \mathbf{D}\mathbf{u}(kT).\end{aligned} \tag{2.129}$$

The form in Eq. (2.129) is a bit awkward to work with and it is usually simplified to:

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k),\\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k),\end{aligned} \tag{2.130}$$

or even to:

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{F}\mathbf{x}_k + \mathbf{G}\mathbf{u}_k,\\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k.\end{aligned} \tag{2.131}$$

The form of the equations in (2.130) will be used throughout this book.

The notation above indicates that the matrices have constant entries, which means that the system is time invariant or, as it is sometimes called, *shift invariant* or *step invariant*. In the time varying case the equations must be written

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{F}(k)\mathbf{x}(k) + \mathbf{G}(k)\mathbf{u}(k),\\ \mathbf{y}(k) &= \mathbf{C}(k)\mathbf{x}(k) + \mathbf{D}(k)\mathbf{u}(k).\end{aligned} \tag{2.132}$$

Note that the matrices in the state equation are denoted **F** and **G** rather than **A** and **B** as in the CTM. If the DTM is a discretized version of the CTM (so that the CTM and the DTM describe the *same system*), the matrices will be

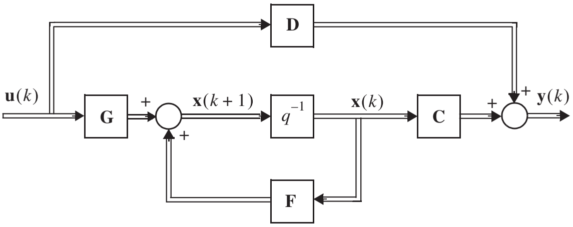


**Fig. 2.22** Block diagram of a general linear discrete time state space model

different, whereas the matrices **C** and **D** in the output equation will be the same. See Sect. 3.3.

A block diagram for the discrete time model looks very much like Fig. 2.3, but the integrator block is replaced by the *time delay* or *backward shift operator* $q^{-1}$. This operator is defined by the expression,

$$\mathbf{x}(k-1) = q^{-1}\mathbf{x}(k). \tag{2.133}$$

Similarly the *forward shift operator, q*, is defined by

$$\mathbf{x}(k+1) = q\mathbf{x}(k). \tag{2.134}$$

With the notation of (2.130) the general state space description for the nonlinear discrete time system will be

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{f}(\mathbf{x}(k), \mathbf{u}(k), k),\\ \mathbf{y}(k) &= \mathbf{g}(\mathbf{x}(k), \mathbf{u}(k), k).\end{aligned} \tag{2.135}$$

## 2.6 Summary

In this chapter it has been seen how continuous time state space models can be derived based on appropriate physical laws governing a system. It has also been seen how state variables can be selected directly from the equations or from a block diagram of the system. The latter possibility is usually the most straightforward method.

Since most system models turn out to be nonlinear and since most of the analysis and design tools are based on linear models, a very important linearization technique has been introduced. Once the general state space model is formulated and the possible stationary states have been found, it is quite easy to derive a linear model which will describes the system properties in the vicinity of the selected stationary state.

It was shown how one can set up state space models of systems with certain special forms and finally discrete time state space models were introduced. These are very important in the many cases where one has to design and implement discrete time controllers with the intention using a computer to perform the desired control tasks.

## 2.7 Problems

### *Problem 2.1*

Figure 2.23 shows a so-called inverted pendulum. If the driving force on the cart, $u$, is the input to the system and the cart position $p$, the output, a model of the system can be derived using Newton's second law. The model can for instance be formulated as two second order differential equations in the position and the pendulum angle $\theta$.

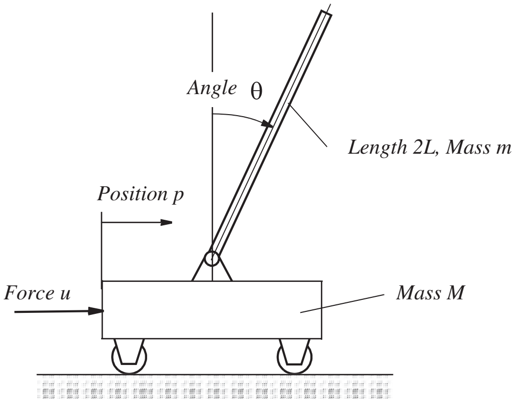


**Fig. 2.23** Inverted pendulum

The equations are

$$\ddot{p} = \frac{\dot{\theta}^2 mL\sin\theta - \frac{3}{4}mg\sin\theta\cos\theta + u}{M + m - \frac{3}{4}m\cos^2\theta},$$
$$\ddot{\theta} = \frac{\frac{3g}{4L}(M+m)\sin\theta - \frac{3m}{4}\dot{\theta}^2\sin\theta\cos\theta - \frac{3}{4L}u\cos\theta}{M + m - \frac{3}{4}m\cos^2\theta}, \tag{2.136}$$

where $g$ is the acceleration due to gravity.

a. Choose an appropriate state vector and derive the nonlinear state equation set,

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, u),$$
$$y = g(\mathbf{x}, u).$$

b. Linearize the model by assuming small angles and velocities, and consequently,

$$\sin\theta \cong \theta, \quad \cos\theta \cong 1, \quad \dot{\theta}^2 \cong 0,$$

and derive the linear state model,

$$\Delta\dot{\mathbf{x}} = \mathbf{A}\Delta\mathbf{x} + \mathbf{B}\Delta u,$$
$$\Delta y = \mathbf{C}\Delta\mathbf{x} + \mathbf{D}\Delta u.$$

c. Show that the linear model can be divided into two models. One model describing the angle (the θ-system) and one model describing the position (the $p$-system) and show that the θ-system is completely decoupled from the $p$-system.

***Problem 2.2***

a. Find the transfer function, $G(s) =$ y(s)/u(s), for the system in Fig. 2.9.
b. Derive companion form 1 and companion form 2 for the system and draw the block diagrams for these state space models.

***Problem 2.3***

a. Linearize the inverted pendulum model from Problem 2.1 a. by using the method from Sect. 2.4 and assume the following stationary state,

$$\mathbf{x}_0 = \begin{bmatrix} \theta_0 & \dot{\theta}_0 & p_0 & \dot{p}_0 \end{bmatrix}^T = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}^T,$$

and use the parameter values,

$$M = 2\,\text{kg}, \quad m = 1\,\text{kg}, \quad L = 0.5\,\text{m}, \quad g \cong 10\frac{m}{\text{sec}^2}.$$

b. Compare the result with the result from problem 2.1 b.

***Problem 2.4***

A simplified car suspension system is shown on Fig. 2.24. The components in the figure represent a quarter of a car (or half a motorcycle). $M$ is the body mass, $m$ is the so-called unsprung mass (wheel, tire, brake etc.), $k_t$ and $k_s$ are the spring constants of the tire and the spring respectively and $b$ is the damping coefficient of the shock absorber which is assumed to be a linear damper (the damping force is proportional to the velocity difference between the two members of the

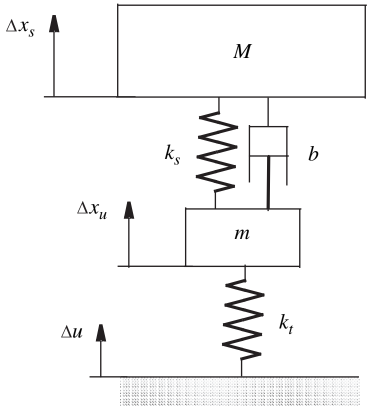


**Fig. 2.24** Car suspension system

damper). $\Delta u$ is the position of the road. Only motion in the vertical direction will be considered.

A model is to be set up which is valid around a stationary state where the springs are in tension and shortened to their stationary length under the influence of gravity.

a. Derive a model of the system using Newton's second law on the two masses
b. Draw a block diagram and choose a set of state variables
c. Formulate a state space model where $\Delta u$ is the input and where the output is the four dimensional vector,

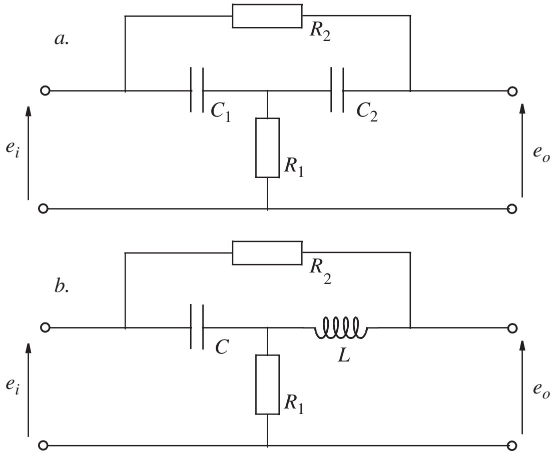
$$\Delta \mathbf{y} = \begin{bmatrix} \Delta x_s \\ \Delta x_u \\ \Delta f_t \\ \Delta \ddot{x}_s \end{bmatrix},$$

where $\Delta f_t$ is the tire force.

***Problem 2.5***

The electrical circuits on Fig. 2.25 are called bridged-T-filters.

**Fig. 2.25** Common electrical circuits (Bridged-T-filters)



a. Use Ohm's and Kirchhoff's laws on the circuits to derive a linear model of the filters.
b. Draw block diagrams of the circuits and choose in each case a set of state variables.
c. Derive state space models for the two filters.

***Problem 2.6***

Figure 2.26 shows a hydraulic servo cylinder. The piston is positioned symmetrically in the cylinder and the oil volume of the two cylinder chambers are

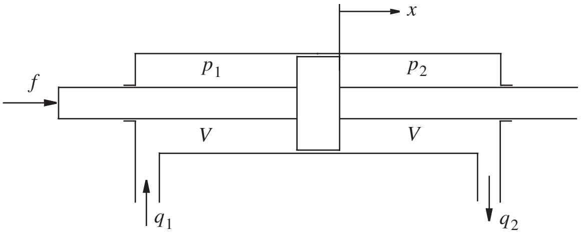


**Fig. 2.26** Hydraulic servo cylinder

therefore of equal size $V$. The ring shaped area of the piston is $A_c$. The mass of piston and piston rod is $M$ and the rod is loaded by an external force $f$.

The servo valve delivering the volume flows $q_1$ and $q_2$ is symmetric and dynamically ideal (massless and frictionless) and the relation between the driving input voltage, $u$, and the flows is assumed to be

$$q_1 = q_2 = ku.$$

The volume flow into a cylinder chamber can be divided into three parts,

$$q = q_{displacement} + q_{compression} + q_{leakage}.$$

For the left chamber one has,

$$q_1 = A_c\dot{x} + \frac{V}{\beta}\dot{p}_1 + C_l(p_1 - p_2),$$

where $\beta$ is the stiffness coefficient for the oil (the so-called bulk modulus) and $C_l$ is a laminar leakage coefficient. Note that the compression flow is proportional to the time derivative of the pressure in the cylinder chamber.

a. Derive a model for the system with input $u$, disturbance $v = f$ and output, $y = x$.
b. Draw a block diagram for the model and set up a linear state model of the form:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u + \mathbf{B}_v v,$$
$$y = \mathbf{C}\mathbf{x} + \mathbf{D}u.$$

***Problem 2.7***

The following discrete time transfer function is given:

$$H(z) = \frac{3z^3 + 2z^2 + z + 5}{4z^3 + 4z^2 + 2z + 8}.$$

Set up state space models in companion form 1 and companion form 2.

***Problem 2.8***

A nonlinear system is given by the equations,

$$\dot{p} = -p + \alpha\sqrt{u},$$
$$m\ddot{y} = pu - y^2 - \dot{y},$$

where $u$ is the input and $y$ is the output.

a. Draw a block diagram of the system and choose a state vector.
b. Derive a nonlinear state space model and find the stationary states of the model.
c. Linearize the system and derive the linear state space model.

***Problem 2.9***

A system with 3 interconnected tanks is shown on Fig. 2.27. The tanks have the cross sectional areas $A_1$, $A_2$ and $A_3$. The input flows are proportional to the input voltages,

$$q_a = ku_1,$$
$$q_b = ku_2.$$

The flows follow the square root law and since the pressures in the bottoms of the tanks are proportional to the level, one has:

$$q_1 = c_1\sqrt{x_1 - x_2},$$
$$q_2 = c_2\sqrt{x_3 - x_2},$$
$$q_0 = c_0\sqrt{x_2}.$$

a. Derive a set of equations describing the system and formulate a nonlinear state model with the two inputs $u_1$ and $u_2$, and the output $y = q_0$.
b. Linearize the model and derive the linear state space model

$$\dot{\Delta\mathbf{x}} = \mathbf{A}\Delta\mathbf{x} + \mathbf{B}\Delta\mathbf{u},$$
$$\Delta y = \mathbf{C}\Delta\mathbf{x}.$$

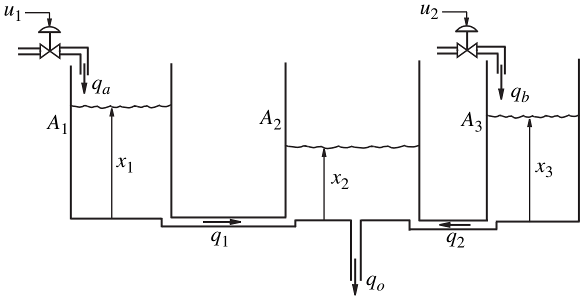
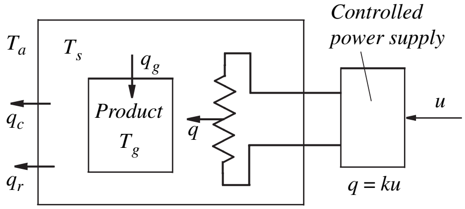


**Fig. 2.27** Tank system

### *Problem 2.10*

A high temperature oven with a product to be heat treated is shown on Figure 2.28. The outer surface temperature $T_s$ of the oven is the same as the temperature in the inner oven space. The ambient temperature is $T_a$ and the product temperature is $T_g$. All the temperatures are assumed to be uniform. The variables denoted $q$ are the heat powers exchanged between the different parts of the system.

**Fig. 2.28** High temperature oven production system



The oven surface temperature is so high that the heat loss to the surroundings is not only due to convection but also to radiation. Assume that the convection power is proportional to the temperature difference, i. e.,

$$q_c = k_c(T_s - T_a) \,\text{and}\, q_g = k_g(T_s - T_g),$$

whereas the radiation heat loss follows the Stefan-Boltzmann law,

$$q_r = k_r(T_s^4 - T_a^4).$$

Note that the temperatures here are *absolute* temperatures. The output is the temperature $T_g$. The total heat capacities of the oven air space and the product are $C_s$ and$C_g$.

a. Derive a nonlinear model of the system and formulate the state space model in the form of Eq. (2.56).
b. Derive a set of equations for determining the stationary states.
c. Linearize the model and derive the matrices **A**, **B**, $\mathbf{B}_v$ and **C** for the system.

# Chapter 3
# Analysis of State Space Models

**Abstract** In this chapter an overview of the properties of the state space models will be given. A basis for the investigation of these properties is the solution of the state equation given appropriate boundary conditions. The important notions of stability, controllability and observability will be introduced and the similarity transformation discussed. This makes possible the construction of state space models with a number of useful properties.

## 3.1 Solution of the Linear State Equation

In this section first the solution of the general time varying state equation will be obtained and then specialized for the time invariant case.

### *3.1.1 The Time Varying System*

A reasonable starting point for the treatment in this chapter is the *homogeneous* (or unforced) state equation,

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t). \tag{3.1}$$

Prior to the search for solutions of this matrix differential equation, it is important to know whether a solution *exists* and whether the solution found is *unique*, given the initial state. $\mathbf{x}(t_0)$ The answers to these questions are not trivial and the conditions required to ensure that Eq. (3.1) has a unique solution will not be derived here. It will only be stated that if $\mathbf{A}(t)$'s elements are piecewise continuous functions of $t$, a unique solution does exist. As a matter of fact, this condition is more than what is required. It is sufficient but not necessary. However, piecewise continuity is not a very restrictive requirement and the vast majority of practical applications will be covered by this assumption. For a more thorough treatment of these questions the reader is referred to Kailath (1980) or Brauer and Nohel (1969).

E. Hendricks et al., *Linear Systems Control*, DOI: 10.1007/978-3-540-78486-9_3,

© Springer-Verlag Berlin Heidelberg 2008

For the system (3.1) one can choose $n$ linearly independent initial condition vectors $\mathbf{x}(t_0)$. An obvious choice would be:

$$\begin{aligned}\mathbf{x}_1(t_0) &= [1\ 0\ 0\ \dots\ 0]^T = \mathbf{e}_1^T,\\ \mathbf{x}_2(t_0) &= [0\ 1\ 0\ \dots\ 0]^T = \mathbf{e}_2^T,\\ &\vdots\\ \mathbf{x}_n(t_0) &= [0\ 0\ 0\ \dots\ 1]^T = \mathbf{e}_n^T.\end{aligned} \tag{3.2}$$

The solutions to (3.1) corresponding to these initial conditions may be arranged in a square matrix,

$$\mathbf{U}(t) = [\mathbf{x}_1(t)\ \mathbf{x}_2(t)\ \dots \mathbf{x}_n(t)]. \tag{3.3}$$

Note that $\mathbf{U}(t_0) = \mathbf{I}$, the identity matrix.

Assume now that the solutions in Eq. (3.3) are linearly dependent. Then, by definition, there exists a nonzero constant vector $\mathbf{z}$ such that

$$[\mathbf{x}_1(t)\ \mathbf{x}_2(t)\ \dots\ \mathbf{x}_n(t)]\mathbf{z} = \mathbf{0}, \tag{3.4}$$

for all $t$. In particular for $t = t_0$,

$$[\mathbf{x}_1(t_0)\ \mathbf{x}_2(t_0)\ \dots\ \mathbf{x}_n(t_0)]\mathbf{z} = [\mathbf{e}_1\ \mathbf{e}_2\ \dots\ \mathbf{e}_n]\mathbf{z} = \mathbf{0}. \tag{3.5}$$

But (3.5) contradicts the fact that all the $\mathbf{e}_i$ vectors are linearly independent. The conclusion is that all the solutions in (3.3) are linearly independent and consequently that $\mathbf{U}(t)$ is regular for all $t$, or, in other words: $\mathbf{U}^{-1}(t)$ exists for all $t$. With the definition of $\mathbf{U}(t)$ it can be seen that:

$$\dot{\mathbf{U}}(t) = \mathbf{A}(t)\mathbf{U}(t). \tag{3.6}$$

Any solution $\mathbf{U}(t)$ to Eq. (3.3) is called a *fundamental matrix* of (3.1).

Now it is claimed that the solution to (3.1) for an *arbitrary* initial condition, $\mathbf{x}(t_0)$, can be written:

$$\mathbf{x}(t) = \mathbf{U}(t)\mathbf{x}(t_0). \tag{3.7}$$

It is obvious that (3.7) holds for $t = t_0$. Differentiating (3.7) with respect to time gives:

$$\dot{\mathbf{x}}(t) = \dot{\mathbf{U}}(t)\mathbf{x}(t_0) = \mathbf{A}(t)\mathbf{U}(t)\mathbf{x}(t_0) = \mathbf{A}(t)\mathbf{x}(t),$$

which shows that (3.7) does in fact satisfy the state equation (3.1).

Now augment equation (3.1) to obtain the nonhomogeneous state equation,

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t). \tag{3.8}$$

From $\mathbf{U}(t)\mathbf{U}^{-1}(t) = \mathbf{I}$ it follows that

$$\frac{d}{dt}(\mathbf{U}(t)\mathbf{U}^{-1}(t)) = \mathbf{0} \tag{3.9}$$

or (omitting the time argument)

$$\frac{d\mathbf{U}}{dt}\mathbf{U}^{-1} + \mathbf{U}\frac{d\mathbf{U}^{-1}}{dt} = \mathbf{0} \tag{3.10}$$

or, solving for the time derivative of the inverse,

$$\frac{d\mathbf{U}^{-1}}{dt} = -\mathbf{U}^{-1}\frac{d\mathbf{U}}{dt}\mathbf{U}^{-1} = -\mathbf{U}^{-1}\mathbf{A}. \tag{3.11}$$

Postmultiplying the last expression with $\mathbf{x}$ yields

$$\frac{d\mathbf{U}^{-1}}{dt}\mathbf{x} = -\mathbf{U}^{-1}\mathbf{A}\mathbf{x}. \tag{3.12}$$

Premultiplying (3.8) with $\mathbf{U}^{-1}$ gives $\mathbf{U}^{-1}\dot{\mathbf{x}} = \mathbf{U}^{-1}\mathbf{A}\mathbf{x} + \mathbf{U}^{-1}\mathbf{B}\mathbf{u}$. Inserting Eq. (3.12) into this expression leads to

$$\mathbf{U}^{-1}\dot{\mathbf{x}} + \frac{d\mathbf{U}^{-1}}{dt}\mathbf{x} = \mathbf{U}^{-1}\mathbf{B}\mathbf{u} \tag{3.13}$$

which can also be written

$$\frac{d}{dt}(\mathbf{U}^{-1}\mathbf{x}) = \mathbf{U}^{-1}\mathbf{B}\mathbf{u}. \tag{3.14}$$

Integrating from $t_0$ to $t$ on both sides of the equal sign yields

$$\int_{t_0}^{t} \frac{d}{d\tau}(\mathbf{U}^{-1}(\tau)\mathbf{x}(\tau))d\tau = \int_{t_0}^{t} \mathbf{U}^{-1}(\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau \tag{3.15}$$

or

$$\mathbf{U}^{-1}(t)\mathbf{x}(t) - \mathbf{U}^{-1}(t_0)\mathbf{x}(t_0) = \int_{t_0}^{t} \mathbf{U}^{-1}(\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau. \tag{3.16}$$

Ordering the terms and premultiplying by **U**(t) gives the solution,

$$\mathbf{x}(t) = \mathbf{U}(t)\mathbf{U}^{-1}(t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{U}(t)\mathbf{U}^{-1}(\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau. \qquad (3.17)$$

Now the *state transition matrix* will be introduced:

$$\phi(t,\tau) = \mathbf{U}(t)\mathbf{U}^{-1}(\tau). \qquad (3.18)$$

The quadratic fundamental matrix $\mathbf{U}(t)$ depends on the initial state, according to the definition in the beginning of this section. If two different fundamental matrices $\mathbf{U}_1(t)$ and $\mathbf{U}_2(t)$ are selected, it is known that all columns in the two matrices are linearly independent and therefore they can act as basis vectors in the n-dimensional vector space. From linear algebra it is known that a constant nonsingular matrix **P** exists such that $\mathbf{U}_2(t)\mathbf{P} = \mathbf{U}_1(t)$. From (3.18) it is clear that

$$\phi(t,\tau) = \mathbf{U}_1(t)\mathbf{U}_1^{-1}(\tau) = \mathbf{U}_2(t)\mathbf{P}\mathbf{P}^{-1}\mathbf{U}_2^{-1}(\tau) = \mathbf{U}_2(t)\mathbf{U}_2^{-1}(\tau). \qquad (3.19)$$

This shows, that the state transition matrix is unique and independent of the specific choice of $\mathbf{U}(t)$.

With the state transition matrix available, the solution (3.17) to the non-homogeneous state equation (3.8) can be written in its final form, noting that $\mathbf{U}(t)\mathbf{U}^{-1}(t_0) = \phi(t,t_0)$ :

$$\mathbf{x}(t) = \phi(t,t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \phi(t,\tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau. \qquad (3.20)$$

The solution has two terms. The first term $\phi(t,t_0)\mathbf{x}(t_0)$ is the solution for $\mathbf{u}(t) = \mathbf{0}$ and it is therefore called the *zero input solution*. The integral term is called the *zero state solution* because it is the solution if the initial state is a zero vector. In other words, the solution is a *superposition* of the effects of the initial conditions $\mathbf{x}(t_0)$ and the effects due to the input $\mathbf{u}(t)$.

From Eqs. (3.6) and (3.18) it can be seen that

$$\frac{\partial}{\partial t}\phi(t,t_0) = \dot{\mathbf{U}}(t)\mathbf{U}^{-1}(t_0) = \mathbf{A}(t)\mathbf{U}(t)\mathbf{U}^{-1}(t_0) = \mathbf{A}(t)\phi(t,t_0). \qquad (3.21)$$

The state transition matrix has some special properties. From the definition in Eq. (3.18) it is obvious that

$$\phi(t_0,t_0) = \mathbf{I}. \qquad (3.22)$$

It is also straightforward to conclude that

$$\phi^{-1}(t, t_0) = \mathbf{U}(t_0)\mathbf{U}^{-1}(t) = \phi(t_0, t). \tag{3.23}$$

Thus, the inverse of $\phi(t, t_0)$ is very easy to find. It is just a matter of interchanging its two arguments. It is equally easy to see that

$$\begin{aligned}\phi(t_2, t_0) = \mathbf{U}(t_2)\mathbf{U}^{-1}(t_0) &= \mathbf{U}(t_2)\mathbf{U}^{-1}(t_1)\mathbf{U}(t_1)\mathbf{U}^{-1}(t_0) \\ &= \phi(t_2, t_1)\phi(t_1, t_0).\end{aligned} \tag{3.24}$$

It should be pointed out that although the solution, Eq. (3.20), to the time varying state equation (3.8) looks simple, it is usually not applicable for practical calculations. The problem is that $\phi(t, t_0)$ can in general not be found analytically given its defining differential equation (3.21) with the boundary conditions (3.22). Except for special cases, one has to be satisfied with numerical solutions, obtained using a computer.

The linear system's output equation is

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t). \tag{3.25}$$

With the solution (3.20) the system's response can thus be written as

$$\begin{aligned}\mathbf{y}(t) &= \mathbf{C}(t)\phi(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{C}(t)\phi(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau + \mathbf{D}(t)\mathbf{u}(t) \\ &= \mathbf{C}(t)\phi(t, t_0)\mathbf{x}(t_0) + \int_{t_0}^{t} [\mathbf{C}(t)\phi(t, \tau)\mathbf{B}(\tau) + \mathbf{D}(\tau)\delta(t-\tau)]\mathbf{u}(\tau)d\tau.\end{aligned} \tag{3.26}$$

Defining

$$\mathbf{g}(t, \tau) = \mathbf{C}(t)\phi(t, \tau)\mathbf{B}(\tau) + \mathbf{D}(\tau)\delta(t-\tau), \tag{3.27}$$

for $\mathbf{x}(t_0) = \mathbf{0}$ Eq. (3.26) can be written,

$$\mathbf{y}(t) = \int_{t_0}^{t} \mathbf{g}(t, \tau)\mathbf{u}(\tau)d\tau. \tag{3.28}$$

$\mathbf{g}(t, \tau)$ is called the *unit impulse response matrix* or just the *unit impulse response* of the system.

If the j'th input variable is a unit impulse at time $t = t_p$ and all other entries of $\mathbf{u}(t)$ are zero, i.e.,

$$\mathbf{u}(t) = \begin{bmatrix}0 & \dots & 0 & \delta(t-t_p) & 0 & \dots & 0\end{bmatrix}^T, \; t_0 < t_p < t,$$

equation (3.28) then gives

$$\mathbf{y}(t) = \int_{t_0}^{t} \begin{bmatrix} g_{1j}(t,\tau)\delta(\tau - t_p) \\ g_{2j}(t,\tau)\delta(\tau - t_p) \\ \vdots \\ g_{rj}(t,\tau)\delta(\tau - t_p) \end{bmatrix} d\tau = \begin{bmatrix} g_{1j}(t,t_p) \\ g_{2j}(t,t_p) \\ \vdots \\ g_{rj}(t,t_p) \end{bmatrix}. \tag{3.29}$$

For the i'th output variable it is found that

$$y_i(t) = g_{ij}(t,t_p). \tag{3.30}$$

### 3.1.2 The Time Invariant System

The solution given in Eq. (3.20) is of course also valid for time invariant systems and it can be simplified and made much more useful in such cases. For linear, time invariant systems (*LTI systems*) the use of the Laplace transform is possible. Start again with the homogeneous equation,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t). \tag{3.31}$$

Suppose now that the initial state is $\mathbf{x}(t_0) = \mathbf{x}(0) = \mathbf{x}_0$. In the time invariant case it can be assumed that $t_0 = 0$ without loss of generality. Laplace transforming the time dependent vectors in (3.31) element by element,

$$s\mathbf{X}(s) - \mathbf{x}_0 = \mathbf{A}\mathbf{X}(s), \tag{3.32}$$

where the Laplace transform is denoted by $\mathbf{X}(s) = \mathscr{L}\{\mathbf{x}(t)\}$.

Rearranging the terms in (3.32) gives

$$(s\mathbf{I} - \mathbf{A})\mathbf{X}(s) = \mathbf{x}_0 \quad \text{or} \quad \mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}_0. \tag{3.33}$$

Now, defining

$$\Psi(s) = (s\mathbf{I} - \mathbf{A})^{-1} \quad \text{and} \quad \Psi(t) = \mathscr{L}^{-1}\{\Psi(s)\}, \tag{3.34}$$

the solution of the state equation can be written as

$$\mathbf{X}(s) = \Psi(s)\mathbf{x}_0, \tag{3.35}$$

or using the inverse Laplace transform,

$$\mathbf{x}(t) = \psi(t)\mathbf{x}_0. \tag{3.36}$$

If the inhomogeneous case is considered,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{3.37}$$

a result similar to that in (3.33) can be found,

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}_0 + (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(s), \tag{3.38}$$

or applying Eq. (3.34),

$$\mathbf{X}(s) = \Psi(s)\mathbf{x}_0 + \Psi(s)\mathbf{B}\mathbf{U}(s). \tag{3.39}$$

The second term is the product of two Laplace transforms and the inverse transform therefore takes the form of a convolution integral:

$$\mathbf{x}(t) = \psi(t)\mathbf{x}_0 + \int_0^t \psi(t-\tau)\mathbf{B}\mathbf{u}(\tau)d\tau. \tag{3.40}$$

This solution has a great deal of resemblance to Eq. (3.20). But in this case it is possible to determine an analytical solution using (3.34). This can be quite laborious, however, since this requires that the inverse of the matrix $(s\mathbf{I} - \mathbf{A})$ is found symbolically.

For numerical calculations it is usually preferable to use another approach based on the well known solution to a scalar, linear, time invariant, first order differential equation

$$\dot{x}(t) = ax(t) \quad \text{with} \quad x(0) = x_0. \tag{3.41}$$

The solution of this equation is

$$x(t) = e^{at}x_0. \tag{3.42}$$

The series expansion of the exponential function is

$$e^{at} = 1 + at + \frac{a^2t^2}{2} + \ \ldots \ = \sum_{k=0}^{\infty} \frac{a^k t^k}{k!}. \tag{3.43}$$

Now *define* a square matrix by a series expansion similar to that in Eq. (3.43),

$$e^{\mathbf{A}t} = \mathbf{I} + \mathbf{A}t + \frac{\mathbf{A}^2 t^2}{2} + \ \ldots \ = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k t^k}{k!}. \tag{3.44}$$

The two series equations, (3.43) and (3.44), converge for all finite values of $t$. $e^{\mathbf{A}t}$ is called the *matrix exponential.*

Briefly, some of the properties of this matrix are:

1. From the definition (3.44) it follows that

$$\begin{aligned}\frac{d}{dt}(e^{\mathbf{A}t}) &= \mathbf{A} + \mathbf{A}^2 t + \frac{1}{2}\mathbf{A}^3 t^2 + \ \ldots \ + \frac{1}{(k-1)!}\mathbf{A}^k t^{k-1} + \ \ldots \\ &= \mathbf{A}\left(\mathbf{I} + \mathbf{A}t + \frac{1}{2}\mathbf{A}^2 t^2 + \ \ldots\right) = \left(\mathbf{I} + \mathbf{A}t + \frac{1}{2}\mathbf{A}^2 t^2 + \ \ldots\right)\mathbf{A}\end{aligned}$$

or

$$\frac{d}{dt}(e^{\mathbf{A}t}) = \mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A}. \tag{3.45}$$

2. $$\begin{aligned}e^{\mathbf{A}t} \cdot e^{\mathbf{A}s} &= \sum_{k=0}^{\infty} \frac{\mathbf{A}^k t^k}{k!} \cdot \sum_{k=0}^{\infty} \frac{\mathbf{A}^k s^k}{k!} \\ &= \left(\mathbf{I} + \mathbf{A}t + \frac{\mathbf{A}^2 t^2}{2} + \frac{\mathbf{A}^3 t^3}{6} + \ \ldots\right) \cdot \left(\mathbf{I} + \mathbf{A}s + \frac{\mathbf{A}^2 s^2}{2} + \frac{\mathbf{A}^3 s^3}{6} + \ \ldots\right) \\ &= \mathbf{I} + \mathbf{A}(t+s) + \frac{1}{2}\mathbf{A}^2(t^2 + 2ts + s^2) + \frac{1}{6}\mathbf{A}^3(t^3 + 3ts^2 + 3t^2 s + s^3) + \ \ldots \\ &= \mathbf{I} + \mathbf{A}(t+s) + \frac{1}{2}\mathbf{A}^2(t+s)^2 + \frac{1}{6}\mathbf{A}^3(t+s)^3 + \ \ldots = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k (t+s)^k}{k!}\end{aligned}$$

or

$$e^{\mathbf{A}t} \cdot e^{\mathbf{A}s} = e^{\mathbf{A}(t+s)}. \tag{3.46}$$

If one lets $s = -t$, it is found that

$$e^{\mathbf{A}t} \cdot e^{-\mathbf{A}t} = e^{\mathbf{A}(t-t)} = e^{\mathbf{A}\cdot 0} = \mathbf{I} \ \text{ or } \ (e^{\mathbf{A}t})^{-1} = e^{-\mathbf{A}t},$$

which means that the matrix exponential is nonsingular (because $e^{-\mathbf{A}t}$ is convergent for all finite $t$). It is inverted by simply changing the sign of the exponent.

3. Note that it can be shown that $e^{(\mathbf{A}+\mathbf{B})t}$ is *not in general* equal to $e^{\mathbf{A}t} \cdot e^{\mathbf{B}t}$ This is *only* the case if **A** and **B** commute, i.e., if $\mathbf{AB} = \mathbf{BA}$.

If comparing Eq. (3.45) with (3.6), one can see that $e^{\mathbf{A}t}$ qualifies as a fundamental matrix. The state transition matrix (3.18) will then be

$$\phi(t, \tau) = e^{\mathbf{A}t} e^{-\mathbf{A}\tau} = e^{\mathbf{A}(t-\tau)}. \tag{3.47}$$

Note, that the state transition matrix is no longer a function of the two *independent* arguments $t$ and $\tau$ as in (3.20), but only a function of the difference $t - \tau$.

For $\tau = t_0 = 0$ one finds that

$$\phi(t, 0) = \phi(t) = e^{\mathbf{A}t}. \tag{3.48}$$

Now it is seen that the solution, Eq. (3.20), of the general linear state equation in this case reduces to:

$$\mathbf{x}(t) = \phi(t)\mathbf{x}_0 + \int_0^t \phi(t - \tau)\mathbf{Bu}(\tau)d\tau. \tag{3.49}$$

Comparing (3.40) and (3.49) it is also seen that the matrix $\Psi(t)$, which was introduced in (3.34), is exactly the same as $\phi(t)$. Therefore one finally has the relation,

$$\phi(t) = \psi(t) = \mathscr{L}^{-1}\{(s\mathbf{I} - \mathbf{A})^{-1}\} = e^{\mathbf{A}t} \tag{3.50}$$

or

$$\Phi(s) = \Psi(s) = (s\mathbf{I} - \mathbf{A})^{-1}. \tag{3.51}$$

The matrix $\Phi(s)$ is called the *resolvent matrix*.

The right hand relation of Eq. (3.50) shows a closed form alternative to the series expansion for $e^{\mathbf{A}t}$. With the matrix exponential, the solution (3.49) can be written,

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{Bu}(\tau)d\tau. \tag{3.52}$$

With the output equation,

$$\mathbf{y}(t) = \mathbf{Cx}(t) + \mathbf{Du}(t), \tag{3.53}$$

one obtains

$$\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}_0 + \mathbf{C}\int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau + \mathbf{D}\mathbf{u}(t). \tag{3.54}$$

The impulse response in Eq. (3.27) can be written

$$\mathbf{g}(t,\tau) = \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B} + \mathbf{D}\delta(t-\tau) = \mathbf{g}(t-\tau). \tag{3.55}$$

After a change of variables this becomes

$$\mathbf{g}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B} + \mathbf{D}\delta(t). \tag{3.56}$$

The resolvent matrix can be written,

$$\Phi(s) = (s\mathbf{I}-\mathbf{A})^{-1} = \frac{adj(s\mathbf{I}-\mathbf{A})}{det(s\mathbf{I}-\mathbf{A})}, \tag{3.57}$$

where *adj* and *det* denote the adjoint and the determinant of the matrix respectively.

The numerator of $\Phi(s)$ is an $n \times n$ matrix. The denominator is a polynomial and it is precisely the same polynomial which may be used for the determination of the *eigenvalues* of the matrix **A**,

$$P_{ch,\mathbf{A}} = det(\lambda\mathbf{I}-\mathbf{A}). \tag{3.58}$$

In this context the polynomial is called the *characteristic polynomial* of **A**.

$\Phi(s)$ can be decomposed in a way similar to the partial fraction decomposition of transfer functions for SISO systems but in this case the residuals will in general be matrices rather than scalars. If it is assumed that **A** has distinct as well as repeated eigenvalues, they can be for example,

$$\lambda_1, \lambda_2, \ \ldots, \ m \text{ times } \lambda_r, \ \ldots, \ \lambda_\sigma,$$

where the eigenvalue $\lambda_r$ is repeated $m$ times. Equation (3.57) can then be written

$$\Phi(s) = \frac{adj(s\mathbf{I}-\mathbf{A})}{(s-\lambda_1)(s-\lambda_2)\ldots(s-\lambda_r)^m\ldots(s-\lambda_\sigma)}. \tag{3.59}$$

The partial fraction decomposition turns out to be

$$\begin{aligned}\Phi(s) =& \frac{\mathbf{Z}_1}{s-\lambda_1} + \frac{\mathbf{Z}_2}{s-\lambda_2} + \ldots \\ &+ \frac{\mathbf{Z}_{r1}}{s-\lambda_r} + \frac{\mathbf{Z}_{r2}}{(s-\lambda_r)^2} + \ldots + \frac{\mathbf{Z}_{rm}}{(s-\lambda_r)^m} + \ldots + \frac{\mathbf{Z}_\sigma}{s-\lambda_\sigma},\end{aligned} \tag{3.60}$$

where the matrices $\mathbf{Z}_i$ have constants elements.

Inverse Laplace transformation of Eq. (3.60) gives the state transition matrix,

$$\begin{aligned}\phi(t) = e^{\mathbf{A}t} = \mathbf{Z}_1 e^{\lambda_1 t} + \mathbf{Z}_2 e^{\lambda_2 t} + \ldots \\ + \mathbf{Z}_{r1} e^{\lambda_r t} + \mathbf{Z}_{r2} t e^{\lambda_r t} + \ldots + \mathbf{Z}_{rm} t^{m-1} e^{\lambda_r t} + \ldots \mathbf{Z}_\sigma e^{\lambda_\sigma t}.\end{aligned} \tag{3.61}$$

The exponential function terms, $e^{\lambda_i t}$ and $t^l\, e^{\lambda_r t}$, are called the *natural modes* of the system.

Computation of the matrix exponential is usually based on some modified form of the series expansion (3.44). It turns out, however, that accurate computation of $e^{\mathbf{A}t}$ is not always simple. The expansion (3.44) converges quite slowly and in some cases the results are difficult to use. This problem is of numerical nature and it is beyond the scope of this text to pursue it in detail. The reader is referred to books like Golub and Van Loan (1989). It is also strongly recommended that the reader to use well-tested computer algorithms like those found in packages like MATLAB for actual computations.

There is a special case where calculation is simple: this is when $\mathbf{A}$ is a diagonal matrix. The diagonal elements are the eigenvalues of the matrix,

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 & 0 \\ 0 & \lambda_2 & \ldots & 0 & 0 \\ : & : & : & : & : \\ 0 & 0 & \ldots & \lambda_{n-1} & 0 \\ 0 & 0 & \ldots & 0 & \lambda_n \end{bmatrix} = \Lambda. \tag{3.62}$$

$\Lambda$ raised to the k'th power is simply

$$\Lambda^k = \begin{bmatrix} \lambda_1^k & 0 & \ldots & 0 & 0 \\ 0 & \lambda_2^k & \ldots & 0 & 0 \\ : & : & : & : & : \\ 0 & 0 & \ldots & \lambda_{n-1}^k & 0 \\ 0 & 0 & \ldots & 0 & \lambda_n^k \end{bmatrix} \tag{3.63}$$

and the series expansion (3.44) can be written

$$e^{\Lambda t} = \sum_{k=0}^{\infty} \frac{\Lambda^k t^k}{k!} = \begin{bmatrix} \sum\limits_{k=0}^{\infty} \frac{\lambda_1^k t^k}{k!} & 0 & \ldots & 0 & 0 \\ 0 & \sum\limits_{k=0}^{\infty} \frac{\lambda_2^k t^k}{k!} & \ldots & 0 & 0 \\ : & : & : & : & : \\ 0 & 0 & \ldots & \sum\limits_{k=0}^{\infty} \frac{\lambda_{n-1}^k t^k}{k!} & 0 \\ 0 & 0 & \ldots & 0 & \sum\limits_{k=0}^{\infty} \frac{\lambda_n^k t^k}{k!} \end{bmatrix}. \tag{3.64}$$

The diagonal elements are the series expansion of the scalar exponential function and consequently it is clear that

$$e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & 0 & \dots & 0 & 0 \\ 0 & e^{\lambda_2 t} & \dots & 0 & 0 \\ : & : & : & : & : \\ 0 & 0 & \dots & e^{\lambda_{n-1} t} & 0 \\ 0 & 0 & \dots & 0 & e^{\lambda_n t} \end{bmatrix}. \tag{3.65}$$

***Example 3.1.*** **Second Order LTI System**

A second Order LTI-System is given by the equations.

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -8 & -6 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u,$$

$$y = [2 \quad 0]\mathbf{x}.$$

The system is SISO and thus $u$ and $y$ are scalars.

The resolvent matrix is

$$\Phi(s) = (s\mathbf{I} - \mathbf{A})^{-1} = \left[\begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ -8 & -6 \end{bmatrix}\right]^{-1} = \begin{bmatrix} s & -1 \\ 8 & s+6 \end{bmatrix}^{-1} = \frac{\begin{bmatrix} s+6 & 1 \\ -8 & s \end{bmatrix}}{s(s+6)+8},$$

$$\Phi(s) = \begin{bmatrix} \frac{s+6}{(s+2)(s+4)} & \frac{1}{(s+2)(s+4)} \\ \frac{-8}{(s+2)(s+4)} & \frac{s}{(s+2)(s+4)} \end{bmatrix} = \begin{bmatrix} \frac{2}{s+2} - \frac{1}{s+4} & \frac{1/2}{s+2} - \frac{1/2}{s+4} \\ -\frac{4}{s+2} + \frac{4}{s+4} & -\frac{1}{s+2} + \frac{2}{s+4} \end{bmatrix}$$

$$= \frac{\begin{bmatrix} 2 & 1/2 \\ -4 & -1 \end{bmatrix}}{s+2} + \frac{\begin{bmatrix} -1 & -1/2 \\ 4 & 2 \end{bmatrix}}{s+4} = \frac{\mathbf{Z}_1}{s+2} + \frac{\mathbf{Z}_2}{s+4}.$$

All the terms in the resolvent matrix can be inverse Laplace transformed and the state transition matrix is readily found to be

$$\phi(t) = \mathscr{L}^{-1}\{\Phi(s)\} = e^{\mathbf{A}t} = \begin{bmatrix} 2e^{-2t} - e^{-4t} & \frac{1}{2}e^{-2t} - \frac{1}{2}e^{-4t} \\ -4e^{-2t} + 4e^{-4t} & -e^{-2t} + 2e^{-4t} \end{bmatrix}.$$

The impulse response matrix (3.56) reduces to a scalar because the system is SISO. $\mathbf{G}(t)$ is the systems response to a unit impulse, $u(t) = \delta(t)$,

$$\mathbf{G}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B} = y(t)$$

$$y(t) = [2 \quad 0]\begin{bmatrix} 2e^{-2t} - e^{-4t} & \frac{1}{2}e^{-2t} - \frac{1}{2}e^{-4t} \\ -4e^{-2t} + 4e^{-4t} & -e^{-2t} + 2e^{-4t} \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = e^{-2t} - e^{-4t}.$$

The unit impulse response can be seen on Fig. 3.1.

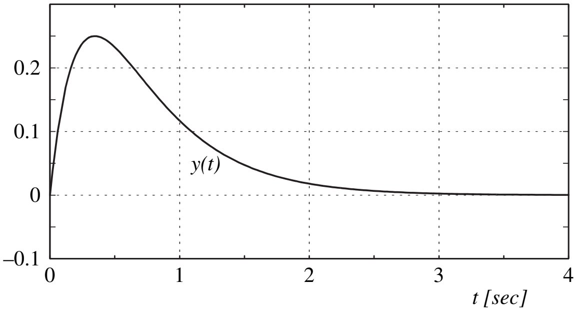


**Fig. 3.1** *Unit impulse response*

For $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $u(t) = 0$ the zero-input response can be calculated from Eq. (3.52):

$$\mathbf{x}_{zi}(t) = e^{\mathbf{A}t}\mathbf{x}_0$$

$$= \begin{bmatrix} 2e^{-2t} - e^{-4t} & \frac{1}{2}e^{-2t} - \frac{1}{2}e^{-4t} \\ -4e^{-2t} + 4e^{-4t} & -e^{-2t} + 2e^{-4t} \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3e^{-2t} - 2e^{-4t} \\ -6e^{-2t} + 8e^{-4t} \end{bmatrix}.$$

These responses are shown on Fig. 3.2.

For $\mathbf{x}_0 = \mathbf{0}$ and $u$ a unit step, i.e., $u(t) = \begin{cases} 0 \text{ for} & t < 0 \\ 1 \text{ for} & t \geq 0 \end{cases}$, the zero-state response is

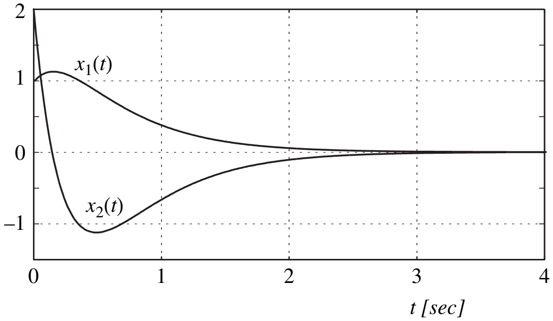


**Fig. 3.2** *Zero-input response for* $\mathbf{x}_0 = [1 \quad 2]^T$

$$\mathbf{x}_{zs}(t) = \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\cdot 1\cdot d\tau = \int_0^t \begin{bmatrix} \frac{1}{2}e^{-2(t-\tau)} - \frac{1}{2}e^{-4(t-\tau)} \\ -e^{-2(t-\tau)} + 2e^{-4(t-\tau)} \end{bmatrix} d\tau$$

$$= \begin{bmatrix} \frac{1}{4}e^{-2(t-\tau)} - \frac{1}{8}e^{-4(t-\tau)} \\ -\frac{1}{2} - e^{-2(t-\tau)} + \frac{1}{2}e^{-4(t-\tau)} \end{bmatrix}_0^t = \begin{bmatrix} \frac{1}{8} - \frac{1}{4}e^{-2t} + \frac{1}{8}e^{-4t} \\ \frac{1}{2}e^{-2t} - \frac{1}{2}e^{-4t} \end{bmatrix}.$$

For $\mathbf{x}_0 = [1 \quad 2]^T$ and $u$ a unit step the results above can be combined using superposition,

$$\mathbf{x}(t) = \mathbf{x}_{zi}(t) + \mathbf{x}_{zs}(t) = \begin{bmatrix} \frac{1}{8} + \frac{11}{4}e^{-2t} - \frac{15}{8}e^{-4t} \\ -\frac{11}{2}e^{-2t} + \frac{15}{2}e^{-4t} \end{bmatrix}.$$

Finally the output can be calculated for the given initial conditions:

$$y(t) = \mathbf{C}\mathbf{x}(t) = [2 \quad 0]\begin{bmatrix} \frac{1}{8} + \frac{11}{4}e^{-2t} - \frac{15}{8}e^{-4t} \\ -\frac{11}{2}e^{-2t} + \frac{15}{2}e^{-4t} \end{bmatrix} = \frac{1}{4} + \frac{11}{2}e^{-2t} - \frac{15}{14}e^{-4t}.$$

A plot of $\mathbf{x}(t)$ and $\mathbf{y}(t)$ is shown on Fig. 3.3. ❐

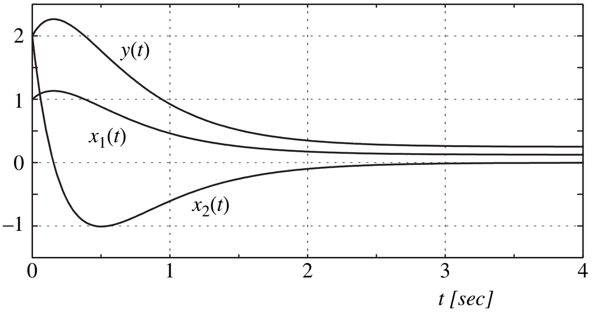


**Fig. 3.3** $\mathbf{x}(f)$ and $\mathbf{y}(f)$ for $\mathbf{x}_0 = [1 \quad 2]^T$ *and* $u(t)$ a unit step

## 3.2 Transfer Functions from State Space Models

In Sect. 2.3 it was seen how to derive a state space model given a transfer function. Here it will be related how this process can be reversed for a LTI state space model.

The impulse response for such a system is expressed by Eq. (3.56) which will be repeated here:

$$\mathbf{g}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B} + \mathbf{D}\delta(t). \tag{3.66}$$

Recalling that Eq. (3.56) is the response given the condition that $\mathbf{x}(t_0) = \mathbf{0}$, it can be Laplace transformed to obtain,

$$\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}, \tag{3.67}$$

where the result, $e^{\mathbf{A}t} = \mathscr{L}^{-1}\{(s\mathbf{I} - \mathbf{A})^{-1}\}$, has be used.

Using Eq. (3.28) with $\mathbf{g}(t, \tau) = \mathbf{g}(t - \tau)$ and $t_0 = 0$, it is seen that

$$\mathbf{y}(t) = \int_0^t \mathbf{g}(t - \tau)\mathbf{u}(\tau)d\tau. \tag{3.68}$$

The right hand side of (3.68) is a convolution integral and Laplace transformation leads to the output transfer function,

$$\mathbf{Y}(s) = \mathbf{G}(s)\mathbf{U}(s), \tag{3.69}$$

and the analogy to the scalar case is obvious. The matrix $\mathbf{G}(s)$ is called the *transfer function matrix*. It is of dimension $r \times m$ and its *i, j*'th element is the scalar transfer function from the *j*'th input to the *i*'th output.

The inverse matrix, $\mathscr{L}\{e^{At}\} = (s\mathbf{I} - \mathbf{A})^{-1}$, in Eq. (3.67) can be written,

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{adj(s\mathbf{I} - \mathbf{A})}{det(s\mathbf{I} - \mathbf{A})}, \tag{3.70}$$

where $adj(s\mathbf{I} - \mathbf{A})$ is the adjoint matrix and $det(s\mathbf{I} - \mathbf{A})$ is the determinant of the matrix $(s\mathbf{I} - \mathbf{A})$. Eq. (3.67) can then be rewritten:

$$\mathbf{G}(s) = \frac{\mathbf{C}adj(s\mathbf{I} - \mathbf{A})\mathbf{B} + det(s\mathbf{I} - \mathbf{A})\mathbf{D}}{det(s\mathbf{I} - \mathbf{A})}. \tag{3.71}$$

The numerator is an $r \times m$-matrix and for SISO systems it is a scalar. The denominator is always a scalar; as a matter of fact, it is a polynomial in $s$. For the usual scalar transfer function of a SISO system the denominator polynomial is called the *characteristic polynomial* and the same notation will be used for MIMO systems. The systems poles are the solutions of the *characteristic equation*:

$$det(s\mathbf{I} - \mathbf{A}) = 0. \tag{3.72}$$

As noted on p. 68, the same equation must be solved to find the eigenvalues of the matrix **A,**

$$det(\lambda\mathbf{I} - \mathbf{A}) = 0. \tag{3.73}$$

This shows that the poles of a transfer function of a LTI system are the same as the eigenvalues of the system matrix of the corresponding state space model.

Note that in this statement the possibility has been neglected that zeroes and poles cancel each other in the transfer functions. If such a cancellation occurs, there are eigenvalues of **A** which are *not* poles in the transfer function. For this reason it is necessary to distinguish between the *eigenvalues* of the **A** matrix and the *poles* of the transfer function matrix.

### *3.2.1 Natural Modes*

If the characteristic polynomial in Eq. (3.71) is factored the transfer function can be written,

$$G(s) = \frac{\mathbf{C}adj(s\mathbf{I} - \mathbf{A})\mathbf{B} + det(s\mathbf{I} - \mathbf{A})\mathbf{D}}{\prod_{i=1}^{n}(s - p_i)}, \tag{3.74}$$

where $p_i$ are the poles. For simplicity it has been assumed that all the eigenvalues are distinct.

This expression can be decomposed (using partial fractions) to a sum of terms

$$\mathbf{G}(s) = \frac{\mathbf{R}_1}{s - p_1} + \frac{\mathbf{R}_2}{s - p_2} + \ldots = \sum_{i=1}^{n} \frac{\mathbf{R}_i}{s - p_i}. \tag{3.75}$$

The constant residual matrices $\mathbf{R}_i$ are of dimension $r \times m$.
Inverse Laplace transformation of (3.75) yields the impulse response of the system,

$$\mathbf{g}(t) = \sum_{i=1}^{n} \mathbf{R}_i e^{p_i t}, \quad t \geq 0. \tag{3.76}$$

As it is the case for scalar systems, the response (3.76) is composed of terms containing the exponential functions $e^{p_i t}$.

It may be recalled from p. 69 that these terms are called the *natural modes* of the system. The natural modes determine the characteristics of the system's dynamic behaviour. The poles/eigenvalues, $p_i$, can be real or complex. In the first case, the system will have a time constant,

$$\tau_i = -\frac{1}{p_i}.$$

In the second case they will appear as complex conjugate pairs, $p_{i,\,i+1} = \alpha \pm j\beta$, and the system will have a natural frequency and a damping ratio associated with each of them,

$$\omega_n = \sqrt{\alpha^2 + \beta^2}, \quad \zeta = -\frac{\alpha}{\sqrt{\alpha^2 + \beta^2}}.$$

If the eigenvalue, $p_i$, is repeated k times, the corresponding natural modes will be

$$e^{p_i t}, te^{p_i t}, t^2 e^{p_i t}, \ldots, t^{k-1} e^{p_i t}.$$

See also Eq. (3.61).

***Example 3.2*. Impulse Response of Second Order System**

Consider the following system:.

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} 0 & 1 \\ -8 & -6 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} u, \\ y &= \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \mathbf{x}. \end{aligned}$$

The system matrix is the same as in Example 3.1, so it is known from that example that

$$\Phi(s) = \begin{bmatrix} \dfrac{2}{s+2} - \dfrac{1}{s+4} & \dfrac{1/2}{s+2} - \dfrac{1/2}{s+4} \\ -\dfrac{4}{s+2} + \dfrac{4}{s+4} & -\dfrac{1}{s+2} + \dfrac{2}{s+4} \end{bmatrix}.$$

The transfer function matrix will be

$$\begin{aligned} \mathbf{G}(s) = \mathbf{C}\Phi(s)\mathbf{B} + \mathbf{D} &= \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} \dfrac{2}{s+2} - \dfrac{1}{s+4} & \dfrac{1/2}{s+2} - \dfrac{1/2}{s+4} \\ -\dfrac{4}{s+2} + \dfrac{4}{s+4} & -\dfrac{1}{s+2} + \dfrac{2}{s+4} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \dfrac{1}{s+2} - \dfrac{1}{s+4} & \dfrac{4}{s+2} - \dfrac{2}{s+4} \\ \dfrac{-3}{s+2} + \dfrac{6}{s+4} & \dfrac{-12}{s+2} + \dfrac{12}{s+4} \end{bmatrix} = \frac{\begin{bmatrix} 1 & 4 \\ -3 & -12 \end{bmatrix}}{s+2} + \frac{\begin{bmatrix} -1 & -2 \\ 6 & 12 \end{bmatrix}}{s+4}. \end{aligned}$$

The impulse response can be found by inverse Laplace transformation,

$$\mathbf{g}(t) = \begin{bmatrix} 1 & 4 \\ -3 & -12 \end{bmatrix} e^{-2t} + \begin{bmatrix} -1 & -2 \\ 6 & 12 \end{bmatrix} e^{-4t}$$

$$= \begin{bmatrix} e^{-2t} - e^{-4t} & 4e^{-2t} - 2e^{-4t} \\ -3e^{-2t} + 6e^{-4t} & -12e^{-2t} + 12e^{-4t} \end{bmatrix}, t \geq 0.$$

❒

## 3.3 Discrete Time Models of Continuous Systems

As will be apparent later, discrete time state equations are easier to solve than the continuous equations, especially if a computer is used to do the job. The discrete time state equations are actually difference equations and their recursive nature makes them very easy to implement directly as computer algorithms (see Sect. 2.5). Discrete time state equations are also important for another reason. As will also be seen later, it is straightforward to design controllers in discrete time and the natural basis for such a design procedure is a discrete time state equation.

When deriving the discrete time counterpart of the continuous state equation, it is usually assumed that all signals in the system are constant between the sample instants. Hence they can only change value at these sample instants. This is equivalent to inserting samplers and zero-order holds in the model's signal lines. One of the consequences is of course that all signal information between the sample instants is discarded but that is a problem that one always has to live with when working in discrete time.

The solution to the LTI state equation (3.37) has been shown earlier to be

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 + e^{\mathbf{A}t} \int_0^t e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)d\tau. \tag{3.77}$$

If the solution of this equation for the two time instants $t_1 = kT$ and $t_2 = (k+1)T$ is calculated then one has

$$\mathbf{x}(kT) = e^{\mathbf{A}kT}\mathbf{x}_0 + e^{\mathbf{A}kT} \int_0^{kT} e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)d\tau \tag{3.78}$$

and

$$\mathbf{x}((k+1)T) = e^{\mathbf{A}(k+1)T}\mathbf{x}_0 + e^{\mathbf{A}(k+1)T}\int_0^{(k+1)T} e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)d\tau \tag{3.79}$$

respectively.

Premultiplying (3.78) by $e^{\mathbf{A}T}$ and subtracting the result from (3.79),

$$\mathbf{x}((k+1)T) = e^{\mathbf{A}T}\mathbf{x}(kT) + e^{\mathbf{A}(k+1)T}\int_{kT}^{(k+1)T} e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)d\tau. \tag{3.80}$$

Since $\mathbf{u}(t)$ is considered constant between the sample instants, one has that $\mathbf{u}(t) = \mathbf{u}(kT)$ on the interval $t \in [kT, ((k+1)T)]$. Then (3.80) can be rewritten

$$\mathbf{x}((k+1)T) = e^{\mathbf{A}T}\mathbf{x}(kT)\int_{kT}^{(k+1)T} e^{\mathbf{A}((k+1)T-\tau)}\mathbf{B}d\tau \cdot \mathbf{u}(kT). \tag{3.81}$$

Changing variables of the integration, $(k+1)T - \tau = t$, finally

$$\mathbf{x}((k+1)T) = e^{\mathbf{A}T}\mathbf{x}(kT) + \int_0^T e^{\mathbf{A}t}\mathbf{B}dt \cdot \mathbf{u}(kT). \tag{3.82}$$

Defining the two constant matrices,

$$\mathbf{F} = e^{\mathbf{A}T} \text{ and } \mathbf{G} = \int_0^T e^{\mathbf{A}t}\mathbf{B}dt. \tag{3.83}$$

Equation (3.82) can be written,

$$\mathbf{x}((k+1)T) = \mathbf{F}\mathbf{x}(kT) + \mathbf{G}\mathbf{u}(kT), \tag{3.84}$$

which is the discrete time state equation. In agreement with (2.130) this is abbreviated to,

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k). \tag{3.85}$$

The corresponding output equation is identical to the continuous counterpart with the exception that the time argument is discrete,

$$\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k) \tag{3.86}$$

Since the derivation above is based on the assumption that all signals change values in a step-wise manner (they are all staircase curves), the procedure is sometimes called the *step-invariant transformation*.

According to Frobenius' theorem[†] the discrete time system matrix has the eigenvalues

$$\lambda_{\mathbf{F}} = e^{\lambda_{\mathbf{A}} T}. \tag{3.87}$$

***Example 3.3*. Transfer Function of a Discrete Time System**

For a SISO system the block diagram and the transfer function $G(s)$, are given on Fig. 3.4.

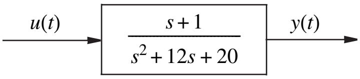


**Fig. 3.4** SISO system

Using the results in Sect. 2.3, it is possible to write down the continuous state equation by inspection of the transfer function,

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -20 & -12 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u,$$
$$y = [1 \ 1]\mathbf{x}.$$

On this basis the resolvent matrix can be calculated to be

$$\Phi(s) = (sI - A)^{-1} = \frac{1}{(s+2)(s+10)} \begin{bmatrix} s+12 & 1 \\ -20 & s \end{bmatrix}$$

and the state transfer matrix can be found by inverse Laplace transformation,

$$e^{\mathbf{A}t} = \phi(t) = \frac{1}{8} \begin{bmatrix} 10e^{-2t} - 2e^{-10t} & e^{-2t} - e^{-10t} \\ -20e^{-2t} + 20e^{-10t} & -2e^{-2t} + 10e^{-10t} \end{bmatrix}.$$

If a sampler and a zero-order hold are added in front of the system on Fig. 3.4 as shown on Fig. 3.5 the system can be discretized by using (3.83).

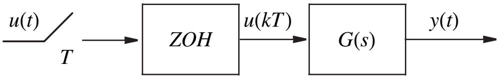


**Fig. 3.5** SISO system with sampler and zero-order hold

[†] Frobenius' theorem: If a matrix **A** has the eigenvalues $\lambda_i$ then the function $f(\mathrm{A})$ has the eigenvalues $f(\lambda_i)$ if the function $f(z)$ is analytic in a region in the complex plane containing the eigenvalues $\lambda_i$. The function $e^{zT}$ is analytic in any finite region of the complex plane.

For a sample period, $T = 1$ sec, the matrices **F** and**G** can be found

$$\mathbf{F} = e^{\mathbf{A}T}|_{T=1} = \begin{bmatrix} 0.169 & 0.0169 \\ -0.338 & -0.0338 \end{bmatrix},$$

the eigenvalues of **A** and **F** are

$$\lambda_{\mathbf{A}} = \begin{cases} -2 \\ -10 \end{cases} \qquad \lambda_{\mathbf{F}} = \begin{cases} 0.1353 \\ 4.54 \cdot 10^{-5} \end{cases}$$

and it can be seen that Frobenius' theorem (3.87) holds for the system .

$$\begin{aligned}\mathbf{G} &= \int_0^T e^{\mathbf{A}t}\mathbf{B}dt = \frac{1}{8}\int_0^1 \begin{bmatrix} e^{-2t} - e^{-10t} \\ -2e^{-2t} + 10e^{-10t} \end{bmatrix} dt \\ &= \frac{1}{80}\begin{bmatrix} -5e^{-2} + e^{-10} + 4 \\ 10e^{-2} - 10e^{-10} \end{bmatrix} = \begin{bmatrix} 0.0415 \\ 0.0169 \end{bmatrix}.\end{aligned}$$

The matrix **C** will be the same as for the continuous system,

$$\mathbf{C} = [1 \;\; 1].$$

This method allows calculation of the exact discrete time matrices but will be very laborious to apply for systems order greater than, say, 2 to 3. The alternative is to use computer computation. ❐

***Example 3.4.* Linearized Oven Process Plant**

In Example 2.9 a linearized time invariant model of a thermal process plant is derived,

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{Ax}(t) + \mathbf{Bu}(t) + \mathbf{B}_v\mathbf{v}(t), \\ \mathbf{y}(t) &= \mathbf{Cx}(t),\end{aligned}$$

where $\mathbf{v}(t)$ is a vector valued disturbance function. Note that the $\Delta$ denoting the incremental states has been omitted and that all variables are deviations from a stationary operating point (see Eq. (2.64)).

The continuous model turned out to be of order 4 and with the given data the system's matrices are

$$\mathbf{A} = \begin{bmatrix} -0.0499 & 0.0499 & 0 & 0 \\ 0.0499 & -0.0667 & 0 & 0 \\ 0 & 0 & -0.0251 & 0 \\ 0 & 0 & 0.0335 & -0.0335 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0.00510 & 0.00510 \\ 0 & 0 \\ 0.0377 & -0.0377 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}_v = \begin{bmatrix} 0 & 0 & 0 \\ -4.177 & 0 & 0 \\ 0 & 0.01255 & 0.01255 \\ 0 & 0 & 0 \end{bmatrix}.$$

The eigenvalues of the system matrix **A** can be found to be

$$\lambda_i = \begin{cases} -0.00769 \\ -0.1089 \\ -0.0335 \\ -0.0251 \end{cases}.$$

This means that the system is characterized by the 4 time constants,

$$\tau_i = \begin{cases} 130 \text{ sec} \\ 9.18 \text{ sec} \\ 29.9 \text{ sec} \\ 39.8 \text{ sec} \end{cases}.$$

If the system is sampled approximately 5 times per $\tau_{min}$, this gives $T = 2$s which corresponds to the sample frequency $f_s = 0.5$ Hz.

MATLAB provides an algorithm for numerical computation of **F** and **G**. The computation is executed by the command ('continuous-to-discrete'):

$$\texttt{[F,G]} = \texttt{c2d(A,B,T)}.$$

The result are

$$\mathbf{F} = \begin{bmatrix} 0.9094 & 0.0890 & 0 & 0 \\ 0.0890 & 0.8795 & 0 & 0 \\ 0 & 0 & 0.951 & 0 \\ 0 & 0 & 0.0632 & 0.9352 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 0.00973 & 0.00973 \\ 0.000471 & 0.000471 \\ 0.0735 & -0.0735 \\ 0.00243 & -0.00243 \end{bmatrix}.$$

The discrete equivalent to $\mathbf{B}_v$ can be computed using the same command as above but with $\mathbf{B}$ replaced by $\mathbf{B}_v$:

$$\texttt{[F,Gv]} = \texttt{c2d(A,Bv,T)},$$

$$\mathbf{G}_v = \begin{bmatrix} -0.386 & 0 & 0 \\ -7.83 & 0 & 0 \\ 0 & 0.0245 & 0.0245 \\ 0 & 0.00081 & 0.00081 \end{bmatrix}.$$

This shows that the numerical structures of the continuous time matrices and their discrete time equivalents resemble each other to some extent. It should be pointed out that such a similarity is not always found. This can also be seen in the next example. ❒

***Example 3.5*. Phase Variable State Space Model**

Considering a state space model in the phase variable form, it is known (see Sect. 2.3) that it will only contain the necessary parameters, namely the coefficients of the nominator and the denominator polynomials. An example could be

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & -5 & -7 & -2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} u, \; y = [1 \quad 3 \quad 0 \quad 0]\mathbf{x}.$$

$\mathbf{A}$'s eigenvalues are

$$\lambda_i = \begin{cases} -0.597 \pm j\,2.3.1 \\ -0.403 \pm j\,0.435 \end{cases},$$

and the largest of the two natural frequencies is $\omega_{n,max} = 2.39\,\text{rad/s}$ or $f_{n,max} = 0.38\,\text{Hz}$. It is desireable to use a sample frequency at least 10 times higher, choose $f_s = 5\,\text{Hz}$ or $T = 0.2$ s here.

The MATLAB command from Example 3.4 generates the result

$$\mathbf{F} = \begin{bmatrix} 0.9999 & 0.1997 & 0.0196 & 0.0012 \\ -0.0024 & 0.9939 & 0.1913 & 0.0172 \\ -0.0343 & -0.0883 & 0.8737 & 0.1570 \\ -0.3140 & -0.8193 & -1.187 & 0.5597 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 0.0000611 \\ 0.001193 \\ 0.01717 \\ 0.1570 \end{bmatrix}.$$

It is clearly seen that the simple structure of $\mathbf{A}$ and $\mathbf{B}$ is lost by the discretization. $\mathbf{A}$ and $\mathbf{B}$ had only 4 elements different from zero or one whereas $\mathbf{F}$ and $\mathbf{G}$ have 20. ❒

## 3.4 Solution of the Discrete Time State Equation

The discrete time and time varying state equation (2.132) is in principle easy to solve because it is in reality a *recursion formula,*

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{F}(k)\mathbf{x}(k) + \mathbf{G}(k)\mathbf{u}(k),\\ \mathbf{y}(k) &= \mathbf{C}(k)\mathbf{x}(k) + \mathbf{D}(k)\mathbf{u}(k).\end{aligned} \tag{3.88}$$

If the initial state is known, the state at the next sample instant can be calculated and so on. If the initial time and state are $k_0$ and $\mathbf{x}(k_0)$ respectively, it is found that

$$\begin{aligned}
&\mathbf{x}(k_0+1) = \mathbf{F}(k_0)\mathbf{x}(k_0) + \mathbf{G}(k_0)\mathbf{u}(k_0),\\
&\mathbf{x}(k_0+2) = \mathbf{F}(k_0+1)\mathbf{x}(k_0+1) + \mathbf{G}(k_0+1)\mathbf{u}(k_0+1)\\
&\ = \mathbf{F}(k_0+1)\mathbf{F}(k_0)\mathbf{x}(k_0) + \mathbf{F}(k_0+1)\mathbf{G}(k_0)\mathbf{u}(k_0) + \mathbf{G}(k_0+1)\mathbf{u}(k_0+1),\\
&\ \vdots\\
&\ \vdots\\
&\mathbf{x}(k_0+k) = \mathbf{F}(k_0+k-1)\mathbf{x}(k_0+k-1) + \mathbf{G}(k_0+k-1)\mathbf{u}(k_0+k-1)\\
&\ = \mathbf{F}(k_0+k-1)\mathbf{F}(k_0+k-2)\ldots\mathbf{F}(k_0)\mathbf{x}(k_0)\\
&\qquad + \mathbf{F}(k_0+k-1)\mathbf{F}(k_0+k-2)\ldots\mathbf{F}(k_0+1)\mathbf{G}(k_0)\mathbf{u}(k_0)\\
&\qquad + \mathbf{F}(k_0+k-1)\mathbf{G}(k_0+k-2)\ldots\mathbf{F}(k_0+2)\mathbf{G}(k_0+1)\mathbf{u}(k_0+1),\\
&\ \vdots\\
&\qquad + \mathbf{F}(k_0+k-1)\mathbf{G}(k_0+k-2)\mathbf{u}(k_0+k-2)\\
&\qquad + \mathbf{G}(k_0+k-1)\mathbf{u}(k_0+k-1).
\end{aligned} \tag{3.89}$$

Comparing Eq. (3.89) with (3.20) the *discrete time state transition matrix* can be defined:

$$\phi(l,m) = \mathbf{F}(l-1)\mathbf{F}(l-2)\ldots\mathbf{F}(m)\ ,\ \phi(l,l) = \mathbf{I}. \tag{3.90}$$

The general solution can then be written,

$$\mathbf{x}(k) = \begin{cases}\mathbf{x}(k_0) \quad \text{for} \quad k = k_0,\\ \phi(k,k_0)\mathbf{x}(k_0) + \sum\limits_{i=k_0}^{k-1}\phi(k,i+1)\mathbf{G}(i)u(i) \quad \text{for} \quad k \geq k_0+1,\end{cases} \tag{3.91}$$

and the output becomes,

$$\mathbf{y}(k) = \begin{cases} \mathbf{C}(k_0)\mathbf{x}(k_0) + \mathbf{D}(k_0)\mathbf{u}(k_0) & \text{for} \quad k = k_0, \\ \mathbf{C}(k)\phi(k, k_0)\mathbf{x}(k_0) & \\ + \sum\limits_{i=k_0}^{k-1} \mathbf{C}(k)\phi(k, i+1)\mathbf{G}(i)\mathbf{u}(i) + \mathbf{D}(k)\mathbf{u}(k) & \text{for} \quad k \geq k_0 + 1. \end{cases} \tag{3.92}$$

In this case the *unit pulse response* can be defined:

$$\mathbf{h}(k, i) = \begin{cases} \mathbf{D}(k) & \text{for} \quad k = i \\ \mathbf{C}(k)\phi(k, i+1)\mathbf{G}(i) & \text{for} \quad k > i \end{cases}. \tag{3.93}$$

For $\mathbf{x}(k_0) = \mathbf{0}$,, the output is

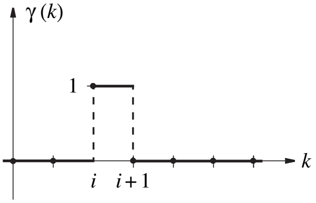
$$\mathbf{y}(k) = \sum_{i=k_0}^{k} \mathbf{h}(k, i)\mathbf{u}(i). \tag{3.94}$$

The matrix element $h_{lj}(k, i)$ is the l'th element of $\mathbf{y}(k)$ when the j'th element of $\mathbf{u}(k)$ is a *unit pulse* at time $i$ and all other elements of $\mathbf{u}$ are zero, i.e.,

$$\mathbf{u}(k) = [0 \ \ldots \ 0 \ \gamma(k) \ 0 \ \ldots \ 0]^T, \tag{3.95}$$

where the unit pulse is defined as seen on Fig. 3.6.

**Fig. 3.6** Unit pulse function



As in the continuous case, the results above are difficult to use for further analytical purposes because a closed form expression for (3.90) cannot in general be found.

### 3.4.1 The Time Invariant Discrete Time System

If the matrices in Eq. (3.88) are constant,

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k), \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k), \end{aligned} \tag{3.96}$$

then the initial time and state vector can be set to $k_0 = 0$ and $\mathbf{x}(k_0) = \mathbf{x}(0) = \mathbf{x}_0$. This simplifies the solution considerably. Recursive calculation yields

$$\begin{aligned}
\mathbf{x}(1) &= \mathbf{F}\mathbf{x}_0 + \mathbf{G}\mathbf{u}(0),\\
\mathbf{x}(2) &= \mathbf{F}\mathbf{x}(1) + \mathbf{G}\mathbf{u}(1) = \mathbf{F}^2\mathbf{x}_0 + \mathbf{F}\mathbf{G}\mathbf{u}(0) + \mathbf{G}\mathbf{u}(1),\\
&\vdots\\
\mathbf{x}(k) &= \mathbf{F}^k\mathbf{x}_0 + \mathbf{F}^{k-1}\mathbf{G}\mathbf{u}(0) + \ldots + \mathbf{F}\mathbf{G}\mathbf{u}(k-2) + \mathbf{G}(k-1).
\end{aligned} \tag{3.97}$$

The expression for the $k$'th step can be rewritten as

$$\mathbf{x}(k) = \mathbf{F}^k\mathbf{x}_0 + \sum_{i=0}^{k-1} \mathbf{F}^{k-1-i}\mathbf{G}\mathbf{u}(i). \tag{3.98}$$

The state transition matrix (3.90) is now reduced to

$$\phi(k) = \mathbf{F}^k. \tag{3.99}$$

This is the discrete time counterpart of the continuous time exponential matrix, $e^{\mathbf{A}t}$.

The parallel to the continuous solution (3.49) can clearly be seen. The first term of (3.98) is the zero-input solution and the second term is the zero-state solution. This sum is a *discrete convolution* of the two discrete time functions, $\mathbf{F}^k$ and $\mathbf{u}(k)$.

For the output one has

$$\mathbf{y}(k) = \mathbf{C}\mathbf{F}^k\,\mathbf{x}_0 + \sum_{i=0}^{k-1} \mathbf{C}\mathbf{F}^{k-1-i}\mathbf{G}\mathbf{u}(i) + \mathbf{D}\mathbf{u}(k). \tag{3.100}$$

The unit pulse response can also be defined here. Setting

$$\mathbf{h}(k) = \begin{cases} \mathbf{D} \quad \text{for} \quad k = 0 \\ \mathbf{C}\mathbf{F}^{k-1}\mathbf{G} \quad \text{for} \quad k \geq 1 \end{cases}, \tag{3.101}$$

equation (3.100) for $\mathbf{x}_0 = \mathbf{0}$ can be written as

$$\mathbf{y}(k) = \sum_{i=0}^{k} \mathbf{h}(k-i)\mathbf{u}(i). \tag{3.102}$$

The matrix element $h_{lj}(k)$ is the l'th element of $\mathbf{y}(k)$ when the j'th element of $\mathbf{u}(k)$ is a unit pulse at time zero and all other elements of $\mathbf{u}$ are zero, see also the expression (3.95) and Fig. 3.6.

### The Z-transform Method

It is possible to derive the solution to the time invariant equation in an alternative way using the Z-transformation. See also Appendix D.

The Z-transform of the discrete time function $f(k)$ is defined by

$$Z\{f(k)\} = F(z) = \sum_{k=0}^{\infty} f(k) z^{-k}. \tag{3.103}$$

The shift properties of the Z-transform need to be used here:

$$Z\{f(k-1)\} = z^{-1}F(z) + f(-1) \text{ (backward shift)}, \tag{3.104}$$

$$Z\{f(k+1)\} = zF(z) - zf(0) \text{ (forward shift)}. \tag{3.105}$$

The discrete time state vector $\mathbf{x}(k)$ can be Z-transformed element by element and the time invariant state equation,

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k), \tag{3.106}$$

is transformed to

$$Z\{\mathbf{x}(k+1)\} = z\mathbf{X}(z) - z\mathbf{x}_0 = \mathbf{F}\mathbf{X}(z) + \mathbf{G}\mathbf{U}(z), \tag{3.107}$$

which can be solved for $\mathbf{X}(z)$,

$$\mathbf{X}(z) = (z\mathbf{I} - \mathbf{F})^{-1} z\mathbf{x}_0 + (z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G}\mathbf{U}(z). \tag{3.108}$$

With the definition,

$$\Psi(z) = (z\mathbf{I} - \mathbf{F})^{-1} z. \tag{3.109}$$

The solution can be written down as

$$\mathbf{X}(z) = \Psi(z)\mathbf{x}_0 + (z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G}\mathbf{U}(z). \tag{3.110}$$

Inverse Z-transformation of this equation leads to

$$\mathbf{x}(k) = Z^{-1}\{(z\mathbf{I} - \mathbf{F})^{-1} z\}\mathbf{x}_0 + Z^{-1}\{(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G}\mathbf{U}(z)\}. \tag{3.111}$$

Comparing (3.111) with (3.98) and using (3.109), the discrete time transition matrix can be found from,

$$\phi(k) = \mathbf{F}^k = Z^{-1}\{(z\mathbf{I} - \mathbf{F})^{-1}z\} = Z^{-1}\{\Psi(z)\}, \tag{3.112}$$

which is the discrete time counterpart of (3.50). The matrix,

$$\phi(z) = (z\mathbf{I} - \mathbf{F})^{-1}z, \tag{3.113}$$

is called the *discrete time resolvent matrix*.

Equation (3.109) can be written as

$$\Psi(z) = \frac{adj(z\mathbf{I} - \mathbf{F}) \cdot z}{det(z\mathbf{I} - \mathbf{F})}. \tag{3.114}$$

With the eigenvalues of **F** found from the characteristic equation,

$$det(\lambda\mathbf{I} - \mathbf{F}) = 0, \tag{3.115}$$

one can, similarly to the continuous case, factor the denominator in (3.114),

$$\Psi(z) = \frac{adj(z\mathbf{I} - \mathbf{F}) \cdot z}{(z-\lambda_1)(z-\lambda_2)\ldots(z-\lambda_r)^m\ldots(z-\lambda_\sigma)}, \tag{3.116}$$

where it is assumed that distinct as well as repeated eigenvalues exist in the system.

Partial fraction decomposition yields

$$\begin{aligned}\frac{1}{z}\Psi(z) &= \frac{\mathbf{Z}_1}{z-\lambda_1} + \frac{\mathbf{Z}_2}{z-\lambda_2} + \ldots \\ &+ \frac{\mathbf{Z}_{r1}}{z-\lambda_r} + \frac{\mathbf{Z}_{r2}}{(z-\lambda_r)^2} + \ldots + \frac{\mathbf{Z}_{rm}}{(z-\lambda_r)^m} + \ldots + \frac{\mathbf{Z}_\sigma}{z-\lambda_\sigma}\end{aligned} \tag{3.117}$$

or

$$\begin{aligned}\Psi(z) &= \frac{z\mathbf{Z}_1}{z-\lambda_1} + \frac{z\mathbf{Z}_2}{z-\lambda_2} + \ldots \\ &+ \frac{z\mathbf{Z}_{r1}}{z-\lambda_r} + \frac{z\mathbf{Z}_{r2}}{(z-\lambda_r)^2} + \ldots + \frac{z\mathbf{Z}_{rm}}{(z-\lambda_r)^m} + \ldots + \frac{z\mathbf{Z}_\sigma}{z-\lambda_\sigma}\end{aligned} \tag{3.118}$$

and inverse Z-transformation gives the state transition matrix,

$$\begin{aligned}\phi(k) = \mathbf{F}^k = \mathbf{Z}_1\lambda_1^k + \mathbf{Z}_2\lambda_2^k + \dots \\ + F(\mathbf{Z}_{r1}, \mathbf{Z}_{r2}, \dots, \mathbf{Z}_{rm}, k, \lambda_r) + \dots \mathbf{Z}_\sigma\lambda_\sigma^k,\end{aligned} \tag{3.119}$$

where the function $F$ contains terms of the form $\mathbf{Z}_{rq}k^{q-1}\lambda_r^{k-k_q}$, where $k_q$ is a positive integer for $q = 1, 2, \dots, m$.

## 3.5 Discrete Time Transfer Functions

Z-transforming the output equation in (3.96) and inserting (3.108) yields

$$\begin{aligned}\mathbf{Y}(z) = \mathbf{CX}(z) + \mathbf{DU}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{F})^{-1}z\mathbf{x}_0 \\ + \mathbf{C}(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{GU}(z) + \mathbf{DU}(z).\end{aligned} \tag{3.120}$$

For the initial condition, $\mathbf{x}_0 = 0$, it is seen that

$$\mathbf{Y}(z) = (\mathbf{C}(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G} + \mathbf{D})\mathbf{U}(z). \tag{3.121}$$

In accordance with the normal practice, the function,

$$\mathbf{H}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G} + \mathbf{D} = \frac{\mathbf{C}adj(z\mathbf{I} - \mathbf{F})\mathbf{G} + det(z\mathbf{I} - \mathbf{F})\mathbf{D}}{det(z\mathbf{I} - \mathbf{F})}, \tag{3.122}$$

is called the *discrete transfer function matrix*. Equation (3.121) can then be written

$$\mathbf{Y}(z) = \mathbf{H}(z)\mathbf{U}(z). \tag{3.123}$$

This could also have been found directly from Eq. (3.102). The right hand side expression of this equation is the discrete convolution of the unit pulse response and the input function. The Z-transform of (3.102) is the same as (3.123).

As for the continuous case, the *poles* of the transfer functions will be the same as the *eigenvalues* of the system matrix **F** if no pole-zero cancellation occurs in the transfer function. This means that the poles are those solutions, $z = p_i$, of the equation

$$det(z\mathbf{I}-\mathbf{F})=0 \tag{3.124}$$

which are not cancelled by zeros.

The terms $p_i^k$ or, in the case of repeated poles, $k^l p_i^{k-k_i}$, are called the *discrete time natural modes*. $p_i$ can be real or complex. For a real pole the continuous system will exhibit a time constant,

$$\tau_i=-\frac{T}{\log_e p_i}.$$

For a complex pair of poles, $p_{i,i+1}=\alpha\pm j\beta$, the continuous system will have the natural frequency and damping ratio,

$$\omega_n=\frac{|\log_e p_i|}{T},\ \zeta=-\cos\angle\log_e p_i.$$

Linear State Equation Solution Overview

| | Continuous Time | Discrete Time |
|---|---|---|
| State equation | $\dot{\mathbf{x}}(t)=\mathbf{A}\mathbf{x}(t)+\mathbf{B}\mathbf{u}(t),\ \mathbf{x}(0)=\mathbf{x}_0$<br>$\mathbf{y}(t)=\mathbf{C}\mathbf{x}(t)+\mathbf{D}\mathbf{u}(t)$ | $\mathbf{x}(k+1)=\mathbf{F}\mathbf{x}(k)+\mathbf{G}\mathbf{u}(k),\ \mathbf{x}(0)=\mathbf{x}_0$<br>$\mathbf{y}(k)=\mathbf{C}\mathbf{x}(k)+\mathbf{D}\mathbf{u}(k)$ |
| Resolvent matrix | $\Phi(s)=(s\mathbf{I}-\mathbf{A})^{-1}$ | $\Phi(z)=(z\mathbf{I}-\mathbf{F})^{-1}z$ |
| State Transition matrix | $\phi(t)=e^{\mathrm{A}t}$ | $\phi(k)=\mathbf{F}^k$ |
| Solution | $\mathbf{x}(t)=e^{\mathbf{A}t}\mathbf{x}_0+\int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau$ | $\mathbf{x}(k)=\mathbf{F}^k\mathbf{x}_0+\sum_{i=0}^{k-1}\mathbf{F}^{k-i-1}\mathbf{G}\mathbf{u}(i)$ |
| Impulse/pulse response | $\mathbf{g}(t)=\mathbf{C}e^{\mathbf{A}t}\mathbf{B}+\mathbf{D}\delta(t)$ | $\mathbf{h}(k)=\begin{cases}\mathbf{D}\text{ for }k=0\\ \mathbf{C}\mathbf{F}^{k-1}\mathbf{G}\text{ for }k\geq 1\end{cases}$ |
| Output | $\mathbf{y}(t)=\int_0^t\mathbf{g}(t-\tau)\mathbf{u}(\tau)d\tau$ | $\mathbf{y}(k)=\sum_{i=0}^{k}\mathbf{h}(k-i)\mathbf{u}(i)$ |
| Eigenvalues | Solutions to $det(\lambda\mathbf{I}-\mathbf{A})=0$ | Solutions to $det(\lambda\mathbf{I}-\mathbf{F})=0$ |
| Natural modi | $e^{\lambda_i t},\ t^{m-1}e^{\lambda_i t}$ | $\lambda_i^k,\ k^l\lambda_i^{k-k_i}$ |
| Transfer function | $\mathbf{G}(s)=\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}$ | $\mathbf{H}(z)=\mathbf{C}(z\mathbf{I}-\mathbf{F})^{-1}\mathbf{G}+\mathbf{D}$ |

***Example 3.6*. Time Response of a Discrete Time System**

Suppose that a system is described by the matrices:

$$\mathbf{A}=\begin{bmatrix}0 & 1\\ -4 & -1\end{bmatrix},\ \mathbf{B}=\begin{bmatrix}0\\ 1\end{bmatrix},\ \mathbf{C}=[\,1\ \ 0\,],\ \mathbf{D}=0.$$

The eigenvalues of **A** are complex conjugates,

$$\lambda_{\mathbf{A}} = -0.5 \pm j1.937,$$

which means that the natural frequency and the damping ratio are

$$\omega_n = 2 \text{ rad/sec}, \ \zeta = 0.25.$$

The system is now discretized according to the rules in Sect. 3.3 using the sample period $T = 0.1$ s. Applying Eq. (3.83) leads to the matrices,

$$\mathbf{F} = \begin{bmatrix} 0.9807 & 0.09453 \\ -0.3781 & 0.8862 \end{bmatrix}, \ \mathbf{G} = \begin{bmatrix} 0.004821 \\ 0.09453 \end{bmatrix}.$$

The eigenvalues of **F** are

$$\lambda_{\mathbf{F}} = 0.9335 \pm j0.1831$$

and it is seen that the relation,

$$\lambda_{\mathbf{F}} = e^{0.1\lambda_{\mathbf{A}}}$$

holds (see Eq. (3.87)).

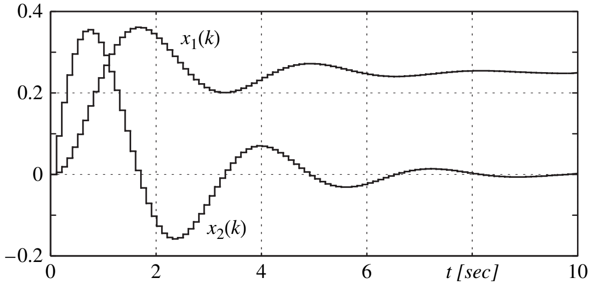
One can calculate $\mathbf{x}$ by using the state equations recursively or by applying Eq. (3.98). Assuming $\mathbf{x}_0 = 0$ with $u(k)$ a unit step, one obtains

$$\begin{aligned} x_1(1) &= 0.9807x_1(0) + 0.09453x_2(0) + 0.004821 \cdot 1 = 0.004821, \\ x_2(1) &= -0.3781x_1(0) + 0.8862x_2(0) + 0.09453 \cdot 1 = 0.09453, \\ y(1) &= 1 \cdot x_1(1) = 0.004821. \end{aligned}$$

$$\begin{aligned} x_1(2) &= 0.9807x_1(1) + 0.09453x_2(1) + 0.004821 \cdot 1 = 0.01849, \\ x_2(2) &= -0.3781x_1(1) + 0.8862x_2(1) + 0.09453 \cdot 1 = 0.1765, \\ y(2) &= 1 \cdot x_1(2) = 0.01849. \end{aligned}$$

$$\begin{aligned} x_1(3) &= 0.9807x_1(2) + 0.09453x_2(2) + 0.004821 \cdot 1 = 0.03963, \\ x_2(3) &= -0.3781x_1(2) + 0.8862x_2(2) + 0.09453 \cdot 1 = 0.2439, \\ y(3) &= 1 \cdot x_1(3) = 0.03963. \\ &\vdots \\ &\vdots \end{aligned}$$

**Fig. 3.7** Discrete time states for a unit step input ($T = 0.1$)



A plot of the two states can be seen on Fig. 3.7.

On Fig. 3.8 the response $y(k) = x_1(k)$ is shown for the same input. For comparison the unit step response $y(t)$ of the continuous system is also shown. Note that the discrete time curves are drawn as staircase curves. From a rigorous point of view this is not correct. The discrete state equation (and the discrete transfer function) provides information on the system behaviour at *the sample instants only*. The staircase curves are used to illustrate that the discrete signals are assumed constant between the sample instants.

It should also be noted that the continuous and the discrete responses coincide with each other at the sample instants. This is most clearly seen on the enlarged section of the plot on Fig. 3.9. The two models obviously describe the same system at these time instants.

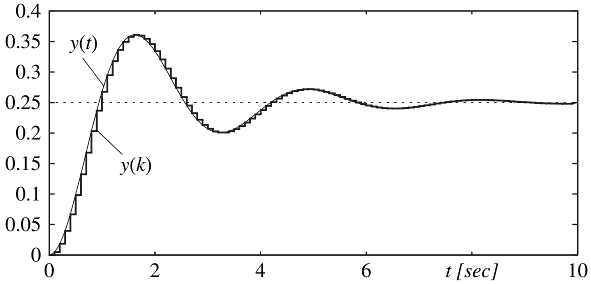


**Fig. 3.8** Continuous and discrete response for $T = 0.1$

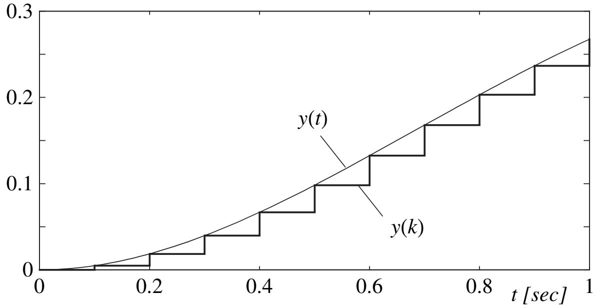


**Fig. 3.9** Enlarged section of Fig. 3.8

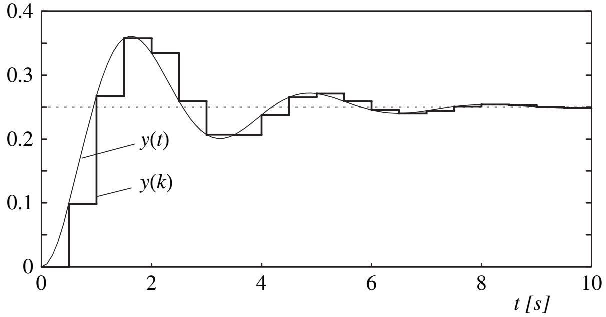


**Fig. 3.10** Response for T = 0.5 s

When generating a discrete time model of a continuous system, one can in principle choose an arbitrary sample period. It should be pointed out, however, that the discrete model's ability to display the original system's properties will deteriorate if $T$ is too large. On Fig. 3.10 the unit step response for $T = 0.5\,\text{s}$ is shown. On this figure the system's oscillations agree reasonably well with those of the continuous system. Figure 3.11 shows the response for $T = 2\,\text{s}$. In this case it is difficult to see the resemblance between the continuous and the discrete responses. Although the discretization is formally correct (note that the responses still coincide at the sample instants), this discrete model is hardly an appropriate tool for further investigations. The sample period $T = 2\,\text{s}$ is more than half the oscillation period of the continuous system (which is 3.14 s) and that is obviously inadequate. An even larger sample period would completely disguise the original system's important oscillatory properties.

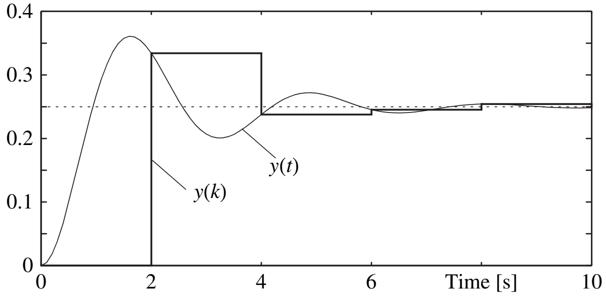


**Fig. 3.11** Response for $T = 2$ s

The transfer function can be found from (3.122). For $T = 0.1$ s one obtains

$$z\mathbf{I} - \mathbf{F} = \begin{bmatrix} z - 0.9807 & -0.09453 \\ 0.3781 & z - 0.8862 \end{bmatrix},\ adj(z\mathbf{I} - \mathbf{F}) = \begin{bmatrix} z - 0.8862 & 0.09453 \\ -0.3781 & z - 0.9807 \end{bmatrix},$$

and

$$det(z\mathbf{I}-\mathbf{F}) = z^2 - 1.8669z + 0.9048.$$

The transfer function becomes

$$\mathbf{H}(z) = \frac{\mathbf{C}adj(z\mathbf{I}-\mathbf{F})\mathbf{G}}{det(z\mathbf{I}-\mathbf{F})} = \frac{0.004821z + 0.004663}{z^2 - 1.8669z + 0.9048}.$$

The natural modes are,

$$m_1 = (0.9335 + j0.1831)^k \quad \text{and} \quad m_2 = (0.9335 - j0.1831)^k,$$

and the corresponding natural frequency and damping ratio are

$$\omega_n = \frac{|\log_e(0.9335 + j0.1831)|}{0.1} = 1.41,$$
$$\zeta = -\cos(\angle(0.9335 + j0.1831)) = 0.707.$$

❐

## 3.6 Similarity Transformations

As it has been shown in Chap. 2, the state space model is not unique. If the state variables are selected in the way as in Examples 2.2 and 2.9, it is possible to generate one of the many possible state models. This ‘natural’ choice of state variables is often quite sensible for all practical purposes in terms of analysis, simulation and design of controllers.

Sometimes it is, however, necessary or practical to change the model to a form different than the one initially selected. In principle it is quite simple to alter the state space model. It is just a matter of changing the choice of state variables. This can be done by *similarity transformations*. If a $n$-dimensional state variable $\mathbf{x}$ is given, one can obtain another simply by multiplying by a nonsingular constant matrix $\mathbf{P}$ of dimension $n \times n$,

$$\mathbf{z} = \mathbf{P}\mathbf{x} \Leftrightarrow \mathbf{x} = \mathbf{P}^{-1}\mathbf{z}. \tag{3.125}$$

The state model for the new state vector is readily found. Whether one considers a continuous or a discrete model:

$$\begin{aligned}&\dot{\mathbf{x}}(t)=\mathbf{Ax}(t)+\mathbf{Bu}(t),\ \mathbf{y}(t)=\mathbf{Cx}(t)+\mathbf{Du}(t),\\&\text{or}\\&\mathbf{x}(k+1)=\mathbf{Fx}(k)+\mathbf{Gu}(k),\ \mathbf{y}(k)=\mathbf{Cx}(k)+\mathbf{Du}(k),\end{aligned}\tag{3.126}$$

one can rewrite the state equations in terms of the new state vector **z** by substituting for **x** in (3.126):

$$\begin{aligned}&\mathbf{P}^{-1}\dot{\mathbf{z}}(t)=\mathbf{AP}^{-1}\mathbf{z}(t)+\mathbf{Bu}(t),\ \mathbf{y}(t)=\mathbf{CP}^{-1}\mathbf{z}(t)+\mathbf{Du}(t),\\&\text{or}\\&\mathbf{P}^{-1}\mathbf{z}(k+1)=\mathbf{FP}^{-1}\mathbf{z}(k)+\mathbf{Gu}(k),\ \mathbf{y}(k)=\mathbf{CP}^{-1}\mathbf{z}(k)+\mathbf{Du}(k).\end{aligned}\tag{3.127}$$

Premultiplying the first equation in each line by **P** yields the result:

$$\begin{aligned}&\dot{\mathbf{z}}(t)=\mathbf{PAP}^{-1}\mathbf{z}(t)+\mathbf{PBu}(t),\ \mathbf{y}(t)=\mathbf{CP}^{-1}\mathbf{z}(t)+\mathbf{Du}(t),\\&\text{or}\\&\mathbf{z}(k+1)=\mathbf{PFP}^{-1}\mathbf{z}(k)+\mathbf{PGu}(k),\ \mathbf{y}(k)=\mathbf{CP}^{-1}\mathbf{z}(k)+\mathbf{Du}(k).\end{aligned}\tag{3.128}$$

The new equations written in the standard form are

$$\begin{aligned}&\dot{\mathbf{z}}(t)=\mathbf{A}_t\mathbf{z}(t)+\mathbf{B}_t\mathbf{u}(t),\ \mathbf{y}(t)=\mathbf{C}_t\mathbf{z}(t)+\mathbf{D}_t\mathbf{u}(t),\\&\text{or}\\&\mathbf{z}(k+1)=\mathbf{F}_t\mathbf{z}(k)+\mathbf{G}_t\mathbf{u}(k),\ \mathbf{y}(k)=\mathbf{C}_t\mathbf{z}(k)+\mathbf{D}_t\mathbf{u}(k),\end{aligned}\tag{3.129}$$

where

$$\begin{aligned}&\mathbf{A}_t=\mathbf{PAP}^{-1}\quad\text{or}\quad\mathbf{F}_t=\mathbf{PFP}^{-1},\\&\mathbf{B}_t=\mathbf{PB}\quad\text{or}\quad\mathbf{G}_t=\mathbf{PG},\\&\mathbf{C}_t=\mathbf{CP}^{-1},\\&\mathbf{D}_t=\mathbf{D}.\end{aligned}\tag{3.130}$$

Usually well defined physical variables are used when the state space model is formulated, e.g. by using the ‘natural’ state variables in the examples of Chap. 2. Applying the transformation (3.125) a new set of state variables is obtained. These will in general be linear combinations of the original state variables and consequently they will usually not be quantities with a specific physical meaning. This is sometimes taken as a disadvantage of the similarity transformation technique, since the person making and working with the model seems to lose the immediate ‘feel’ for what is going on in the model.

However, as will be apparent later, transformation can provide considerable benefits when it is applied.

As will be seen here and in what follows, the transformed system does retain the essential dynamic properties of the original system. In Sects. 3.2.1 and 3.5 the eigenvalues and the corresponding natural modes were found to be of great importance for the system's behaviour. Eigenvalues for the system matrix are found from the characteristic equation,

$$det(\lambda \mathbf{I} - \mathbf{A}) = 0, \tag{3.131}$$

or from the analogous equation involving $\mathbf{F}$. Writing the characteristic equation for the transformed system leads to

$$\begin{aligned} det(\lambda_t \mathbf{I} - \mathbf{A}_t) &= det(\lambda_t \mathbf{P}\mathbf{P}^{-1} - \mathbf{P}\mathbf{A}\mathbf{P}^{-1}) = det(\mathbf{P}(\lambda_t \mathbf{I} - \mathbf{A})\mathbf{P}^{-1}) \\ &= det(\mathbf{P}) \cdot det(\lambda_t \mathbf{I} - \mathbf{A}) \cdot det(\mathbf{P}^{-1}) = 0 \\ &\Rightarrow det(\lambda_t \mathbf{I} - \mathbf{A}) = 0. \end{aligned} \tag{3.132}$$

The last implication follows from the fact that $\mathbf{P}$ and $\mathbf{P}^{-1}$ are both nonsingular and hence their determinants are nonzero. Equation (3.132) shows that $\mathbf{A}_t$ and $\mathbf{A}$ have the same eigenvalues.

For the continuous time state transition matrix (3.48) which uses the transformed system matrix one obtains (using the series expansion for the matrix exponential)

$$\begin{aligned} \phi_t(t) = e^{\mathbf{A}_t t} = e^{\mathbf{P}\mathbf{A}\mathbf{P}^{-1}t} &= \mathbf{I} + \mathbf{P}\mathbf{A}\mathbf{P}^{-1}t + \frac{1}{2}\mathbf{P}\mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{A}\mathbf{P}^{-1}t^2 + \dots \\ &= \mathbf{P}\left(\mathbf{I} + \mathbf{A}t + \frac{1}{2}\mathbf{A}^2 t^2 + \dots\right)\mathbf{P}^{-1} = \mathbf{P}e^{\mathbf{A}t}\mathbf{P}^{-1} = \mathbf{P}\phi(t)\mathbf{P}^{-1}. \end{aligned} \tag{3.133}$$

Similarly for the discrete time transition matrix (3.99) it is clear that

$$\phi_t(k) = \mathbf{F}_t^k = (\mathbf{P}\mathbf{E}\mathbf{P}^{-1})^k = \underbrace{(\mathbf{P}\mathbf{F}\mathbf{P}^{-1})(\mathbf{P}\mathbf{F}\mathbf{P}^{-1})\dots}_{k \text{ factors}} = \mathbf{P}\mathbf{F}^k\mathbf{P}^{-1} = \mathbf{P}\phi(k)\mathbf{P}^{-1}. \tag{3.134}$$

### Diagonal transformation

Every constant quadratic matrix with distinct eigenvalues has a full set of linearly independent eigenvectors. Denoting the eigenvectors $\mathbf{v}_i$, the so called *modal matrix,*

$$\mathbf{M} = [\, \mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n \,], \tag{3.135}$$

is nonsingular.

For matrices with multiple eigenvalues it may be found that some eigenvectors are linearly dependent and consequently the modal matrix is singular. Matrices with this peculiarity are called *defective*.

From the expression defining eigenvectors with the eigenvalue equation,

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i \ , \ \mathbf{v}_i \neq 0 \ , \ i = 1, \ldots, n, \tag{3.136}$$

it is seen that

$$\mathbf{A}[\mathbf{v}_1 \ \mathbf{v}_2 \ \ldots \ \mathbf{v}_n] = [\mathbf{v}_1 \ \mathbf{v}_2 \ \ldots \ \mathbf{v}_n] \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}. \tag{3.137}$$

Designating the diagonal matrix as $\Lambda$,

$$\mathbf{AM} = \mathbf{M}\Lambda. \tag{3.138}$$

Assuming that **A** is non-defective one observes that

$$\mathbf{A} = \mathbf{M}\Lambda\mathbf{M}^{-1} \Leftrightarrow \Lambda = \mathbf{M}^{-1}\mathbf{AM}. \tag{3.139}$$

Equation (3.139) shows that the system matrix can be transformed to *diagonal form* if the similarity transformation matrix $\mathbf{P} = \mathbf{M}^{-1}$ is used, i.e.,

$$\mathbf{z} = \mathbf{Px} = \mathbf{M}^{-1}\mathbf{x}. \tag{3.140}$$

The systems (3.126) will be transformed to

$$\begin{aligned} &\dot{\mathbf{z}}(t) = \Lambda\mathbf{z}(t) + \mathbf{B}_t\mathbf{u}(t), \ \mathbf{y}(t) = \mathbf{C}_t\mathbf{z}(t) + \mathbf{D}_t\mathbf{u}(t), \\ &\text{or} \\ &\mathbf{z}(k+1) = \Lambda\mathbf{z}(k) + \mathbf{G}_t\mathbf{u}(k), \ \mathbf{y}(k) = \mathbf{C}_t\mathbf{z}(k) + \mathbf{D}_t\mathbf{u}(k), \end{aligned} \tag{3.141}$$

where

$$\begin{aligned} &\Lambda = \mathbf{M}^{-1}\mathbf{AM} \ \text{ or } \ \Lambda = \mathbf{M}^{-1}\mathbf{FM}, \\ &\mathbf{B}_t = \mathbf{M}^{-1}\mathbf{B} \ \text{ or } \ \mathbf{G}_t = \mathbf{M}^{-1}\mathbf{G}, \\ &\mathbf{C}_t = \mathbf{CM}, \\ &\mathbf{D}_t = \mathbf{D}. \end{aligned} \tag{3.142}$$

The continuous state equation in (3.141) can be written

$$\dot{\mathbf{z}} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ : & : & : & : \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} \mathbf{z} + \mathbf{B}_t\mathbf{u}\,. \tag{3.143}$$

For the *i*'th state variable one has,

$$\dot{z}_i = \lambda_i z_i + \sum_{j=1}^{n} b_{ij} u_j\,, \tag{3.144}$$

which shows that the states are completely decoupled from each other. For this reason the diagonal form is especially convenient for system analysis purposes.

If the system matrix has repeated eigenvalues and is defective the diagonal transformation is not possible. Based on the notion of generalized eigenvectors it can be shown however (see Sect. B.2 of Appendix B) that a non-singular transformation matrix can be generated which is close to the diagonal form in this case. If the system matrix has *l* distinct eigenvalues, the transformed system matrix will have the form

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & 0 & \dots & 0 \\ 0 & \mathbf{J}_2 & \dots & 0 \\ : & : & : & : \\ 0 & 0 & 0 & \mathbf{J}_l \end{bmatrix} \tag{3.145}$$

where

$$\mathbf{J}_i = \begin{bmatrix} \lambda_i & * & 0 & \dots & 0 \\ 0 & \lambda_i & * & \dots & 0 \\ : & : & : & : & 0 \\ 0 & 0 & 0 & \lambda_i & * \\ 0 & 0 & 0 & 0 & \lambda_i \end{bmatrix}_{m_i \times m_i}. \tag{3.146}$$

The * in the superdiagonal can be 0 or 1.

The state space model with the system matrix (3.145) is called the *Jordan normal form* or the *modified diagonal form*. The submatrices $\mathbf{J}_1$ are called *Jordan* *blocks*. They have the dimension $m_i \times m_i$ where $m_i$ is the number of times the i'th eigenvalue occurs.

***Example 3.7*. Modal Matrix of a Continuous System**

The following system is given:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & -1 \\ -6 & -11 & 6 \\ -6 & -11 & 5 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} u, \quad .$$

$$y = [\,0\ 0\ 1\,]$$

The characteristic equation is:

$$det(\lambda \mathbf{I} - \mathbf{A}) = det \begin{bmatrix} \lambda & -1 & 1 \\ 6 & \lambda + 11 & -6 \\ 6 & 11 & \lambda - 5 \end{bmatrix} = 0,$$

or

$$\lambda(\lambda + 11)(\lambda - 5) + 36 + 66 - 6(\lambda + 11) + 6(\lambda - 5) + 66\lambda = 0$$
$$\Rightarrow \lambda^3 + 6\lambda^2 + 11\lambda + 6 = 0.$$

This equation has the solutions

$$\lambda = \begin{cases} -1 \\ -2 \\ -3. \end{cases}$$

The eigenvector $\mathbf{v}_1$ for $\lambda_1 = -1$ can be found by and inserting in (3.136)

$$\begin{bmatrix} 0 & 1 & -1 \\ -6 & -11 & 6 \\ -6 & -11 & 5 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{21} \\ v_{31} \end{bmatrix} = -1 \cdot \begin{bmatrix} v_{11} \\ v_{21} \\ v_{31} \end{bmatrix}$$

or

$$v_{11} + v_{21} - v_{31} = 0,$$
$$6v_{11} + 10v_{21} - 6v_{31} = 0, .$$
$$6v_{11} + 11v_{21} - 6v_{31} = 0.$$

From these equations clearly

$$v_{11} = v_{31} \ \text{ and } \ v_{21} = 0,$$

so a possible solution for the first eigenvector is

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Similarly one finds that

$$\mathbf{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} \text{ and } \ \mathbf{v}_3 = \begin{bmatrix} 1 \\ 6 \\ 9 \end{bmatrix}.$$

The modal matrix and its inverse become,

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 6 \\ 1 & 4 & 9 \end{bmatrix}, \ \mathbf{M}^{-1} = \begin{bmatrix} 3 & \frac{5}{2} & -2 \\ -3 & -4 & 3 \\ 1 & \frac{3}{2} & -1 \end{bmatrix},$$

and with the similarity transformation $\mathbf{z} = \mathbf{M}^{-1}\mathbf{x}$ the new state model can be computed

$$\dot{\mathbf{z}} = \Lambda z + \mathbf{B}_t u, \ y = \mathbf{C}_t \mathbf{z}.$$

where

$$\Lambda = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix}, \ \mathbf{B}_t = \mathbf{M}^{-1}\mathbf{B} = \begin{bmatrix} -1 \\ 3 \\ -1 \end{bmatrix}, \ \mathbf{C}_t = \mathbf{CM} = [1 \ 4 \ 9].$$

Using $\mathbf{M}^{-1}$ as a transformation matrix implies that the new set of state variables is

$$\begin{aligned} z_1 &= 3x_1 + \frac{5}{2}x_2 - 2x_3, \\ z_2 &= -3x_1 - 4x_2 + 3x_3, \\ z_3 &= x_1 + \frac{3}{2}x_2 - x_3. \end{aligned}$$

Since the two state models above are models of the same physical system, they will be equivalent to the same transfer function. This can be verified using Eq. (3.71). In both cases the transfer function is

$$G(s) = \frac{2s^2 + 16s + 12}{s^3 + 6s^2 + 11s + 6} = \frac{2(s + 7.162)(s + 0.8277)}{(s + 1)(s + 2)(s + 3)}.$$

□

***Example 3.8*. Modal Matrix of a Discrete System**

Discretizing the continuous system,

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0.5 \\ 1 & 0 & 0 \\ -8 & -4 & -1 \end{bmatrix}, \; \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix}, \; \mathbf{C} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \; \mathbf{D} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

with the sample interval $T = 1$ s, the following matrices for the discrete time system are found,

$$\mathbf{F} = \begin{bmatrix} -0.2468 & -0.5153 & 0.1246 \\ 0.5069 & 0.7838 & 0.1288 \\ -3.0245 & -0.9970 & -0.4960 \end{bmatrix}, \; \mathbf{G} = \begin{bmatrix} 0.5069 & -0.4325 \\ 0.3658 & 1.8787 \\ -2.4936 & -2.0611 \end{bmatrix}.$$

Here the MATLAB `c2d`-function has been used.

The eigenvalues of **A** and **F** are

$$\lambda_{\mathbf{A}} = \begin{cases} -0.2334 \pm j1.9227 \\ -0.5332 \end{cases},$$

$$\lambda_{\mathbf{F}} = e^{\lambda_{\mathbf{A}} \cdot 1} = \begin{cases} -0.2729 \pm j0.7433 \\ 0.5867 \end{cases}.$$

The eigenvectors for complex eigenvalues will in general have complex entries and so will the modal matrix for **F**. To avoid the slightly cumbersome manipulations with complex numbers, MATLAB will be used to compute the modal matrix. The command,

$$\texttt{[M,E]} = \texttt{eig(F)},$$

provides **M** as well as with the matrix **E** (which is diagonal with the eigenvalues in the diagonal),

$$\mathbf{M} = \begin{bmatrix} 0.1578 + j0.1912 & 0.1578 - j0.1912 & -0.4205 \\ 0.08816 - j0.09280 & 0.08816 + j0.09280 & 0.7887 \\ -0.8088 + j0.5177 & -0.8088 - j0.5177 & 0.4484 \end{bmatrix}.$$

Note that MATLAB normalizes the eigenvectors, i.e., all columns in **M** have unity 2-norm,

$$\|\mathbf{v}_i\|_2 = \sqrt{\sum_{j=1}^{n} |v_{ji}|^2} = 1,$$

which can easily be verified.

Diagonal-transformation this system using (3.142) one finds that

$$\Lambda = \begin{bmatrix} -0.2729 + j0.7433 & 0 & 0 \\ 0 & -0.2729 - j0.7433 & 0 \\ 0 & 0 & 0.5867 \end{bmatrix},$$

$$\mathbf{G}_t = \begin{bmatrix} 1.6415 - j0.1037 & 1.6099 + j0.3188 \\ 1.6415 + j0.1037 & 1.6099 - j0.3188 \\ 0.1212 & 1.9471 \end{bmatrix},$$

$$\mathbf{C}_t = \begin{bmatrix} -0.8088 + j0.5177 & -0.8088 - j0.5177 & 0.4484 \\ 0.1578 + j0.1912 & 0.1578 - j0.1912 & -0.4205 \end{bmatrix}.$$

Although the states of this state space model are decoupled, this is not necessarily convenient for practical computations. However the transfer functions will of course still have real polynomials. For this MIMO system with two inputs and two outputs there are four transfer functions,

$$\mathbf{Y}(z) = \mathbf{H}(z)\mathbf{U}(z),$$

where

$$\mathbf{H}(z) = \begin{bmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \end{bmatrix}.$$

The transfer function matrix $\mathbf{H}(s)$ can be found from Eq. (3.122) based on the matrices above. The MATLAB command is for input number 1 (see the last right hand side argument),

```
[num, den] = ss2tf(Lambda, Gt, Ct, D, 1).
```

The denominator polynomial is the same for all four transfer functions,

$$A(z) = z^3 - 0.04094z^2 + 0.3068z - 0.3679.$$

The numerator polynomials are

$$\begin{aligned}
B_{11}(z) &= -2.4936z^2 - 0.5587z + 1.2564,\\
B_{12}(z) &= -2.0611z^2 + 0.5418z + 1.5193,\\
B_{21}(z) &= 0.5069z^2 - 0.6451z + 0.1382,\\
B_{22}(z) &= -0.4325z^2 - 1.1005z - 0.2629.
\end{aligned}$$

❒

## 3.7 Stability

Stability is one of the most important properties of dynamic systems. In most cases the stability of a system is a necessary condition for the practical applicability of the system.

There are several possible definitions of stability. Most of them involve the notion of an *equilibrium point*. For a system governed by the state equation,

$$\begin{aligned}
&\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{v}(t), t)\\
&\text{or}\\
&\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k), \mathbf{u}(k), \mathbf{v}(k), k).
\end{aligned} \tag{3.147}$$

A point $\mathbf{x}_e$ is said to be an equilibrium point or state for the system if the state is $\mathbf{x}_e$ at some initial time $t_0$ or $k_0$ and at all future times in the absence of inputs or disturbances. This means that

$$\begin{aligned}
&\mathbf{0} = \mathbf{f}(\mathbf{x}_e, \mathbf{0}, \mathbf{0}, t) \text{ for } t \geq t_0\\
&\text{or}\\
&\mathbf{x}(k+1) = \mathbf{x}_e = \mathbf{f}(\mathbf{x}_e, \mathbf{0}, \mathbf{0}, k) \text{ for } k \geq k_0.
\end{aligned} \tag{3.148}$$

In other words if the system is positioned in an equilibrium point, it will remain there if not influenced by any input or disturbance. At this early stage of the investigation, it is intuitively obvious that the existence of an equilibrium point does not assure stability. If for instance the simple dynamic system shown on Fig. 3.12 is considered, it is noticed that both balls are in an equilibrium point but only the left ball is in a *stable* equilibrium point.

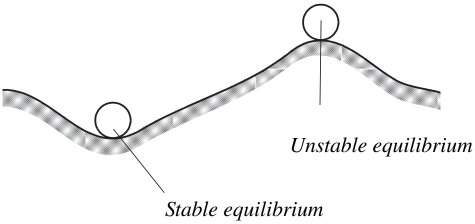


**Fig. 3.12** Stable and unstable equilibrium

Comparing the above definition of an equilibrium point with the definition of a stationary point in Sect. 2.4, it is seen that an equilibrium point is the same as a stationary point in the special case $\mathbf{u}_0 = \mathbf{0}$ and $\mathbf{v}_0 = \mathbf{0}$ (compare Eq. (3.148) with (2.63)).

For the linear systems the equilibrium points can be found by solving:

$$\begin{aligned} &\mathbf{0} = \mathbf{A}(t)\mathbf{x}_e \\ &\text{or} \\ &\mathbf{x}_e = \mathbf{F}(k)\mathbf{x}_e. \end{aligned} \tag{3.149}$$

$\mathbf{x}_e = \mathbf{0}$ is always an equilibrium state for the linear system but there may be others. In the continuous case the origin is the *only* equilibrium state if $\mathbf{A}$ is nonsingular for all $t$. Such an equilibrium point is called *isolated*. If $\mathbf{A}$ is singular (which means that it has at least one zero eigenvalue and $det(\mathbf{A}) = 0$ ) there will be infinitely many equilibrium points. Similarly if the discrete system matrix $\mathbf{F}$ has the eigenvalue 1 for some $k$. In these cases Eq. (3.149) can be written

$$\begin{aligned} \mathbf{A}(t)\mathbf{x}_e &= 0 \cdot \mathbf{x}_e, \\ \mathbf{F}(k)\mathbf{x}_e &= 1 \cdot \mathbf{x}_e, \end{aligned}$$

which shows that all the infinitely many eigenvectors $a\mathbf{x}_e$ are equilibrium states.

For nonlinear systems the matter can be much more complicated. In Fig. 3.13 a so-called *phase plane* plot is shown. The curves are possible trajectories for a second order nonlinear system with the states $x_1$ and $x_2$. The time is a parameter along the trajectories. One can see several trajectories for different values of the initial states. The unshaded area is an *unstable region*. Two isolated equilibrium points are apparent: one is stable and one is unstable. Trajectories started near the closed contour in the bottom of the unstable region (in the top of the fourth quadrant) will eventually follow the contour. Such a contour is called a *limit cycle*. If an initial state is in the unshaded region, the trajectory will go towards such a limit cycle or disappear into the infinite distance. Trajectories started in the shaded *stable region* will end at the stable equilibrium point.

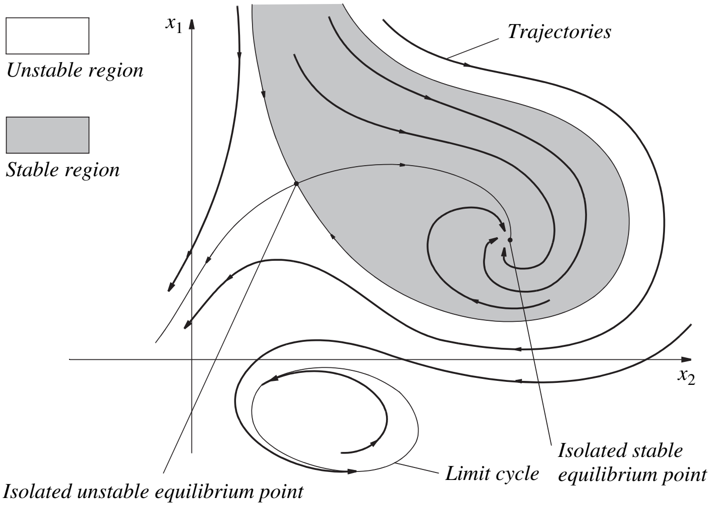

Unstable region
Stable region
$x_1$
$x_2$
Trajectories
Limit cycle
Isolated stable equilibrium point
Isolated unstable equilibrium point

**Fig. 3.13** Stable and unstable regions

From the Figs. 3.12 and 3.13 it is obvious that it is necessary to distinguish between local and global stability. The systems on the figures are stable only for certain values of the states but not for other combinations of the state variables. In such cases one talk about *local stability*. If the entire state space is a stable region, the system is said to be *globally stable*.

Now some more formal definitions of stability will be given. Note that the definitions below are stated for continuous systems but they are also perfectly valid for discrete time systems. It is only a matter of replacing the continuous time $t$ with the discrete time $k$.

### Stability Definition 1

The state $\mathbf{x}_e$ is a stable equilibrium point at the time $t = t_0$ if for any given value $\varepsilon > 0$ there exists a number $\delta(\varepsilon, t_0)$ such that if $\|\mathbf{x}(t_0) - \mathbf{x}_e\| < \delta$, then $\|\mathbf{x}(t) - \mathbf{x}_e\| < \varepsilon$ for all $t \geq t_0$. A system which is stable according to this definition is also called *stable in the sense of Lyapunov* or *stable i.s.L.* For a second order system the definition can be clarified as shown on Fig. 3.14. Stability i.s.L. implies that if the system is started somewhere within the disk with radius $\delta$ then the trajectory will remain within the disk with radius $\varepsilon$ at all times.

If $\delta$ is independent of $t_0$ the system is said to be *uniformly stable i.s.L.*

### Stability Definition 2

The state $\mathbf{x}_e$ is an *asymptotically stable* equilibrium point at the time $t = t_0$ if it is stable i.s.L. *and* if there exists a number $\delta_1(t_0)$ such that if $\|\mathbf{x}(t_0) - \mathbf{x}_e\| < \delta_1$, then $\lim_{t \to \infty} \|\mathbf{x}(t) - \mathbf{x}_e\| = 0$. If $\delta_1$ is independent of $t_0$ the system is *uniformly asymptotically stable*.

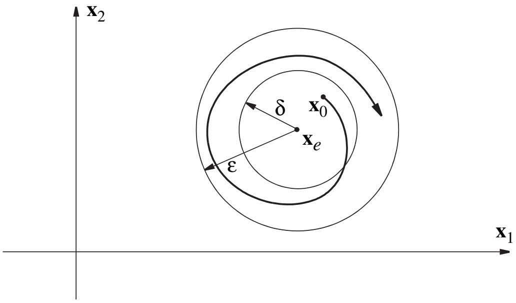


**Fig. 3.14** Stability in the sense of Lyapunov

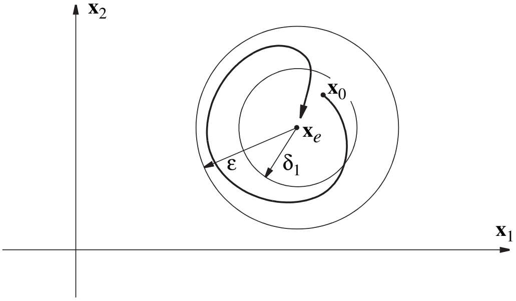


**Fig. 3.15** Asymptotic stability

On Fig. 3.15 is shown what this means for the second order system example. If the systems' initial state is within the disk with radius $\delta_1$ then the trajectory will remain within the disk with radius $\varepsilon$ (because asymptotic stability also implies stability i.s.L.) and the state will tend to the equilibrium state $\mathbf{x}_e$ as $t$ goes to infinity.

The two stability definitions above only depend on the properties of the homogeneous (unforced) state equation. They deal with *zero-input* stability. An alternative definition, which also includes the input, is the following:

**Stability Definition 3**

If any input satisfying the condition $\|\mathbf{u}(t)\| \leq k_1 < \infty$ (i.e. the input is *bounded*) for all $t$ results in an output satisfying the analogous condition $\|\mathbf{y}(t)\| \leq k_2 < \infty$ for all $t$, then the system is *bounded-input-bounded-output stable* or *BIBO stable*.

### *3.7.1 Stability Criteria for Linear Systems*

Based on the definitions above a criteria for stability of linear systems is derived.

Consider the equilibrium point $\mathbf{x}_e = \mathbf{0}$ (the origin).

**Stability Theorem 1A**

The state $\mathbf{x}_e = \mathbf{0}$ of the homogeneous linear system

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) \tag{3.150}$$

is stable i.s.L. if and only if there exists a finite constant $M < \infty$ such that

$$\|\phi(t, t_0)\| \leq M \text{ for all } t_0 \text{ and } t \geq t_0. \tag{3.151}$$

The *matrix norm* will be used here:

$$\|\mathbf{Q}\| = \sup_{\|\mathbf{x}\| \neq 0} \frac{\|\mathbf{Q}\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\|=1} \|\mathbf{Q}\mathbf{x}\|. \tag{3.152}$$

The ‘if’-part (sufficiency) will be proved first. In other words, if (3.151) is true, then it will be proved that the system is stable i.s.L.

The solution of (3.150) is

$$\mathbf{x}(t) = \phi(t, t_0)\mathbf{x}_0. \tag{3.153}$$

From Eq. (3.152)

$$\|\mathbf{Q}\mathbf{x}\| \leq \|\mathbf{Q}\| \cdot \|\mathbf{x}\|. \tag{3.154}$$

From (3.153), (3.154) and (3.151) one finds

$$\|\mathbf{x}(t)\| = \|\phi(t, t_0)\mathbf{x}_0\| \leq \|\phi(t, t_0)\| \cdot \|\mathbf{x}_0\| \leq M \cdot \|\mathbf{x}_0\|. \tag{3.155}$$

If for a given positive $\varepsilon$ one chooses $\delta = \varepsilon/M$, from (3.155) it is clear that

$$\|\mathbf{x}_0\| \leq \delta \Rightarrow \|\mathbf{x}(t)\| \leq M \cdot \frac{\varepsilon}{M} = \varepsilon, \tag{3.156}$$

which shows that the system is stable i.s.L. according to definition 1.

It is also easy to see that condition (3.151) is necessary for stability. Assume that the system is stable i.s.L. but (3.151) does *not* hold, i.e., the norm $\|\phi(t, t_0)\|$ can take arbitrarily large values for some $t$. This means that for given values of $\varepsilon$ and $\delta$ one can be sure that for instance,

$$\|\phi(t, t_0)\| > \frac{\varepsilon}{\delta_1} \quad \text{where} \quad 0 < \delta_1 < \delta. \tag{3.157}$$

Now choose the vector $\mathbf{v}$ with $\|\mathbf{v}\| = 1$ and such that

$$\|\phi(t, t_0)\mathbf{v}\| = \|\phi(t, t_0)\|. \tag{3.158}$$

Furthermore, choose

$$\mathbf{x}_0 = \delta_1 \mathbf{v}. \tag{3.159}$$

This means that

$$\|\mathbf{x}_0\| < \delta \tag{3.160}$$

which is the condition which is desired for the initial state.

But now

$$\begin{aligned} \|\mathbf{x}(t)\| &= \|\phi(t,t_0)\mathbf{x}_0\| = \delta_1\|\phi(t,t_0)\mathbf{v}\| \\ &= \delta_1\|\phi(t,t_0)\| > \delta_1\frac{\varepsilon}{\delta_1} = \varepsilon, \end{aligned} \tag{3.161}$$

which contradicts the assumption that the system is stable i.s.L and the theorem is proven.

Along the same lines one can prove:

**Stability Theorem 1B**

The state $\mathbf{x}_e = \mathbf{0}$ of the homogeneous linear system,

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t), \tag{3.162}$$

is asymptotically stable if and only if there exists a finite constant $M < \infty$ such that,

$$\|\phi(t,t_0)\| \leq M \text{ for all } t_0 \text{ and } t \geq t_0, \tag{3.163}$$

and

$$\lim_{t\to\infty}\|\phi(t,t_0)\| = 0 \text{ for all } t_0. \tag{3.164}$$

The theorems above seem simple and their practical uses are somewhat limited, since they are based on the norm of the state transition matrix for which a closed form analytical expression does not exist in general. However, for time invariant systems, the theorems lead to very useful stability rules.

## *3.7.2 Time Invariant Systems*

The theorems in the last section are valid for continuous as well as for discrete time systems and in the following specific rules will be derived for each of these systems.

**Continuous Time Systems**

For the time invariant system,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t), \tag{3.165}$$

one has (see Eqs. (3.47) and (3.48))

$$\phi(t, t_0) = \phi(t) = e^{\mathbf{A}t}. \tag{3.166}$$

From equation (3.61), if **A** has no repeated eigenvalues, then the norm of $\phi(t)$ can be written

$$\|\phi(t)\| = \left\| \sum_i \mathbf{Z}_i e^{\lambda_i t} \right\| \leq \sum_i |e^{\lambda_i t}| \cdot \|Z_i\|, \tag{3.167}$$

where the following matrix norm property has been used

$$\|\mathbf{A} + \mathbf{B} + \mathbf{C} + \ldots\| \leq \|\mathbf{A}\| + \|\mathbf{B}\| + \|\mathbf{C}\| + \ldots$$

The matrices $\mathbf{Z}_i$ are constant and so are their norms. The eigenvalues are in general complex, $\lambda_{i,i+1} = a \pm jb$ (and of course are real for $b = 0$) and the exponential functions can therefore be written in the form,

$$e^{\lambda_i t} = e^{at} e^{jbt}.$$

If the real part is nonpositive, i.e. $a \leq 0$, then all the exponential functions in (3.167) are bounded and one can choose

$$M = \sup_t \sum_i |e^{\lambda_i t}| \cdot \|\mathbf{Z}_i\| \tag{3.168}$$

and immediately it is clear that the condition of theorem 1A,

$$\|\phi(t)\| < M, \tag{3.169}$$

is fulfilled.

If **A** has repeated eigenvalues, Eq. (3.61) will contain terms of the form $\mathbf{Z}_{ri} t^l e^a t e^{jbt}$, where $l$ is a positive integer. If $a \leq 0$, the result of the procedure above will be unchanged because

$$\lim_{t \to \infty} t^l e^{at} e^{jbt} = 0 \text{ for } a < 0. \tag{3.170}$$

If $a = 0$ this is not true since

$$|t^l e^{jbt}| \to \infty \text{ for } t \to \infty \tag{3.171}$$

and the condition in Eq. (3.151) can obviously not be met.

The result of this investigation is as follows:

- The state $\mathbf{x}_e = \mathbf{0}$ of the time invariant system (3.165) is *stable i.s.L.* if and only if the eigenvalues of the system matrix $\mathbf{A}$ have nonpositive real parts and if eigenvalues on the imaginary axis are simple (nonrepeated).

If the real parts of $\mathbf{A}$'s eigenvalues are strictly negative it is obvious from the development above that the condition for asymptotic stability,

$$\lim_{t\to\infty} \|\phi(t)\| = 0, \tag{3.172}$$

is satisfied and it can be stated that:

- The state $\mathbf{x}_e = \mathbf{0}$ of the time invariant system (3.165) is *asymptotically stable* if and only if the eigenvalues of the system matrix $\mathbf{A}$ have negative real parts.

***Example 3.9.* Stability of a Phase Variable System**

In this example consider a continuous time 4th order system with three different eigenvalue locations,

$$\text{a.}\begin{cases} -1.5 \\ -2 \\ -0.3 \pm j1.5 \end{cases} \quad \text{b.}\begin{cases} 0 \\ -2 \\ -0.3 \pm j1.5 \end{cases} \quad \text{c.}\begin{cases} 0 \\ 0 \\ -0.3 \pm j1.5 \end{cases}$$

The matrices are

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \ \mathbf{C} = [1 \quad 0 \quad 0 \quad 0], \ \mathbf{D} = 0.$$

As a matter of fact only the system matrix $\mathbf{A}$ is important here since the stability will be investigated according to stability definitions 1 and 2.

The characteristic polynomial of $\mathbf{A}$ is

$$P(\mathbf{A}) = \lambda^4 + a_3\lambda^3 + a_2\lambda^2 + a_1\lambda + a_0$$

with the coefficients

| | $a_3$ | $a_2$ | $a_1$ | $a_0$ |
|---|---|---|---|---|
| System a | 4.1 | 7.44 | 9.99 | 7.02 |
| System b | 2.6 | 3.54 | 4.68 | 0 |
| System c | 0.6 | 2.34 | 0 | 0 |

System a has all it's eigenvalues strictly in the left half plane and it is therefore asymptotically stable.

System b is stable i.s.L. because it has an eigenvalue on the imaginary axis.

System c has a double eigenvalue on the imaginary axis and hence it is neither asymptotically stable nor stable i.s.L. It is unstable.

A simulation (numerical solution of the system equations) in system a. for the initial state $\mathbf{x}_0 = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}^T$ and $u(t) = 0$ results in the responses shown on Fig. 3.16. On the left plot the phase plane plot is given for the first two states, which means that $x_2$ is drawn as a function of $x_1$. The right plot shows the usual time response plot of the two states. As expected the states tend to zero in time since the system is asymptotically stable and the only equilibrium point is the zero state (see the remarks following Eq. (3.149)).

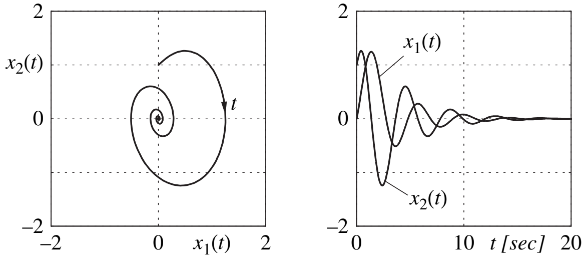


**Fig. 3.16** Responses for system a

The responses for system b. are shown in Fig. 3.17. The system matrix for system b. is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -4.64 & -3.54 & -2.6 \end{bmatrix}.$$

The eigenvector associated with the eigenvalue $\lambda_1 = 0$ is $\mathbf{v}_1 = \begin{bmatrix} q & 0 & 0 & 0 \end{bmatrix}^T$ where $q$ is any number and consequently any initial vector equal to this eigenvector will be an equilibrium point,

$$\mathbf{x}_0 = \mathbf{v}_1 = \mathbf{x}_e.$$

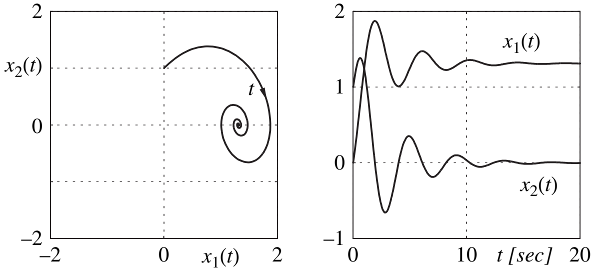


**Fig. 3.17** Responses for system b

If this initial state is chosen for some $q$, the system will remain in that state at all times. However, the initial state selected here is not an equilibrium point, which is clearly seen on Fig. 3.17.

The simulation gives the final state,

$$\mathbf{x}_f = \lim_{t\to\infty} \mathbf{x}(t) = [\,1.312 \quad 0 \quad 0 \quad 0]^T,$$

which is the equilibrium point found above for $q = 1.312$. If it is desired that the system come to rest closer to the origin (i.e., if a smaller value of ε is required), this can be achieved by choosing an initial state closer to the origin (picking a smaller δ). As a matter of fact ε can be chosen arbitrarily small. This is in agreement with stability definition 1: The system is stable i.s.L.

From the plots for system c. (Fig. 3.18) it is obvious that the system is unstable. The phase plane plot as well as the plot of $x_1(t)$ vs. $t$ goes to infinity with time. This is also in agreement with the stability criterion above. One or more eigenvalues in the right half plane would give a similar result. ❐

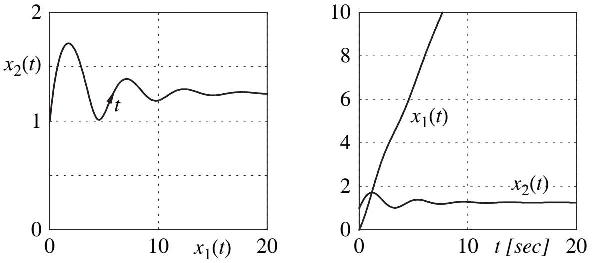


**Fig. 3.18** *Responses for system c*

### *Example 3.10.* Eigenvalues of a Robot Arm

The two-link robot in Example 2.10 is a nonlinear system but earlier a linearized model for a specific pair of link angles was derived. It was found that the eigenvalues of the 4. order system matrix with the link angles $\theta_1 = 45^\circ$ and $\theta_2 = -30^\circ$ were

$$\lambda_{\mathbf{A}} = \begin{cases} \pm 4.933 \\ \pm 1.988 \end{cases} .$$

Linearizing the system for other angle combinations, other quite different sets of eigenvalues would be found. Fixing the angle $\theta_2 = -30^\circ$ and varying $\theta_1$, the results as shown on Fig. 3.19 can be obtained. Apparently the system is unstable for the two positive values of the angle $\theta_1$ since it has two eigenvalues in the right half plane. For the two negative values of $\theta_1$, the picture changes completely. The eigenvalues are now purely imaginary (and nonrepeated) and the system will be stable i.s.L. in these configurations.

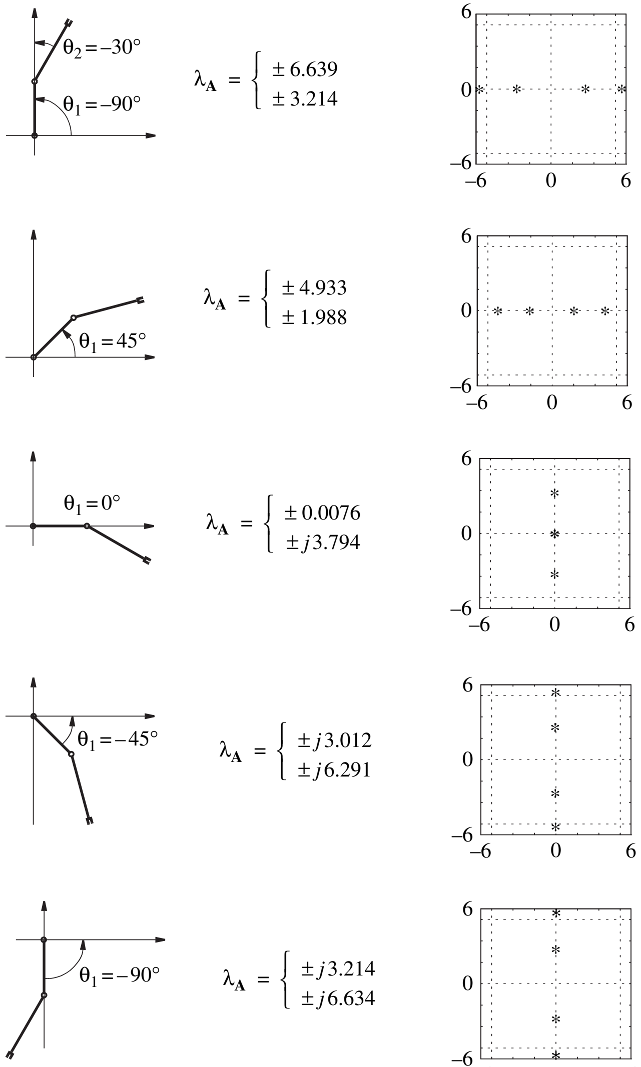


**Fig. 3.19** *Eigenvalues for two-link robot*

It is obvious that the system's properties will change considerably when the angle configuration and therefore also the basis for the linearization is altered. This is typical of nonlinear systems. ❒

**Discrete Time Systems**

For the discrete time system which is time invariant,

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k), \tag{3.173}$$

the solution (3.98) was found earlier with the state transition matrix given by (3.119).

If **F** has only simple eigenvalues then

$$\phi(k) = \mathbf{F}^k = \sum_i \mathbf{Z}_i \lambda_i^k. \tag{3.174}$$

For $\lambda_{i,i+1}$ a complex conjugate pole pair one can write,

$$\lambda_i = a + jb = |\lambda_i| e^{j\angle\lambda_i}, \tag{3.175}$$

thus

$$\|\phi(k)\| = \left\| \sum_i \mathbf{Z}_i \lambda_i^k \right\| \leq \sum_i |\lambda_i|^k \cdot \|\mathbf{Z}_i\|. \tag{3.176}$$

If the magnitude of all the eigenvalues has the property,

$$|\lambda_i| \leq 1, \tag{3.177}$$

all terms in (3.176) will be bounded. Setting

$$M = \sup_k \sum_i |\lambda_i|^k \cdot \|\mathbf{Z}_i\| \tag{3.178}$$

one finds that

$$\|\phi(k)\| < M. \tag{3.179}$$

From (3.119) it is clear that this is still true even if **F** has repeated poles provided they have magnitude less than 1 because the terms originating from repeated eigenvalues will disappear with time:

$$\lim_{k \to \infty} k^l |\lambda_r|^k = 0. \tag{3.180}$$

Only if $|\lambda_r| = 1$ for some repeated eigenvalue will there be difficulties because in that case the term will go to infinity with time,

$$|k^l| \to \infty \text{ for } k \to \infty. \tag{3.181}$$

It can be concluded that

- The state $\mathbf{x}_e = \mathbf{0}$ of the time invariant system, Eq. (3.173), is *stable i.s.L.* if and only if the eigenvalues of the system matrix **F** are not located outside the unit circle and if eigenvalues on the unit circle are simple.

If **F**'s eigenvalues are all strictly inside the unit circle, then it is certain that

$$\lim_{k\to\infty} \|\phi(k)\| = 0 \tag{3.182}$$

and hence:

- The state $\mathbf{x}_e = \mathbf{0}$ of the time invariant system (3.173) is *asymptotically stable* if and only if the eigenvalues of the system matrix **F** are strictly within the unit circle.

***Example 3.11.* Eigenfrequencies of a Time Varying System**

Examples have been found which show that the stability rules involving the eigenvalues of *time invariant* systems should not be applied to other classes of systems. One such example is the following:

Consider the continuous time, time varying and unforced system,

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} -1+\alpha\cos^2(t) & 1-\alpha\sin(t)\cos(t) \\ -1-\alpha\sin(t)\cos(t) & -1+\alpha\sin^2(t) \end{bmatrix}\mathbf{x}(t).$$

The eigenvalues are obtained from the characteristic equation,

$$\lambda^2 + (2-\alpha)\lambda + 2 - \alpha = 0,$$

and the result is

$$\lambda = -\frac{2-\alpha}{2} \pm \sqrt{\frac{(2-\alpha)^2}{4} + \alpha - 2}.$$

It can be seen that the eigenvalues are real and negative or have negative real parts for $\alpha<2$.

The solution of the state equation above for $t_0 = 0$ is

$$\mathbf{x}(t) = \phi(t,0)\mathbf{x}_0$$

where

$$\phi(t,0) = \begin{bmatrix} e^{(\alpha-1)t}\cos(t) & e^{-t}\sin(t) \\ -e^{(\alpha-1)t}\sin(t) & e^{-t}\cos(t) \end{bmatrix}.$$

This can be seen by direct substitution in the state equation.

If $\alpha > 1$ the elements in the first column of $\phi(t,0)$ will clearly take on arbitrarily large values as $t \to \infty$ and hence the system is unstable even though its' eigenvalues are strictly in the left half plane for $\alpha > 2$. ❐

The results from this and similar examples should not lead to the presumption that stability evaluations based on eigenvalues always will give wrong results for time varying systems. As a matter of fact, the eigenvalues will often give reliable information on stability also in these cases. This subject will not be pursued further here but the interested reader is referred to more specialized textbooks such as Rugh (1996).

### *3.7.3 BIBO Stability*

According to stability definition 3 on p. 104 the *input-output stability* or *BIBO stability* problem can be treated in a way similar to the development in Sect. 3.7.1 and 3.7.2. In this case the derivations will be omitted and only the final results stated. For a more detailed discussion the reader is again referred to Rugh (1996).

Note that in the stability theorem below the influence of the **D**-term of the output eqs. (3.25) or (3.88) is not mentioned. It is obvious, that if **D** is bounded it will not affect the boundedness of the output and if it is not bounded, then the output is not bounded either. The theorem is based on the unit impulse response (3.28) or, in the discrete time case, the unit pulse response (3.93).

**Stability Theorem 2**

The continuous time linear state equation (3.8) and (3.25) is BIBO-stable if and only if there exists a finite constant $c$ such that the unit impulse response satisfies the condition:

$$\int_{t_0}^{t} \|\mathbf{g}(t,\tau)\| d\tau \leq c \text{ for } t \geq t_0. \tag{3.183}$$

The discrete time linear state equation (3.88) is BIBO-stable if and only if there exist a finite constant $c$ such that the unit pulse response satisfies the condition:

$$\sum_{i=k_0}^{k-1} \|\mathbf{g}(k,i)\| \leq c \quad \text{for} \quad k \geq k_0. \tag{3.184}$$

Note, that for linear systems in general, BIBO-stability and zero-input stability of the origin as discussed in Sect. 3.7.1 are *not* necessarily equivalent.

**Time Invariant Systems**

- The continuous time LTI system (3.37) and (3.53) is BIBO-stable if and only if the poles of the transfer function matrix (3.71) are strictly in the left half plane
- The discrete time LTI system (3.96) is BIBO-stable if and only if the poles of the transfer function matrix (3.122) are strictly within the unit circle.

As can be judged from the above stability rules, zero-input stability and BIBO-stability are equivalent for LTI-systems if the systems' eigenvalues and poles are the same. This is not always the case and therefore one must distinguish between *internal* and *external stability*.

### *3.7.4 Internal and External Stability*

The state space model is an internal model and the stability investigation based on the state model will reveal whether or not the system is *internally stable*. For LTI systems this property is determined by the eigenvalues of the system matrix.

If the model is represented by a transfer function matrix (Eq. (3.71) or (3.122)), then one is dealing with an external model and the placement of the poles gives information on the system's *external stability*. If an unstable eigenvalue/pole is cancelled by a zero, the system may be externally stable although it is internally unstable. Conversely, an internally stable LTI-system will also be externally stable.

***Example 3.12*. Externally Stable and Internally Unstable System**

Consider the SISO continuous time LTI system,

$$\dot{\mathbf{x}} = \begin{bmatrix} -3 & 4 & 12 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u,$$

$$y = [1 \ \ -1 \ \ -2]\mathbf{x}.$$

The eigenvalues which are found from the expression

$$det \begin{bmatrix} \lambda + 3 & -4 & -12 \\ -1 & \lambda & 0 \\ 0 & -1 & \lambda \end{bmatrix} = 0$$

are

$$\lambda = \begin{cases} -3 \\ -2 \\ 2 \end{cases}$$

and it is obvious that the system is not internally (asymptotically) stable.

The external model is the transfer function

$$\frac{y(s)}{u(s)} = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \frac{s^2 - s - 2}{s^3 + 3s^2 - 4s - 12}$$

$$= \frac{(s+1)(s-2)}{(s+3)(s+2)(s-2)} = \frac{(s+1)}{(s+3)(s+2)}.$$

Thanks to the cancellation of the pole-zero pair, the systems poles are

$$s = \begin{cases} -3 \\ -2 \end{cases}$$

and the system is externally stable.

If $u(t)$ is a unit step applied to the system and the response simulated the results on Figs. 3.20 and 3.21 are obtained. All three states go to infinity for $t \to \infty$ but the output $y(t) = x_1(t) - x_2(t) - 2x_3(t)$ remains finite at all times.

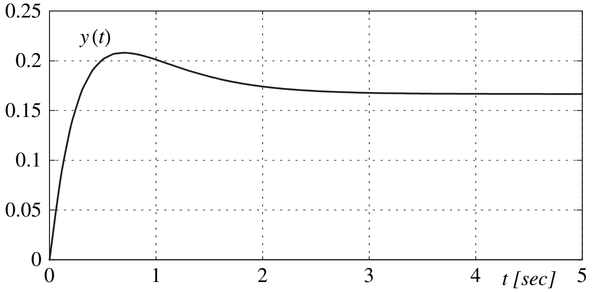


**Fig. 3.20** Output of externally stable system

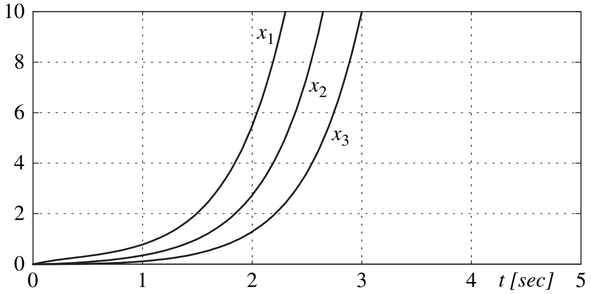


**Fig. 3.21** States of internally unstable system

Consequently it is reasonable to state that the system is internally unstable but externally stable. In spite of this interesting result, one should keep in mind that exact cancellation of pole-zero pairs is more often found in textbook examples than in real practical systems.

With matrix entries differing slightly from the integers in this example one might see zeroes and poles very close to each other but not exactly equal. In that case the unstable eigenvalue would remain as a pole in the transfer function and the response $y(t)$ would contain the corresponding unstable natural mode although possibly with a very small weight factor. The first 5 s of the response could very well look almost like that in Fig. 3.20 but at some later time it would diverge and go to infinity. ❐

### 3.7.5 Lyapunov's Method

Except for LTI-systems, stability analysis can be quite difficult. As indicated above, there are several definitions of stability and a variety of stability criteria for different classes of systems can be found. The Russian mathematician Lyapunov formulated one of the most general stability criteria known. Only a brief overview of the *Lyapunov's second* (or *direct*) *method* will be given here. For more details please refer to Vidyasagar (1978) or to Middleton and Goodwin (1990).

The Lyapunov method is based on the notion of *positive definite function*s of the state vector $\mathbf{x}$. The continuous scalar function $W(\mathbf{x})$ is positive definite if and only if

1. $W(\mathbf{0}) = 0$,
2. $W(\mathbf{x}) > 0$ for all nonzero $\mathbf{x}$ and
3. $W(\mathbf{x}) \to \infty$ for $\|\mathbf{x}\| \to \infty$.

The scalar time dependent functions $V(\mathbf{x}, t)$ and $V(\mathbf{x}, k)$ are positive definite if $W(\mathbf{x})$ is positive definite and if and only if

$V(\mathbf{x}, t) \geq W(\mathbf{x})$ for all $\mathbf{x}$ and $t \geq t_0$ (continuous time)

or

$V(\mathbf{x}, k) \geq W(\mathbf{x})$ for all $\mathbf{x}$ and $k \geq k_0$ (discrete time).

**Stability Theorem 3A**

The systems,

$$\begin{gathered} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t) \\ \text{or} \\ \mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k), k) \end{gathered} \tag{3.185}$$

are stable i.s.L. if $V(\mathbf{x}, t)$ or $V(\mathbf{x}, k)$ is a positive definite function and if and only if

$$\begin{gathered} \frac{d}{dt} V(\mathbf{x}, t) \leq 0 \\ \text{or} \\ V(\mathbf{x}(k+1), k+1) - V(\mathbf{x}(k), k) \leq 0. \end{gathered} \tag{3.186}$$

**Stability Theorem 3B**

The systems (3.185) are asymptotically stable if $V(\mathbf{x}, t)$ or $V(\mathbf{x}, k)$ is a positive definite function and if and only if

$$\begin{gathered} \frac{d}{dt} V(\mathbf{x}, t) < 0 \\ \text{or} \\ V(\mathbf{x}(k+1), k+1) - V(\mathbf{x}(k), k) < 0. \end{gathered} \tag{3.187}$$

The positive definite functions fulfilling the conditions of Eqs. (3.186) or (3.187) are called*Lyapunov functions*. Although these criteria look simple, they are not necessarily easy to use in practical situations. The problem is to find a Lyapunov function for the system at hand. It should be pointed out that lack of success in finding such a function is of course not a proof of instability.

An example of a Lyapunov function with an obvious physical significance is the energy function of a system. The total energy of a system is always positive. If the (unforced) system dissipates energy with time, the total energy will decrease as expressed in (3.186) or (3.187) and the system will be stable.

***Example 3.13*. Stability i.s.L. for a Nonlinear System**

Consider the mechanical system on Fig. 3.22. A block with the mass $M$ is moving on a horizontal surface and it is assumed that the friction between the block and the surface is of the Coulomb type ('dry' friction). The block is connected to the surface frame via the linear spring with the stiffness constant $k$.

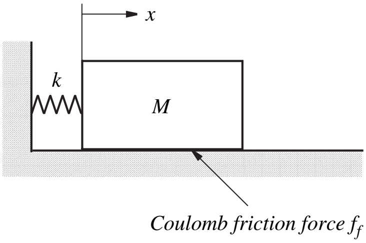


**Fig. 3.22** Mass-spring system with Coulomb-friction

When the block moves, the Coulomb friction force can be written

$$f_f = d \cdot sign(\dot{x}). \tag{3.188}$$

The friction force has the constant numerical value $d$ and it is always directed opposite to the velocity. Note that the *sign*-function is only defined for $\dot{x} \neq 0$. For $\dot{x} = 0$ the friction force can assume any value in the range,

$$-d \leq f_f \leq d. \tag{3.189}$$

The system model can be found simply by applying Newton's second law to the block's mass,

$$M\ddot{x} = -kx - d \cdot sign(\dot{x}), \quad \dot{x} \neq 0. \tag{3.190}$$

If the velocity is zero at $t = 0$, it will remain so for all future times if the spring force is smaller than the maximum friction force $d$ because one has

$$M\ddot{x} = 0, \ \dot{x} = 0, \ |kx| < d. \tag{3.191}$$

If for some t, $\dot{x} = 0$ and $|kx| \geq d$, the friction force will assume its maximum value and in that instant of time Newton's second law will be

$$M\ddot{x} = -kx + d \cdot sign(x). \tag{3.192}$$

The total energy is the sum of the kinetic energy of the block and the potential energy accumulated in the spring.

Define now the state vector,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix}. \tag{3.193}$$

Since the spring force is proportional to the deformation, the expression for the total energy is

$$E(\mathbf{x}) = \frac{1}{2}mx_2^2 + \frac{1}{2}kx_1^2. \tag{3.194}$$

$E(\mathbf{x})$ is quadratic in both state variables so it is found that

$$\begin{aligned} &E(\mathbf{x}) = 0 \text{ for } \mathbf{x} = \mathbf{0}, \\ &E(\mathbf{x}) > 0 \text{ for } \mathbf{x} \neq \mathbf{0}, \end{aligned} \tag{3.195}$$

and $E(\mathbf{x})$ is a positive definite function.

The time derivative of $E(\mathbf{x})$ is seen to be

$$\frac{d}{dt}E(\mathbf{x}) = mx_2\dot{x}_2 + kx_1\dot{x}_1. \tag{3.196}$$

For $x_2 \neq 0$ the following expressions for the state equations are valid,

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= \frac{1}{m}(-kx_1 - d \cdot sign(x_2)), \end{aligned} \tag{3.197}$$

which, inserted into (3.196), leads to

$$\frac{d}{dt}E(\mathbf{x}) = -dx_2 \cdot sign(x_2) < 0 \text{ for all } x_2 \neq 0. \tag{3.198}$$

For $x_2 = 0$ one has

$$\frac{d}{dt}E(\mathbf{x}) = 0. \tag{3.199}$$

Equations (3.198) and (3.199) can be combined to show that

$$\frac{d}{dt}E(\mathbf{x}) \leq 0. \tag{3.200}$$

This means, according to theorem 3A above, that the system is stable i.s.L.

Note that this system has infinitely many equilibrium points, namely all points in the state plane (the system is second order) where

$$x_2 = 0 \text{ and } -\frac{d}{k} \leq x_1 \leq \frac{d}{k}.$$

❐

## Lyapunov's Direct Method for LTI Systems

For the homogeneous LTI system,

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t),\\ \mathbf{x}(k{+}1) &= \mathbf{F}\mathbf{x}(k),\end{aligned} \tag{3.201}$$

a *quadratic form* can be formed by choosing a symmetric, constant matrix **P**,

$$V(\mathbf{x}) = \mathbf{x}^T\mathbf{P}\mathbf{x}. \tag{3.202}$$

If **P** is positive definite, Eq. (3.202) is a Lyapunov function *candidate*.

For the continuous and the discrete time cases respectively,

$$\begin{aligned}\frac{d}{dt}V(\mathbf{x}) &= \dot{\mathbf{x}}^T\mathbf{P}\mathbf{x} + \mathbf{x}^T\mathbf{P}\dot{\mathbf{x}},\\ \Delta V = V(\mathbf{x}(k+1)) - V(\mathbf{x}(k)) &= \mathbf{x}^T(k+1)\mathbf{P}\mathbf{x}(k+1) - \mathbf{x}^T(k)\mathbf{P}\mathbf{x}(k).\end{aligned} \tag{3.203}$$

Inserting the expressions (3.201) into (3.203) gives,

$$\begin{aligned}\frac{d}{dt}V(\mathbf{x}) &= \mathbf{x}^T(\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A})\mathbf{x},\\ \Delta V &= \mathbf{x}^T(k)(\mathbf{F}^T\mathbf{P}\mathbf{F} - \mathbf{P})\mathbf{x}(k).\end{aligned} \tag{3.204}$$

Applying theorem 3B it can now be seen that if it is possible to find another positive definite matrix **Q** such that

$$\begin{gathered}\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}\\ \text{or}\\ \mathbf{F}^T\mathbf{P}\mathbf{F} - \mathbf{P} = -\mathbf{Q}\end{gathered} \tag{3.205}$$

then the systems (3.201) are asymptotically stable. These two equations above are called the *Lyapunov equations*.

## 3.8 Controllability and Observability

In the effort to design controllers for dynamic systems one is often confronted with the question whether or not it is possible to give a closed loop system an appropriate set of properties. This question may be divided into several subquestions such as: can a controller be designed which will stabilize an unstable system? Can one be sure that the system state can be forced to achieve any desired value in state space? Can one be sure that no state in the system can achieve undesired values? Is it possible to determine all states in the system from the knowledge available in the measured system output?

Answers to such essential questions can be found by considering a systems *controllability* and *observability*. Before proceeding to formal definitions of these concepts an introductory example will be given.

***Example 3.14***[†]**. Stability with Mixed Controllability/Observability Characteristics**

Consider a fourth order LTI SISO system with the following state space model:

$$\dot{\mathbf{x}} = \begin{bmatrix} 2 & 3 & 2 & 1 \\ -2 & -3 & 0 & 0 \\ -2 & -2 & -4 & 0 \\ -2 & -2 & -2 & -5 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ -2 \\ 2 \\ -1 \end{bmatrix} u, \tag{3.206}$$

$$y = [7\ 6\ 4\ 2]\mathbf{x}.$$

The system transfer function can be found from Eq. (3.71),

$$G(s) = \frac{s^3 + 9s^2 + 26s + 24}{s^4 + 10s^3 + 35s^2 + 50s + 24}. \tag{3.207}$$

The zeroes and the poles can be found from the polynomials of $G(s)$,

$$z_i = \begin{cases} -2 \\ -3 \\ -4 \end{cases}, \ p_i = \begin{cases} -1 \\ -2 \\ -3 \\ -4 \end{cases}. \tag{3.208}$$

[†] Example 3.14 is borrowed from Friedland (1987).

Three of the poles are cancelled by the zeroes and the transfer function is in reality not of fourth order but only first order as

$$G(s) = \frac{1}{s+1}. \tag{3.209}$$

A block diagram of the system is shown on Fig. 3.23. It is not possible to see from the model in Eq. (3.207) or from the block diagram that the system has an order less than 4.

However if the system is transformed into the diagonal form (see p. 94), some insight can be gained into the nature of the problem. Since all eigenvalues are distinct, the modal matrix is regular and the similarity transformation can be carried out without difficulties. With the modal matrix **M** a new state vector will be introduced given by

$$\mathbf{v} = \mathbf{M}^{-1}\mathbf{x}. \tag{3.210}$$

Using MATLAB it is found that

$$\Lambda = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & -4 \end{bmatrix}, \ \mathbf{B}_t = \begin{bmatrix} 1.4142 \\ 0 \\ 2.4495 \\ 0 \end{bmatrix}, \tag{3.211}$$

$$\mathbf{C}_t = [0.7071 \ 0.4082 \quad 0 \quad 0].$$

With this model representation a quite different block diagram on Fig. 3.24 can be obtained. If the transfer function from the two block diagrams are found they give exactly the same result as (3.209). The input-output description is the same.

From Fig. 3.24 one can see that the four states have different status in the system. The state variable $v_1$ can be influenced directly via the input $u$ and it can be observed directly at the output. The state $v_2$ is not coupled to the input and since all state are decoupled from each other, $v_2$ can not be affected by any other state either. $v_2$ can be observed at the output just like $v_1$. With $v_3$ the opposite is the case. It can be influenced from $u$ but can not be 'seen' at the output. The last state can neither be controlled via $u$ nor be observed at $y$.

It seems natural to divide the system into the four first order subsystems visible on Fig. 3.24:

- Subsystem with $v_1$ is *controllable* and *observable,*
- Subsystem with $v_2$ is *not controllable* but *observable,*
- Subsystem with $v_3$ is *controllable* but *not observable,*
- Subsystem with $v_4$ is *not controllable* and *not observable.*

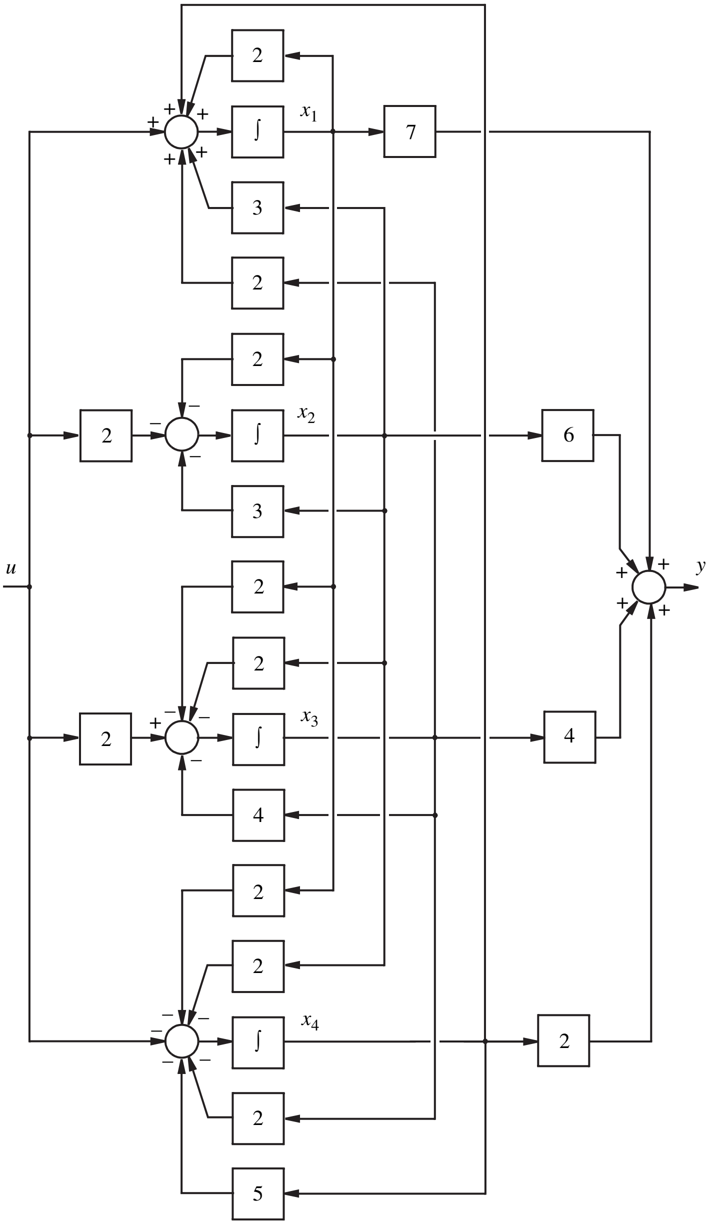


**Fig. 3.23** Block diagram of fourth order SISO system

Only one of the four states is controllable as well as observable and, as will be seen later, this is the reason why the transfer function is in reality only of first order. In the diagonalized system on Fig 3.24 it is obvious what consequences the lack of controllability or observability have. In this case the four eigenvalues are uniquely

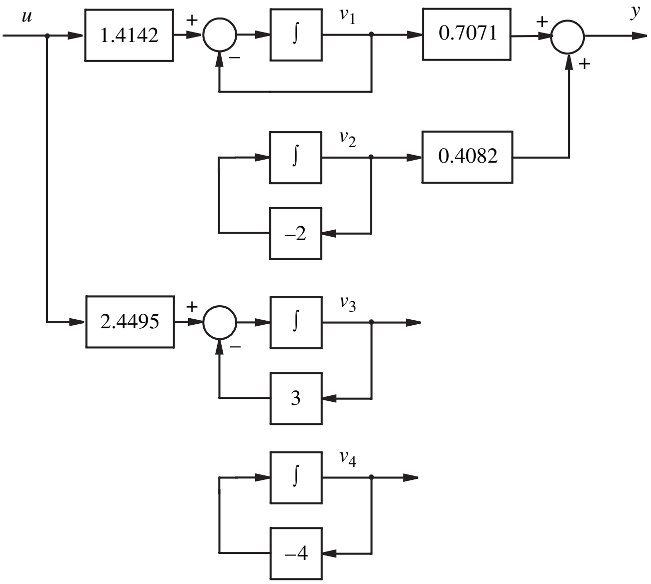


**Fig. 3.24** Block diagram of system on diagonal form

connected to each of the four subsystems. One can see that for instance the subsystem with the eigenvalue $v_2$ can not in any way be affected from the input by which one can hope to control the system and this eigenvalue's influence on the output must be accepted as it is. Similarly the two unobservable subsystems with eigenvalues $v_3$ and $v_4$ will be invisible from outside the system. Since the system is internally (and therefore also externally) stable, one can probably live with this state of affairs. Even if the initial values of the states $v_2$ and $v_4$ are nonzero, they will decay to zero with time and they will do no harm. The state $v_3$ is affected by the input but it will remain within finite limits for all times. ❐

Note that a system where all uncontrollable states are stable is called *stabilizable*, because any unstable state can be affected via the input and therefore it may be stabilized. A system where the unobservable states are stable is said to be *detectable*.

## *3.8.1 Controllability (Continuous Time Systems)*

**Controllability Definition**

The linear system,

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \qquad (3.212)$$

is controllable on the finite time interval $[t_0, t_f]$ if there exists an input $\mathbf{u}(t)$ which will drive the system from any initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ to the zero state $\mathbf{x}(t_f) = \mathbf{0}$.

Controllability is a property connected to the inner structure (i.e., the couplings between the states) and the way the states are coupled to the input variables. Controllability has nothing to do with the system outputs and therefore only the state equation (3.212) is of interest for judging this property. For obvious reasons, controllability defined as above is sometimes called *controllability-to-the-origin*. See also Sect. 3.8.3.

There are several ways in which it is possible to determine whether a system is controllable or not. Some of the criteria which can be used are closely related to the type of direct inspection used in Example 3.14 but others are much more abstract in nature. The study of the matter will start with a criterion of the last category.

**Controllability Theorem CC1**

The linear state space model (3.212) is controllable on $[t_0, t_f]$ if and only if the quadratic $n \times n$ matrix,

$$\mathbf{W}_c(t_0, t_f) = \int_{t_0}^{t_f} \phi(t_0, t)\mathbf{B}(t)\mathbf{B}^T(t)\phi^T(t_0, t)dt, \tag{3.213}$$

is regular. The matrix $\mathbf{W}_c(t_0, t_f)$ is called the *controllability Gramian*. It will first be shown, that the test condition is sufficient, i.e., if the Gramian is regular, then the system is controllable.

Assume that $\mathbf{W}_c(t_0, t_f)$ is regular and for a given arbitrary initial state vector $\mathbf{x}_0$, let the input vector be

$$\mathbf{u}(t) = -\mathbf{B}^T(t)\phi^T(t_0, t)\mathbf{W}_c^{-1}(t_0, t_f)\mathbf{x}_0 \ , \ t \in [t_0, t_f]. \tag{3.214}$$

From Eq. (3.20) it is clear that the solution to the state equation at the final time can be written

$$\begin{aligned} \mathbf{x}(t_f) &= \phi(t_f, t_0)\mathbf{x}_0 + \int_{t_0}^{t_f} \phi(t_f, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau \\ &= \phi(t_f, t_0)\mathbf{x}_0 - \int_{t_0}^{t_f} \phi(t_f, \tau)\mathbf{B}(\tau)\mathbf{B}^T(\tau)\phi^T(t_0, \tau)\mathbf{W}_c^{-1}(t_0, t_f)\mathbf{x}_0 d\tau. \end{aligned} \tag{3.215}$$

Property (3.24) of the state transition matrix means that one can write

$$\phi(t_f, \tau) = \phi(t_f, t_0)\phi(t_0, \tau). \tag{3.216}$$

Using this expression in the integral of (3.215) yields

$$\begin{aligned} \mathbf{x}(t_f) &= \phi(t_f, t_0)\mathbf{x}_0 - \phi(t_f, t_0)\int_{t_0}^{t_f} \phi(t_0, \tau)\mathbf{B}(\tau)\mathbf{B}^T(\tau)\phi^T(t_0, \tau)d\tau\mathbf{W}_c^{-1}(t_0, t_f)\mathbf{x}_0 \\ &= \mathbf{0}. \end{aligned} \tag{3.217}$$

To see that the regularity condition is necessary, it can be assumed that it is not and obtain a contradiction. In other words, suppose that the system is controllable but $\mathbf{W}_c(t_0, t_f)$ is singular.

If the Gramian is singular then there exists a nonzero vector $\mathbf{x}_0$ such that the quadratic form

$$\mathbf{x}_0^T \mathbf{W}_c(t_0, t_f)\mathbf{x}_0 = 0 \tag{3.218}$$

or

$$\mathbf{x}_0^T \mathbf{W}_c(t_0, t_f)\mathbf{x}_0 = \int_{t_0}^{t_f} \mathbf{x}_0^T \phi(t_0, t)\mathbf{B}(t)\mathbf{B}^T(t)\phi^T(t_0, t)\mathbf{x}_0 dt = 0. \tag{3.219}$$

Defining

$$\mathbf{z}(t) = \mathbf{B}^T(t)\phi^T(t_0, t)\mathbf{x}_0, \tag{3.220}$$

the integrand can be written

$$\mathbf{x}_0^T \phi(t_0, t)\mathbf{B}(t)\mathbf{B}^T(t)\phi^T(t_0, t)\mathbf{x}_0 = \mathbf{z}^T(t)\mathbf{z}(t) = \|\mathbf{z}(t)\|^2. \tag{3.221}$$

So the integrand is the square of the norm of a vector and therefore always nonnegative. An integral with a nonnegative integrand can only be zero if the integrand is identically zero over the entire integration interval,

$$\mathbf{x}_0^T \phi(t_0, t)\mathbf{B}(t) = \mathbf{0}\ ,\ \ t \in [t_0, t_f]. \tag{3.222}$$

Since the system is controllable an input vector, $\mathbf{u}(t)$, can be found such that

$$\mathbf{0} = \phi(t_f, t_0)\mathbf{x}_0 + \int_{t_0}^{t_f} \phi(t_f, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau \tag{3.223}$$

or using (3.23) and (3.24),

$$\mathbf{x}_0 = -\int_{t_0}^{t_f} \phi(t_0, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau. \tag{3.224}$$

Premultiplying by $\mathbf{x}_0^T$ and applying (3.222) gives the result,

$$\mathbf{x}_0^T\mathbf{x}_0 = -\int_{t_0}^{t_f} \mathbf{x}_0^T \phi(t_0, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau = 0, \tag{3.225}$$

which contradicts the fact that $\mathbf{x}_0$ is a nonzero vector and the necessity of the condition above has been proved.

The quadratic controllability Gramian is clearly symmetric and furthermore it can be shown that it is in general positive semidefinite. If the system is controllable, the Gramian must be positive definite.

In the LTI case the transition matrix is

$$\phi(t,\tau) = e^{\mathbf{A}(t-\tau)} \tag{3.226}$$

and the controllability Gramian becomes

$$\mathbf{W}_c(t_0,t_f) = \int_{t_0}^{t_f} e^{\mathbf{A}(t_0-t)}\mathbf{B}\mathbf{B}^T e^{\mathbf{A}^T(t_0-t)}dt \tag{3.227}$$

or, if one lets $t_0 = 0$,

$$\mathbf{W}_c(t_f) = \int_0^{t_f} e^{-\mathbf{A}t}\mathbf{B}\mathbf{B}^T e^{-\mathbf{A}^T t}dt. \tag{3.228}$$

Controllability theorem CC1 is not convenient to use for practical controllability tests but fortunately for LTI systems, an alternative criterion is available.

**Controllability Theorem CC2**

The LTI system model,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{3.229}$$

is controllable if and only if the *controllability matrix*,

$$\mathbf{M}_c = \begin{bmatrix} \mathbf{B} & \mathbf{A}\mathbf{B} & \mathbf{A}^2\mathbf{B} & \dots & \mathbf{A}^{n-1}\mathbf{B} \end{bmatrix}, \tag{3.230}$$

has rank $n$, i.e., full rank.

It will now be demonstrated that this rank requirement fails if and only if the controllability Gramian is singular. If $\mathbf{W}_c(t_f)$ is singular the condition, Eq. (3.222), is satisfied. For the LTI system the transition matrix is $\phi(t_0,t) = \phi(t) = e^{\mathbf{A}t}$ and (3.222) becomes

$$\mathbf{z}^T(t) = \mathbf{x}_0^T e^{\mathbf{A}t}\mathbf{B} = \mathbf{0}\ ,\ \ t \in [0,\ t_f]. \tag{3.231}$$

Since this vector is identically zero on the entire time interval, its time derivatives must also be zero:

$$\begin{aligned} d\mathbf{z}^T(t)/dt &= \mathbf{x}_0^T e^{\mathbf{A}t}\mathbf{A}\mathbf{B} = \mathbf{0}, \\ d^2\mathbf{z}^T(t)/dt^2 &= \mathbf{x}_0^T e^{\mathbf{A}t}\mathbf{A}^2\mathbf{B} = \mathbf{0}, \\ &\vdots \\ d^{n-1}\mathbf{z}^T(t)/dt^{n-1} &= \mathbf{x}_0^T e^{\mathbf{A}t}\mathbf{A}^{n-1}\mathbf{B} = \mathbf{0}. \end{aligned} \tag{3.232}$$

This means that

$$\mathbf{x}_0^T e^{\mathbf{A}t}\begin{bmatrix}\mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} & \dots & \mathbf{A}^{n-1}\mathbf{B}\end{bmatrix} = \mathbf{x}_0^T e^{\mathbf{A}t}\mathbf{M}_c = \mathbf{0}. \tag{3.233}$$

$\mathbf{x}_0^T e^{\mathbf{A}t}$ is a nonzero 1 × $n$ vector whose elements are time functions

$$\mathbf{x}_0^T e^{\mathbf{A}t} = [a_1(t)\ a_2(t)\ \dots\ a_n(t)]. \tag{3.234}$$

The $n$ rows are called $\mathbf{M}_c$,

$$\mathbf{M}_c = \begin{bmatrix}\mathbf{r}_1^T \\ \mathbf{r}_2^T \\ \vdots \\ \mathbf{r}_n^T\end{bmatrix}, \tag{3.235}$$

and from (3.233) it is seen that

$$\sum_{i=1}^{n} a_i(t)\mathbf{r}_i^T = \mathbf{0}. \tag{3.236}$$

The last expression shows that the $n$ rows of $\mathbf{M}_c$ are not linearly independent and this means that $\mathbf{M}_c$ has less than full rank.

The 'only if'-part can be proven as follows. Using the series expansion (3.44) of the matrix exponential and additionally the Cayley-Hamilton theorem, $e^{-\mathbf{A}t}$ can be expressed as the finite series

$$e^{-\mathbf{A}t} = \sum_{k=0}^{\infty}\frac{\mathbf{A}^k(-t)^k}{k!} = \sum_{k=0}^{n-1}\alpha_k(t)\mathbf{A}^k. \tag{3.237}$$

Postmultiplying by the $n \times m$-dimensional $\mathbf{B}$-matrix gives the $n \times m$-dimensional matrix,

$$e^{-\mathbf{A}t}\mathbf{B} = \sum_{k=0}^{n-1}\alpha_k(t)\mathbf{A}^k\mathbf{B} = \begin{bmatrix}\mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} & \dots & \mathbf{A}^{n-1}\mathbf{B}\end{bmatrix}\begin{bmatrix}\mathbf{I}_m\alpha_0(t) \\ \mathbf{I}_m\alpha_1(t) \\ \vdots \\ \mathbf{I}_m\alpha_{n-1}(t)\end{bmatrix}, \tag{3.238}$$

where $\mathbf{I}_m$ is the $m$-dimensional identity matrix.

The transpose of this matrix is

$$\mathbf{B}^T e^{-\mathbf{A}^T t} = [\,\mathbf{I}_m\alpha_0(t) \quad \mathbf{I}_m\alpha_1(t) \quad \ldots \quad \mathbf{I}_m\alpha_{n-1}(t)\,] \begin{bmatrix} \mathbf{B}^T \\ \mathbf{B}^T\mathbf{A}^T \\ \vdots \\ \mathbf{B}^T(\mathbf{A}^T)^{n-1} \end{bmatrix}. \tag{3.239}$$

Inserting (3.238) and (3.239) into (3.228) one obtains the result,

$$\begin{aligned} \mathbf{W}_c(t_f) &= \mathbf{M}_c \int_0^{t_f} \begin{bmatrix} \mathbf{I}_m\beta_{11}(t) & \mathbf{I}_m\beta_{12}(t) & \ldots & \mathbf{I}_m\beta_{1n}(t) \\ \mathbf{I}_m\beta_{21}(t) & \mathbf{I}_m\beta_{22}(t) & \ldots & \mathbf{I}_m\beta_{2n}(t) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{I}_m\beta_{n1}(t) & \mathbf{I}_m\beta_{n2}(t) & \ldots & \mathbf{I}_m\beta_{nn}(t) \end{bmatrix} dt\mathbf{M}_c^T \\ &= \mathbf{M}_c\mathbf{Q}\mathbf{M}_c^T, \end{aligned} \tag{3.240}$$

where the β-functions are products of the α-functions above.

This equation shows that even if $\mathbf{Q}$ has full rank (i.e., *nm),* the quadratic $n \times n$ matrix product to the right can not have rank larger that the rank of $\mathbf{M}_c$. In other words, if $\mathbf{M}_c$ has less than full rank, the Gramian will be singular.

Note that from the development above, controllability due to theorem CC2 is independent of the final time $t_f$. This means that if the system (3.229) has a controllability matrix (3.230) of full rank, then the controllability matrix (3.228) will be regular for *any* $t_f$.

In addition to the two controllability theorems above there is a third one which is based on certain properties of the *eigenvectors* of the system matrix **A**.

The usual *i*'th eigenvector corresponding to the *i*'th eigenvalue of **A** is defined by

$$\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i, \ \mathbf{v}_i \neq 0, \ i = 1, \ldots, n. \tag{3.241}$$

Because of the order of multiplication in this equation the eigenvector $\mathbf{v}_i$ is sometimes called the *right eigenvector*. Similarly the *left eigenvector is* $\mathbf{w}_\mathrm{i}$ defined by the relation

$$\mathbf{w}_i^T\mathbf{A} = \lambda_i\mathbf{w}_i{}^T, \ \mathbf{w}_i \neq \mathbf{0}, \ i = 1, \ldots, n. \tag{3.242}$$

Taking transpose on both sides of the equal sign in (3.242) gives

$$\mathbf{A}^T\mathbf{w}_i = \lambda_i\mathbf{w}_i \tag{3.243}$$

and it is seen that the left eigenvectors of **A** are the right eigenvectors of $\mathbf{A}^T$.

All of the expressions (3.242) can be written in a compact form,

$$\begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_n^T \end{bmatrix} \mathbf{A} = \Lambda \begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_n^T \end{bmatrix} \text{ or } \mathbf{QA} = \Lambda\mathbf{Q}. \tag{3.244}$$

Assuming that the left eigenvectors are linearly independent (which will be the case unless **A** is defective), it is seen from (3.244) that

$$\mathbf{AQ}^{-1} = \mathbf{Q}^{-1}\Lambda. \tag{3.245}$$

Comparing this expression with Eq. (3.138) one has that

$$\mathbf{Q}^{-1} = \mathbf{M} \text{ or } \mathbf{M}^{-1} = \mathbf{Q} = \begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_n^T \end{bmatrix}. \tag{3.246}$$

**Controllability Theorem CC3**

Part 1

The LTI system (3.229) is controllable if and only if no left eigenvector of **A** exists such that:

$$\mathbf{w}_i^T\mathbf{B} = \mathbf{0}, \tag{3.247}$$

which means that no left eigenvector of **A** must be orthogonal to all the columns of **B.**

Part 2

The LTI system (3.229) is controllable if and only if the $n \times (n+m)$ -dimensional matrix,

$$\mathbf{R}_c = [\,s\mathbf{I} - \mathbf{A} \quad \mathbf{B}\,], \tag{3.248}$$

has rank $n$ for any complex scalar $s$. This theorem is proved in RUGH (1996).

The controllability test in the two-part theorem CC3 is called the *Popov-Belevitch-Hautus test* or the*PBH test*. Based on theorem CC3 a controllability test for systems can finally be found with diagonal system matrices which is particularly easy to use since it can answer the controllability question simply, by inspection.

**Controllability Theorem CC4**

A diagonal LTI system with distinct eigenvalues is controllable if and only if the **B** matrix has no zero rows.

The state equation for such a system can be written:

$$\dot{\mathbf{x}} = \Lambda\mathbf{x} + \mathbf{B}\mathbf{u} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} \mathbf{x} + \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{bmatrix} \mathbf{u}, \qquad (3.249)$$

so $\Lambda = \Lambda^T$ and therefore right and the left eigenvectors are the same. Equation (3.242) becomes

$$\begin{bmatrix} w_{1i} & w_{2i} & \dots & w_{ni} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} = \begin{bmatrix} \lambda_i w_{1i} & \lambda_i w_{2i} & \dots & \lambda_i w_{ni} \end{bmatrix} \; (3.250)$$

and the elements of $i$'th eigenvector are determined from the equations

$$w_{ki}\lambda_k = \lambda_i w_{ki}. \qquad (3.251)$$

Since all eigenvalues are distinct it is found that

$$\begin{aligned} w_{ki} &= 0 \text{ for } k \neq i, \\ w_{ki} &= q_i \text{ for } k = i, \end{aligned} \qquad (3.252)$$

or

$$\mathbf{w}_i^T = [0 \quad 0 \quad \dots \quad 0 \quad q_i \quad 0 \quad \dots \quad 0 \quad 0], \; q_i \neq 0. \qquad (3.253)$$

The product (3.247) will have an appearance like

$$\mathbf{w}_i^T\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 \\ q_i b_{i1} & q_i b_{i2} & q_i b_{i3} & \dots & q_i b_{im} \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}. \qquad (3.254)$$

It is evident from (3.254) that $\mathbf{w}_i^T\mathbf{B} = \mathbf{0}$ if and only if the $i$'th row of $\mathbf{B}$ is a zero row.

If there are repeated eigenvalues, other eigenvectors than (3.253) can be generated and the system can be uncontrollable even if **B** has no zero rows. Applying theorem CC2, it can be proved that controllability is preserved by a similarity transformation.

## *3.8.2 Controllability and Similarity Transformations*

Using the transformation matrix **P** on a system with the matrices **A** and **B**, Eq. (3.130) can be used to find the matrices of the transformed system,

$$\begin{aligned}\mathbf{A}_t &= \mathbf{PAP}^{-1},\\ \mathbf{B}_t &= \mathbf{PB},\end{aligned} \tag{3.255}$$

and the controllability matrix is

$$\mathbf{M}_{ct} = \begin{bmatrix}\mathbf{PB} & \mathbf{PAP} & \mathbf{PA}^2\mathbf{B} & \ldots & \mathbf{PA}^{n-1}\mathbf{B}\end{bmatrix} = \mathbf{PM}_c \tag{3.256}$$

where $\mathbf{M}_c$ is the controllability matrix for the original system. The quadratic **P**-matrix is regular and has full rank and therefore Eq. (3.256) shows that $\mathbf{M}_c$ and $\mathbf{M}_{ct}$ have the same rank.

In other words, controllability is preserved during similarity transformations.

## *3.8.3 Reachability (Continuous Time Systems)*

In addition to the definition of controllability on p. 124, there is a similar definition of *reachability*.

### Reachability Definition

The system (3.212) is said to be reachable on the finite time interval $[t_0, t_f]$ if there exists an input $\mathbf{u}(t)$ which will drive the system from any initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ to any final state $\mathbf{x}(t_f)$. Sometimes the initial state is taken to be the origin and in that case reachability is denoted *controllability-from-the-origin*. For continuous time systems the issue of reachability does not call for further investigations, since controllability and reachability are equivalent for systems in continuous time. The theorems CC1-CC4 can therefore also be used as tests for reachability for this class of systems.

***Example 3.15*. Controllability Analysis with the PBH Test**

The system in Example 3.14 can be tested for controllability by applying the theorems above.

With the system matrices

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 2 & 1 \\ -2 & -3 & 0 & 0 \\ -2 & -2 & -4 & 0 \\ -2 & -2 & -2 & -5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ -2 \\ 2 \\ -1 \end{bmatrix},$$

the controllability matrix $\mathbf{M}_c$ can be calculated as

$$\mathbf{M}_c = \begin{bmatrix} \mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} & \mathbf{A}^3\mathbf{B} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 & -1 \\ -2 & 4 & -10 & 28 \\ 2 & -6 & 18 & -54 \\ -1 & 3 & -9 & 27 \end{bmatrix}.$$

The matrix has rank 2 and it is concluded that the system is not controllable.

The left eigenvectors corresponding to the eigenvalues are

$$\begin{aligned} \mathbf{w}_1^T &= [4 \quad 3 \quad 2 \quad 1] \text{ for } \lambda_1 = -1, \\ \mathbf{w}_2^T &= [3 \quad 3 \quad 2 \quad 1] \text{ for } \lambda_2 = -2, \\ \mathbf{w}_3^T &= [2 \quad 2 \quad 2 \quad 1] \text{ for } \lambda_3 = -3, \\ \mathbf{w}_4^T &= [1 \quad 1 \quad 1 \quad 1] \text{ for } \lambda_4 = -4. \end{aligned}$$

It can be seen immediately that $\mathbf{w}_4^T\mathbf{B} = \mathbf{0}$ and again, this time according to the PBH-test, the conclusion is that the system is not controllable.

This result is plausible in the light of the block diagram on Fig. 3.24. It was seen above that controllability is not changed by the similarity transformation: if the diagonal system is not controllable, neither is the original system. That the diagonal system is not controllable is evident not only from Fig. 3.24 but also by theorem CC4 which can be applied here since the eigenvalues are distinct. The $\mathbf{B}_t$-matrix has two zero rows and the system is certainly not controllable. ❐

***Example 3.16*. Controllability of an Electrical Network**

Consider the electrical system on Fig. 3.25.

It is desired to create a model of the system with the voltage $u$ as input and the current $y$ as output.

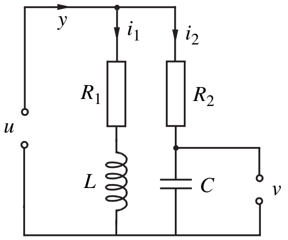


**Fig. 3.25** Passive electrical circuit

Using Ohm's and Kirchhoff's laws and the relations for inductive and capacitive impedances on the two branches of the circuit,

$$\begin{aligned} u &= R_1 i_1 + L\frac{di_1}{dt}, \\ u &= R_2 i_2 + v, \\ v &= \frac{1}{C}\int i_2 dt. \end{aligned} \tag{3.257}$$

Differentiating the third equation and inserting for $i_2$ in the second one yields

$$\begin{aligned} u &= R_1 i_1 + L\frac{di_1}{dt}, \\ u &= R_2 C\frac{dv}{dt} + v. \end{aligned} \tag{3.258}$$

Choosing the state vector, $\mathbf{x} = [i_1 \;\; v]^T$, the state equation for the circuit can easily be found:

$$\dot{\mathbf{x}} = \begin{bmatrix} -\dfrac{R_1}{L} & 0 \\ 0 & -\dfrac{1}{R_2 C} \end{bmatrix}\mathbf{x} + \begin{bmatrix} \dfrac{1}{L} \\ \dfrac{1}{R_2 C} \end{bmatrix} u. \tag{3.259}$$

The output $y$ is the sum of the two currents,

$$y = i_1 + i_2 = i_1 + \frac{u - v}{R_2} = x_1 - \frac{1}{R_2}x_2 + \frac{1}{R_2}u = \begin{bmatrix} 1 \dfrac{1}{R_2} \end{bmatrix}\mathbf{x} + \frac{1}{R_2}u. \tag{3.260}$$

The system is by its nature in diagonal form and the controllability theorem CC4 can be applied if the eigenvalues are distinct.

From (3.259) it can be seen directly that the eigenvalues are

$$\begin{aligned} \lambda_1 &= -\frac{R_1}{L}, \\ \lambda_2 &= -\frac{1}{R_2 C}. \end{aligned} \tag{3.261}$$

Since both rows of $\mathbf{B}$ are always nonzero, it is obvious that the system is controllable if the eigenvalues are distinct, i.e., if

$$\frac{R_1}{L} \neq \frac{1}{R_2 C}, \tag{3.262}$$

the controllability matrix is

$$\mathbf{M}_c = [\mathbf{B} \quad \mathbf{AB}] = \begin{bmatrix} \dfrac{1}{L} & -\dfrac{R_1}{L^2} \\ \dfrac{1}{R_2 C} & -\dfrac{1}{R_2^2 C^2} \end{bmatrix}. \tag{3.263}$$

This quadratic matrix has rank 2 if the determinant is nonzero, which is precisely the case if the condition (3.262) is fulfilled.

If the component values

$$\begin{aligned} R_1 &= 100\Omega, \\ R_2 &= 1\,\mathrm{k}\Omega, \\ L &= 1\,\mathrm{H}, \\ C &= 10\,\mu\mathrm{F}, \end{aligned}$$

are selected the model below emerges:

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} -100 & 0 \\ 0 & -100 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 \\ 100 \end{bmatrix}u, \\ y &= [\,1 \quad -0.001\,]\mathbf{x} + 0.001u. \end{aligned} \tag{3.264}$$

In this case the eigenvalues are equal and theorem CC4 does not apply. Thus, even though the **B** matrix has no zero rows, one *cannot* conclude that the system is controllable. On the contrary, the controllability matrix $\mathbf{M}_c$ does not have full rank and the system is clearly uncontrollable.

In this case it is simple to see why this must be the case. A block diagram of the system is seen on Fig. 3.26.

The two states are decoupled and the two independent first order differential equations,

$$\begin{aligned} \dot{x}_1 &= -100x_1 + u, \\ \dot{x}_2 &= -100x_2 + 100u, \end{aligned} \tag{3.265}$$

can easily be solved. The solutions are

$$\begin{aligned} x_1(t) &= e^{-100t}(x_{10} + \mu(t)), \\ x_2(t) &= e^{-100t}(x_{20} + 100\mu(t)), \end{aligned} \tag{3.266}$$

where $x_{10}$ and $x_{20}$ are the initial values and

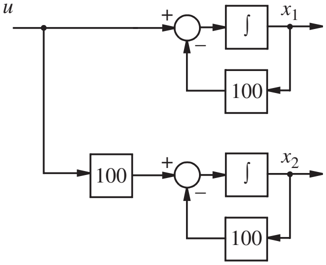


**Fig. 3.26** Block diagram of the uncontrollable system

$$\mu(t) = \int_0^t e^{100\tau} u(\tau) d\tau. \tag{3.267}$$

If the situation is considered where it is desired to drive the system from an initial state to the zero state,

$$\mathbf{x}(t_f) = \begin{bmatrix} x_1(t_f) \\ x_2(t_f) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

for some final time $t_f$, it is seen from (3.266) that this is only possible if $x_{20} = 100x_{10}$. But this is in disagreement with the definition of controllability, which requires that the system is taken from *any* initial state to the zero state in a given finite period of time. Consequently, the system is not controllable. ❒

***Example 3.17*. Controllability of the Water Tank Process**

In the tank Example 2.9 a linearized MIMO model was derived for a specific stationary state. The inputs to the system were the two input voltages $u_1$ and $u_1$ to the control valves, the outputs were the measured level $H_2$ and temperature $T_2$ of the R-tank and the state vector was composed of the two levels and the two temperatures,

$$\mathbf{x}(t) = \begin{bmatrix} H_1 \\ H_2 \\ T_1 \\ T_2 \end{bmatrix}.$$

The system matrices turned out to be:

$$\mathbf{A} = \begin{bmatrix} -0.0499 & 0.0499 & 0 & 0 \\ 0.0499 & -0.0667 & 0 & 0 \\ 0 & 0 & -0.0251 & 0 \\ 0 & 0 & 0.0335 & -0.0355 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0.00510 & 0.00510 \\ 0 & 0 \\ 0.0377 & -0.0377 \\ 0 & 0 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}.$$

Using MATLAB the controllability matrix can be found,

$$\mathbf{M}_c = \begin{bmatrix} 509.5 & 509.5 & -25.45 & -25.45 & 2.541 & 2.541 & -0.2711 & -0.2711 \\ 0 & 0 & 25.45 & 25.45 & -2.968 & -2.968 & 0.3249 & 0.3249 \\ 3766 & -3766 & -94.56 & 94.56 & 2.374 & -2.374 & -0.05961 & 0.05961 \\ 0 & 0 & 126.3 & -126.3 & -7.408 & 7.408 & 0.3291 & -0.3291 \end{bmatrix} \cdot 10^{-5}.$$

The first 4 columns of $\mathbf{M}_c$ are linearly independent and therefore the matrix has full rank and the system is controllable. Canceling the first column of the **B**-matrix, which means discarding the first input voltage $u_1$, one might expect difficulties with the controllability. However, testing the controllability from the matrix-pair **A** (unchanged) and the reduced input matrix,

$$\mathbf{B}_{red} = \begin{bmatrix} 0.00510 \\ 0 \\ -0.0377 \\ 0 \end{bmatrix},$$

it is discoverd (using MATLAB's `rank` function) that the system is still controllable. ❐

### *3.8.4 Controllability (Discrete Time Systems)*

A definition of controllability exists for discrete time systems and it is quite similar to the definition on p. 124. However, whereas for continuous systems the properties of controllability and reachability are equivalent (controllability implies reachability and vice versa), this is not necessarily the case for discrete time systems.

If a discrete time system is reachable one can find a sequence of control inputs which will bring the system from any *initial* state to any *final* state. Choosing the origin as the final state, the system fulfills the requirements for controllability and it is seen that if the system is reachable, it is also controllable. The following simple example shows that the converse need not be true.

***Example 3.18.*** **Non-Reachable System**

Consider now the second order discrete time LTI system

$$\mathbf{x}(k+1) = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}(k) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(k). \tag{3.268}$$

For the individual states one can write

$$\begin{aligned} x_1(k+1) &= x_1(k) + x_2(k) + u(k), \\ x_2(k+1) &= 0. \end{aligned} \tag{3.269}$$

The system is clearly controllable because if for an arbitrary initial state one has

$$\mathbf{x}_0 = [a \quad b]^T$$

one can set

$$u(k) = -(a+b), 0, 0, \ldots$$

and find that

$$\mathbf{x}(k) = \mathbf{0} \text{ for } k \geq 1,$$

which shows that the system is controllable. However, the second state $x_2(k)$ is zero for all $k \geq 1$ no matter what the input is and it is obviously not possible to move it somewhere else. Therefore the system is not reachable. ❐

It turns out that this problem only arises if the discrete time system matrix $\mathbf{F}$ is singular. If $\mathbf{F}$ is regular, controllability implies reachability as in the continuous time case.

Although singular system matrices may be rare, the consequences of the finding above must be accepted and the conclusion is that reachability is the better measure for characterizing discrete time systems. Thus, in the following the theorems and tests presented will be in terms of reachability.

### *3.8.5 Reachability (Discrete Time Systems)*

**Reachability Definition**

The discrete time linear system,

$$\mathbf{x}(k+1) = \mathbf{F}(k)\mathbf{x}(k) + \mathbf{G}(k)\mathbf{u}(k), \tag{3.270}$$

is said to be reachable on the finite time interval $[k_0, k_f]$ if there exists an input sequence $\mathbf{u}(k)$ which will drive the system from the initial state $\mathbf{x}(k_0) = \mathbf{0}$ to any final state $\mathbf{x}(k_f)$. This property can also be denoted *controllability-from-the-origin*. Sometimes the initial state is taken to be any vector $\mathbf{x}(k_0)$ in the state space. In the theorems and proofs below it is assumed that $\mathbf{x}(k_0) = \mathbf{0}$ but only minor and immaterial changes will occur if another initial state is selected.

**Reachability Theorem RD1**

The system (3.270) is reachable on $[k_0, k_f]$ if and only if the quadratic *reachability Gramian,*

$$\mathbf{W}_r(k_0, k_f) = \sum_{i=k_0}^{k_f-1} \phi(k_f, i+1)\mathbf{G}(i)\mathbf{G}^T(i)\phi^T(k_f, i+1), \tag{3.271}$$

is regular.

It can be proved that the condition is sufficient by assuming that $\mathbf{W}_r(k_0, k_f)$ is regular and then, for $\mathbf{x}(k_0) = \mathbf{0}$ and an arbitrary final state $\mathbf{x}(k_f)$, choose the sequence of input vectors,

$$\mathbf{u}(k) = \mathbf{G}^T(k)\phi^T(k_f, k+1)\mathbf{W}_r^{-1}(k_0, k_f)\mathbf{x}(k_f)\,,\; k = k_0,\, k_1,\, \ldots,\, k_f - 1. \tag{3.272}$$

According to Eq. (3.91) the solution to the state equation at the final time can be written as

$$\begin{aligned} \mathbf{x}(t_f) &= \sum_{i=k_0}^{k_f-1} \phi(k_f, i+1) \\ &= \sum_{i=k_0}^{k_f-1} \phi(k_f, i+1)\mathbf{G}(i)\mathbf{G}^T(i)\phi^T(k_f, i+1) \\ &= \mathbf{x}(k_f) \end{aligned} \tag{3.273}$$

and this shows that the system is reachable on $[k_0, k_f]$.

Conversely, assume that the system is reachable but the Gramian is singular. If so, a nonzero vector $\mathbf{x}_a$ can be found such that it's quadratic form with $\mathbf{W}_r(k_0, k_f)$ becomes zero,

$$\mathbf{x}_a^T\mathbf{W}_r(k_0, k_f)\mathbf{x}_a = 0, \tag{3.274}$$

or

$$\mathbf{x}_a^T\mathbf{W}_r(k_0, k_f)\mathbf{x}_a = \sum_{i=k_0}^{k_f-1} \mathbf{x}_a^T\phi(k_f, i+1)\mathbf{G}(i)\mathbf{G}^T(i)\phi^T k_f, (i+1)\mathbf{x}_a = 0. \tag{3.275}$$

Defining

$$\mathbf{z}(i) = \mathbf{G}^T(i)\phi^T(k_f, i+1)\mathbf{x}_a\,,\; i = k_0,\, k_1,\, \ldots,\, k_f - 1, \tag{3.276}$$

it is seen that the summand can be written,

$$\mathbf{x}_a^T\phi(k_f, i+1)\mathbf{G}(i)\mathbf{G}^T(i)\phi^T(k_f, i+1)\mathbf{x}_a = \mathbf{z}^T(i)\mathbf{z}(i) = \|\mathbf{z}(i)\|^2. \tag{3.277}$$

Since the norm is nonnegative, $\mathbf{z}(i)$ must be identically zero for (3.275) to hold.

It was assumed that the system is reachable, so choosing $\mathbf{x}(k_f) = \mathbf{x}_a$, it is known that an input sequence $\mathbf{u}_a(k)$ can be found such that

$$\mathbf{x}_a = \mathbf{x}(t_f) = \sum_{i=k_0}^{k_f-1} \phi(k_f, i+1)\mathbf{G}(i)\mathbf{u}_a(i). \tag{3.278}$$

Premultiplying by $\mathbf{x}_a^T$ gives

$$\mathbf{x}_a^T\mathbf{x}_a = \sum_{i=k_0}^{k_f-1} \mathbf{x}_a^T \phi(k_f, i+1)\mathbf{G}(i)\mathbf{u}_a(i) = \sum_{i=k_0}^{k_f-1} \mathbf{z}^T(i)\mathbf{u}_a(i) = 0 \tag{3.279}$$

which contradicts the assumption that $\mathbf{x}_a \neq \mathbf{0}$. The conclusion is that if the system is reachable, then the Gramian must be regular.

The Gramian is in general symmetric and positive semidefinite. If the system is reachable, the Gramian is positive definite.

In the time invariant case,

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k). \tag{3.280}$$

The corresponding state transfer matrix is (see Eq. (3.90))

$$\phi(k_f, i+1) = \mathbf{F}^{k_f-1-i} \tag{3.281}$$

and the reachability Gramian can be written (for $k_0 = 0$)

$$\mathbf{W}_r(k_f) = \sum_{i=0}^{k_f-1} \mathbf{F}^{k_f-1-i}\mathbf{G}\mathbf{G}^T(\mathbf{F}^T)^{k_f-1-i} \tag{3.282}$$

or, changing the summation index to $j = k_f - 1 - i$,

$$\mathbf{W}_r(k_f) = \sum_{j=0}^{k_f-1} \mathbf{F}^j\mathbf{G}\mathbf{G}^T(\mathbf{F}^T)^j. \tag{3.283}$$

**Reachability Theorem RD2**

The LTI system model,

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k), \tag{3.284}$$

is reachable if and only if the *reachability matrix,*

$$\mathbf{M}_r = \begin{bmatrix} \mathbf{G} & \mathbf{F}\mathbf{G} & \mathbf{F}^2\mathbf{G} & \dots & \mathbf{F}^{n-1}\mathbf{G} \end{bmatrix}, \tag{3.285}$$

has rank $n$, i.e., full rank.

The solution to (3.284) at the final time $k_f$ is given by Eq. (3.98),

$$\mathbf{x}(k_f) = \sum_{i=0}^{k_f-1} \mathbf{F}^{k_f-1-i}\mathbf{G}\mathbf{u}(i) \text{ for } \mathbf{x}(0) = \mathbf{0}. \tag{3.286}$$

The solution can be written

$$\begin{aligned}\mathbf{x}(k_f) &= \mathbf{F}^{k_f-1}\mathbf{G}\mathbf{u}(0) + \ldots + \mathbf{F}\mathbf{G}\mathbf{u}(k_f-2) + \mathbf{G}\mathbf{u}(k_f-1) \\ &= \begin{bmatrix}\mathbf{G} & \mathbf{F}\mathbf{G} & \mathbf{F}^2\mathbf{G} & \ldots & \mathbf{F}^{k_f-1}\mathbf{G}\end{bmatrix}\begin{bmatrix}\mathbf{u}(k_f-1) \\ \vdots \\ \mathbf{u}(2) \\ \mathbf{u}(1) \\ \mathbf{u}(0)\end{bmatrix}.\end{aligned} \tag{3.287}$$

For reachability it is required that $\mathbf{x}(k_f)$ can be assigned any value in the $n$-dimensional state space and with the input sequence in (3.287), it is therefore required that the vectors in the matrix immediately to the right of the last equal sign have at least $n$ linearly independent columns. If this is not the case for some $k_f$, one may improve the situation by letting $k_f$ increase, thus adding more columns to the matrix. However, according to Cayley-Hamilton's theorem all powers $\mathbf{F}^p$, for $p \geq n$ can be expressed as a linear combination of the powers up to $n-1$, so one does not gain anything further by letting $k_f$ grow larger than $n$. This shows that Eq. (3.287) has a unique solution if and only if $\mathbf{M}_r$ has full rank. The reachability matrix is precisely the same as the controllability matrix (3.230) for continuous time systems and very often $\mathbf{M}_r$ is called the controllability matrix for the system (3.284).

**Reachability Theorem RD3**

The PBH test is also valid for discrete time systems.

Part 1
The LTI system (3.280) is controllable if and only if no left eigenvector of **F** exists such that

$$\mathbf{w}_i^T\mathbf{G} = \mathbf{0} \tag{3.288}$$

which means that no left eigenvector of **F** must be orthogonal to all the columns of **G**.

Part 2
The LTI system (3.280) is controllable if and only if the $n \times (n+m)$-dimensional matrix,

$$\mathbf{R}_r = [z\mathbf{I} - \mathbf{F} \quad \mathbf{G}], \tag{3.289}$$

has rank $n$ for all complex scalars $z$.

**Reachability Theorem RD4**

A diagonal LTI system with distinct eigenvalues is controllable if and only if the **G** matrix has no zero rows.

The proofs for theorems RD3 and RD4 are similar to the proofs for theorems CC3 and CC4.

## 3.8.6 Observability (Continuous Time Systems)

**Observability Definition**

The linear system,

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t),\ \ \mathbf{x}(t_0) = \mathbf{x}_0, \\ \mathbf{y}(t) &= \mathbf{C}(t)\mathbf{x}(t), \end{aligned} \tag{3.290}$$

is said to be observable on the finite time interval $[t_0, t_f]$ if any initial state $\mathbf{x}_0$ is uniquely determined by the output $\mathbf{y}(t)$ over the same time interval.

Observability has nothing to do with the input to the system but only with the way the states are interconnected and the way the output is connected to the states. Consequently only the unforced state space model is considered.

**Observability Theorem OC1**

The linear state model (3.290) is observable if and only if the quadratic $n \times n$ matrix,

$$\mathbf{W}_o(t_0, t_f) = \int_{t_0}^{t_f} \phi^T(t, t_0)\mathbf{C}^T(t)\mathbf{C}(t)\phi(t, t_0)dt, \tag{3.291}$$

is regular. The matrix $\mathbf{W}_o(t_0, t_f)$ is called the *observability Gramian.*

To see that the condition is sufficient consider the output solution (see Eq. (3.26)),

$$\mathbf{y}(t) = \mathbf{C}(t)\phi(t, t_0)\mathbf{x}_0. \tag{3.292}$$

Premultiplying by $\phi^T(t, t_0)\mathbf{C}^T(t)$ and integrating on both sides of the equal sign yields

$$\int_{t_0}^{t_f} \phi^T(t, t_0)\mathbf{C}^T(t)\mathbf{y}(t)dt = \mathbf{W}_o(t_0, t_f)\mathbf{x}_0 \tag{3.293}$$

which is a set of linear equations in the elements of $\mathbf{x}_0$. If $\mathbf{W}_o(t_0, t_f)$ is regular, $\mathbf{x}_0$ is uniquely determined.

Conversely, if the Gramian is singular, a nonzero vector $\mathbf{x}_a$ can be found such that

$$\mathbf{W}_0(t_0, t_f)\mathbf{x}_a = \mathbf{0} \tag{3.294}$$

and therefore also

$$\mathbf{x}_a^T \mathbf{W}_o(t_0, t_f)\mathbf{x}_a = 0. \tag{3.295}$$

But then

$$\begin{aligned} \mathbf{x}_a^T \mathbf{W}_o(t_0, t_f)\mathbf{x}_a^T &= \int_{t_0}^{t_f} \mathbf{x}_a^T \phi^T(t, t_0)\mathbf{C}^T(t)\mathbf{C}(t)\phi(t, t_0)\mathbf{x}_a dt \\ &= \int_{t_0}^{t_f} \mathbf{z}^T(t)\mathbf{z}(t)dt = 0 \end{aligned} \tag{3.296}$$

where $\mathbf{z}(t)$ has been defined as

$$\mathbf{z}(t) = \mathbf{C}(t)\phi(t, t_0)\mathbf{x}_a. \tag{3.297}$$

The integrand in the right hand term of (3.297) is the square of the norm of $\mathbf{z}(t)$ and since the norm is nonnegative, it can be concluded that

$$\mathbf{C}(t)\phi(t, t_0)\mathbf{x}_a = \mathbf{0}\ ,\ \ t \in [t_0, t_f]. \tag{3.298}$$

Note that $\mathbf{z}(t)$ is precisely the output of the system for $\mathbf{x}(t_0) = \mathbf{x}_a$ and since this output is identically zero over the entire time interval, $\mathbf{x}(t_0)$ cannot be determined from $\mathbf{y}(t)$ and the system is clearly not observable. In other words, the regularity of the Gramian is also a necessary condition for observability.

For LTI systems one has $\phi(t, t_0) = e^{\mathbf{A}(t-t_0)}$ and for $t_0 = 0$ the observability Gramian becomes

$$\mathbf{W}_o(t_f) = \int_0^{t_f} e^{\mathbf{A}^T t}\mathbf{C}^T\mathbf{C}e^{\mathbf{A}t}dt. \tag{3.299}$$

**Observability Theorem OC2**

The LTI system model,

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t), \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t), \end{aligned} \tag{3.300}$$

is observable if and only if the *observability matrix*

$$\mathbf{M}_o = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ : \\ \mathbf{CA}^{n-1} \end{bmatrix}, \tag{3.301}$$

has rank $n$, i.e., full rank.

It can be shown that this rank requirement fails if and only if the observability Gramian is singular. If $\mathbf{W}_0(t_f)$ is singular the condition (3.298) is applicable. For the LTI system it is known that $\phi(t_0, t) = \phi(t) = e^{\mathbf{A}t}$ and (3.298) becomes

$$\mathbf{z}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}_a = \mathbf{0}\ ,\ \ t \in [0,\ t_f], \tag{3.302}$$

Since this vector is identically zero on the entire time interval, its time derivatives must also be zero

$$\begin{aligned} d\mathbf{z}(t)/dt &= \mathbf{CA}e^{\mathbf{A}t}\mathbf{x}_a = \mathbf{0}, \\ d^2\mathbf{z}(t)/dt^2 &= \mathbf{CA}^2e^{\mathbf{A}t}\mathbf{x}_a = \mathbf{0}, \\ &\vdots \\ d^{n-1}\mathbf{z}(t)/dt^{n-1} &= \mathbf{CA}^{n-1}\,e^{\mathbf{A}t}\mathbf{x}_a = \mathbf{0}. \end{aligned} \tag{3.303}$$

This means that

$$\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix} e^{\mathbf{A}t}\mathbf{x}_a = \mathbf{M}_o e^{\mathbf{A}t}\mathbf{x}_a = \mathbf{0}. \tag{3.304}$$

$e^{\mathbf{A}t}\mathbf{x}_a$ is a nonzero $n \times 1$ vector whose elements are time functions,

$$e^{\mathbf{A}t}\mathbf{x}_a = \begin{bmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_n(t) \end{bmatrix}. \tag{3.305}$$

If the $n$ columns of $\mathbf{M}_o$ are denoted $\mathbf{M}_c$,

$$\mathbf{M}_c = [\mathbf{c}_1\ \mathbf{c}_2\ \ldots \mathbf{c}_n]. \tag{3.306}$$

It is seen from (3.304) that

$$\sum_{i=1}^{n} a_i(t)\mathbf{c}_i = \mathbf{0}. \tag{3.307}$$

The last expression shows that the $n$ columns of $\mathbf{M}_o$ are not linearly independent and this means that $\mathbf{M}_o$ has less than full rank.

The ‘only if’-part can be proven as follows.

Using the series expansion (3.44) of the matrix exponential and additionally the Cayley-Hamilton theorem, $e^{\mathbf{A}t}$ can be expressed as the finite series,

$$e^{\mathbf{A}t} = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k (t)^k}{k!} = \sum_{k=0}^{n-1} \gamma_k(t)\, \mathbf{A}^k. \tag{3.308}$$

Premultiplying with the $r \times n$-dimensional $\mathbf{C}$-matrix gives the $r \times m$-dimensional matrix,

$$\mathbf{C}e^{\mathbf{A}t} = \sum_{k=0}^{n-1} \gamma_k(t)\mathbf{C}\mathbf{A}^k = [\mathbf{I}_r\gamma_0(t) \quad \mathbf{I}_r\gamma_1(t) \quad \ldots \quad \mathbf{I}_r\gamma_{n-1}(t)] \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ : \\ \mathbf{CA}^{n-1} \end{bmatrix}, \tag{3.309}$$

where $\mathbf{I}_r$ is the $r$-dimensional identity matrix. The transpose of this matrix is

$$e^{\mathbf{A}^T t}\mathbf{C}^T = [\mathbf{C}^T \quad \mathbf{A}^T\mathbf{C}^T \quad \ldots \quad (\mathbf{A}^T)^{n-1}\mathbf{C}^T] \begin{bmatrix} \mathbf{I}_r\gamma_0(t) \\ \mathbf{I}_r\gamma_1(t) \\ : \\ \mathbf{I}_r\gamma_{n-1}(t) \end{bmatrix}. \tag{3.310}$$

Inserting (3.309) and (3.310) into (3.299) results in the matrix,

$$\mathbf{W}_o(t_f) = \mathbf{M}_o^T \int_0^{t_f} \begin{bmatrix} \mathbf{I}_r\beta_{11}(t) & \mathbf{I}_r\beta_{12}(t) & \ldots & \mathbf{I}_r\beta_{1n}(t) \\ \mathbf{I}_r\beta_{21}(t) & \mathbf{I}_r\beta_{22}(t) & \ldots & \mathbf{I}_r\beta_{2n}(t) \\ : & : & : & : \\ \mathbf{I}_r\beta_{n1}(t) & \mathbf{I}_r\beta_{n2}(t) & \ldots & \mathbf{I}_r\beta_{nn}(t) \end{bmatrix} dt\mathbf{M}_o = \mathbf{M}_o^T\mathbf{Q}\mathbf{M}_o, \tag{3.311}$$

where the β-functions are products of the $\gamma$-functions above. This equation shows that even if $\mathbf{Q}$ has full rank (i.e., $rn$), the quadratic $n \times n$ matrix product on the right can not have rank larger that the rank of $\mathbf{M}_o$. In other words, if $\mathbf{M}_o$ has less than full rank, the Gramian will be singular.

If the observability matrix $\mathbf{M}_o$ has full rank, then the system is observable for all values of the final time $t_f$.

The PBH test for observability is expressed in the next theorem. In this case the ordinary eigenvectors (right eigenvectors) are used:

$$\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i. \tag{3.312}$$

**Observability Theorem OC3**

Part 1
The LTI system (3.300) is observable if and only if no right eigenvector of **A** exists such that

$$\mathbf{C}\mathbf{v}_i = \mathbf{0} \tag{3.313}$$

which means that no right eigenvector of **A** must be orthogonal to all the rows of **C**.

Part 2
The LTI system (3.300) is observable if and only if the $(r+n) \times n$-dimensional matrix,

$$\mathbf{R}_o = \begin{bmatrix} \mathbf{C} \\ s\mathbf{I} - \mathbf{A} \end{bmatrix}, \tag{3.314}$$

has rank $n$ for any complex scalar $s$.

The theorem is proved in Middleton and Goodwin (1990).

For a diagonal system the following theorem is valid

**Observability Theorem OC4**

A diagonal LTI system with distinct eigenvalues is observable if and only if the **C** matrix has no zero columns.

The proof can be based on observability theorem OC3 and follows the same lines as the proof of controllability theorem CC4 on p. 130.

### 3.8.7 Observability and Similarity Transformations

Using the transformation matrix **P** on a system with the matrices **A** and **C**, Eq. (3.130) can be used and find the matrices for a transformed system,

$$\begin{aligned} \mathbf{A}_t &= \mathbf{P}\mathbf{A}\mathbf{P}^{-1}, \\ \mathbf{C}_t &= \mathbf{C}\mathbf{P}^{-1}, \end{aligned} \tag{3.315}$$

and the controllability matrix,

$$\mathbf{M}_{ot} = \begin{bmatrix} \mathbf{C}\mathbf{P}^{-1} \\ \mathbf{C}\mathbf{A}\mathbf{P}^{-1} \\ \mathbf{C}\mathbf{A}^2\mathbf{P}^{-1} \\ \vdots \\ \mathbf{C}\mathbf{A}^{n-1}\mathbf{P}^{-1} \end{bmatrix} = \mathbf{M}_o\mathbf{P}^{-1}, \tag{3.316}$$

where $\mathbf{M}_o$ is the observability matrix for the original system. The quadratic $\mathbf{P}^{-1}$ matrix is regular and has full rank and therefore Eq. (3.316) shows that $\mathbf{M}_o$ and $\mathbf{M}_{ot}$ have the same rank.

In other words, observability is preserved during similarity transformations.

***Example 3.19.*** **Observability of the Water Tank Process**

Consider now the observability property for the tank system in Example 2.9 (see also Example 3.17).

The observability matrix can be computed to be

$$\mathbf{M_0} = \begin{bmatrix} 0 & 2\cdot 10^5 & 0 & 0 \\ 0 & 0 & 0 & 10^4 \\ 9987 & -1.134\cdot 10^4 & 0 & 0 \\ 0 & 0 & 335.4 & -335.4 \\ -1165 & 1389 & 0 & 0 \\ 0 & 0 & -19.67 & 11.25 \\ 127.5 & -150.1 & 0 & 0 \\ 0 & 0 & 0.8712 & -0.3773 \end{bmatrix} .10^{-5}.$$

By using MATLAB it is found that $\mathbf{M}_o$ has full rank (the first four rows are linearly independent) and it is concluded that the system is observable.

Omitting the temperature $T_2$ as an output, a reduced output matrix is obtained,

$$\mathbf{C}_{red} = [0\ 2\ 0\ 0],$$

and the new observability matrix for the pair $(\mathbf{A}, \mathbf{C}_{red})$ is

$$\mathbf{M}_{o,\,red} = \begin{bmatrix} 0 & 2\cdot 10^5 & 0 & 0 \\ 9987 & -1.344\cdot 10^4 & 0 & 0 \\ -1165 & 1389 & 0 & 0 \\ 127.5 & -150.8 & 0 & 0 \end{bmatrix} .10^{-5}.$$

This matrix is obviously singular (as it has zero columns) and the system with only the single output $y_1$ is not observable.

The system matrix **A** has the eigenvalues,

$$\lambda_{\mathbf{A}} = \begin{cases} -0.03354 \\ -0.02511 \\ -0.109 \\ -0.007686 \end{cases},$$

and the corresponding right eigenvectors are

$$\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0.6459 & -0.7634 \\ 0 & 0 & -07934 & -0.6459 \\ 0 & 0.2438 & 0 & 0 \\ 1 & 0.9698 & 0 & 0 \end{bmatrix}.$$

Using the PBH-test for observability, the following products are found:

$$\mathbf{C}\mathbf{v}_1 = \begin{bmatrix} 0 \\ 0.1 \end{bmatrix}, \mathbf{C}\mathbf{v}_2 = \begin{bmatrix} 0 \\ 0.097 \end{bmatrix}, \mathbf{C}\mathbf{v}_3 = \begin{bmatrix} 0 \\ -1.527 \end{bmatrix}, \mathbf{C}\mathbf{v}_4 = \begin{bmatrix} -1.292 \\ 0 \end{bmatrix}.$$

None of these vectors are zero vectors and this shows that the original system with two outputs is observable.

Carrying out the same calculation based on the reduced output matrix, $\mathbf{C}_{red}$, one finds

$$\mathbf{C}_{red}\mathbf{v}_1 = 0, \ \mathbf{C}_{red}\mathbf{v}_2 = 0 \ \mathbf{C}_{red}\mathbf{v}_3 = -1.527, \ \mathbf{C}_{red}\mathbf{v}_4 = -1.292.$$

Since two of these products (in this case scalars) are zero, it can be concluded as before, that the system with only one output is not observable. ❐

### 3.8.8 Observability (Discrete Time Systems)

As it is the case for controllability/reachability, the proofs are often easier to carry out in discrete time than in continuous time and therefore the proofs are not given in this section. The proofs can be found in Kailath(1980), Rugh(1996) or in Middleton and Goodwin (1990). The proofs can also be carried out as a useful exercise for the reader.

**Observability Definition (Discrete Time)**

The discrete time linear system,

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{F}(k)\mathbf{x}(k), \ \mathbf{x}(k_0) = \mathbf{x}_0, \\ \mathbf{y}(k) &= \mathbf{C}(t)\mathbf{x}, \end{aligned} \tag{3.317}$$

is said to be observable on the finite time interval $[k_0 \quad k_f]$ if any initial state $\mathbf{x}_0$ is uniquely determined by the output $\mathbf{y}(k)$ over the same time interval.

**Observability Theorem OD1**

The state model (3.317) is observable if and only if the quadratic $n \times n$ observability Gramian,

$$\mathbf{W}_o(k_0, k_f) = \sum_{i=k_0}^{k_f-1} \phi^T(i, k_0)\mathbf{C}^T(k)\mathbf{C}\phi(i, k_0), \tag{3.318}$$

is regular. $\mathbf{W}_o(k_0, k_f)$ is symmetric and in general positive semidefinite. If the system is observable, the Gramian is positive definite.

For an LTI system with $\phi(i, k_0) = \mathbf{F}^{i-k_0}$ and with $k_0 = 0$ the Gramian is

$$\mathbf{W}_o(k_f) = \sum_{i=0}^{k_f-1} (\mathbf{F}^T)^i \mathbf{C}^T \mathbf{C}\mathbf{F}^i. \tag{3.319}$$

**Observability Theorem OD2**

The LTI system,

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{F}\mathbf{x}(k), \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k), \end{aligned} \tag{3.320}$$

is observable if and only if theobservability matrix,

$$\mathbf{M}_o = \begin{bmatrix} \mathbf{C} \\ \mathbf{CF} \\ \mathbf{CF}^2 \\ \vdots \\ \mathbf{CF}^{n-1} \end{bmatrix}, \tag{3.321}$$

has rank $n$ (full rank).

**Observability Theorem OD3**

The PBH test.

Part 1

The LTI system (3.320) is observable if and only if no right eigenvector of $\mathbf{F}$ exists such that

$$\mathbf{C}\mathbf{v}_i = \mathbf{0} \tag{3.322}$$

which means that no right eigenvector of **F** is orthogonal to all the rows of **C**.

Part 2

The LTI system (3.320) is observable if and only if the $(r+n) \times n$-dimensional matrix,

$$\mathbf{R}_o = \begin{bmatrix} \mathbf{C} \\ z\mathbf{I} - \mathbf{F} \end{bmatrix}, \tag{3.323}$$

has rank n for all complex scalars $z$.

**Observability Theorem OD4**

The diagonal theorem is the same as the continuous time observability theorem OC4.

On the following pages an overview of the controllability/reachability and the observability theorems is given.

### 3.8.9 Duality

If the two LTI state space models below are considered:

$$S_x : \begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases} \tag{3.324}$$

with the model

$$S_z : \begin{cases} \dot{\mathbf{z}} = \mathbf{A}^T\mathbf{z} + \mathbf{C}^T\mathbf{u} \\ \mathbf{y} = \mathbf{B}^T\mathbf{z} \end{cases} \quad . \tag{3.325}$$

The controllability and observability matrices for the latter systems are:

$$\mathbf{M}_{c,z} = \begin{bmatrix} \mathbf{C}^T & \mathbf{A}^T\mathbf{C}^T & \ldots & (\mathbf{A}^T)^{n-1} & \mathbf{C}^T \end{bmatrix} = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix}^T = \mathbf{M}_{o,x}^T, \tag{3.326}$$

$$\mathbf{M}_{o,z} = \begin{bmatrix} \mathbf{B}^T \\ \mathbf{B}^T\mathbf{A}^T \\ \vdots \\ \mathbf{B}^T(\mathbf{A}^T)^{n-1} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & \mathbf{AB} & \ldots & \mathbf{A}^{n-1}\mathbf{B} \end{bmatrix}^T = \mathbf{M}_{c,x}^T. \tag{3.327}$$

This shows that the system $S_z$ is controllable if and only if $S_x$ is observable and that the system $S_z$ is observable if and only if $S_x$ is controllable. This property is called *duality*.

### 3.8.10 Modal Decomposition

In the solutions of the state equations will be again investigated here in the light of controllability and observability.

Controllability/Reachability Overview

| | Controllability/Reachability<br>The continuous system: $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$ is controllable/reachable if and only if: | | Reachability<br>The discrete system: $\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k)$ is reachable if and only if |
|---|---|---|---|
| CC1 | $\mathbf{W}_c(t_0, t_f) = \int_{t_0}^{t_f} \phi(t_0, t)\mathbf{B}(t)\mathbf{B}^T(t)\phi^T(t_0, t)dt$<br>is regular | RD1 | $\mathbf{W}_r(k_0, k_f) = \sum_{i-k_0}^{k_f-1} \phi(k_f, i+1)\mathbf{G}(i)\mathbf{G}^T(i)\phi^T(k_f, i+1)$<br>is regular |
| CC2 | The matrix $\mathbf{M}_c = \begin{bmatrix} \mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} & \dots & \mathbf{A}^{n-1}\mathbf{B} \end{bmatrix}$<br>has full rank | RD2 | The matrix $\mathbf{M}_c = \begin{bmatrix} \mathbf{G} & \mathbf{FG} & \mathbf{F}^2\mathbf{G} & \dots & \mathbf{F}^{n-1}\mathbf{G} \end{bmatrix}$<br>has full rank |
| CC3 | PBH-test : $\begin{cases} \text{No } \mathbf{w}_i \text{ exists such that } \mathbf{w}_i^T\mathbf{B} = 0 \text{ or} \\ \text{rank}[s\mathbf{I} - \mathbf{A} \quad \mathbf{B}] = n \text{ for all } s \end{cases}$ | RD3 | PBH-test : $\begin{cases} \text{No } \mathbf{w}_i \text{ exists such that } \mathbf{w}_i^T\mathbf{G} = 0 \text{ or} \\ \text{rank}[z\mathbf{I} - \mathbf{F} \quad \mathbf{G}] = n \text{ for all } z \end{cases}$ |
| CC4 | For diagonal system with distinct eogmevalues:<br>No zero-rows in **B** | RD4 | For a diagonal system with distinct eigenvalues:<br>No zero-rows in **G** |

Observability Overview

| | The continuous system: $\begin{array}{l}\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}\\ \mathbf{y} = \mathbf{Cx} + \mathbf{Du}\end{array}$<br>is observable if and only if: | | The discrete system: $\begin{array}{r}\mathbf{x}\,(k+1) = \mathbf{Fx}(k) + \mathbf{Gu}(k)\\ \mathbf{y}\,(k) = \mathbf{Cx}(k) + \mathbf{Du}(k)\end{array}$<br>is observable if and only if: |
|---|---|---|---|
| OC1 | $\mathbf{W}_o(t_0, t_f) = \int_{t_0}^{t_f} \phi^T(t, t_0)\mathbf{C}^T(t)\mathbf{C}(t)\phi(t, t_0)dt$<br>is regular | OD1 | $\mathbf{W}_0(k_0, k_f) = \sum_{i=k_0}^{k_f-1} \phi^T(i, k_0)\mathbf{C}^T(k)\mathbf{C}(k)\phi(i, k_0)$<br>is regular |
| OC2 | The matrix<br>$\mathbf{M}_o^T = \left[\mathbf{C}^T \ (\mathbf{CA})^T \ \ (\mathbf{CA}^2)^T \ldots (\mathbf{CA}^{n-1})^T\right]^T$<br>has full rank | OD2 | The matrix<br>$\mathbf{M}_o^T = \left[\mathbf{C}^T \quad (\mathbf{CF})^T \quad (\mathbf{CF}^2)^T \quad \ldots \quad (\mathbf{CF}^{n-1})^T\right]^T$<br>has full rank |
| OC3 | PBH-test : $\begin{cases} \text{No } \mathbf{v}_i \text{ exists such that } \mathbf{Cv}_i = 0 \ \text{ or} \\ \text{rank}\begin{bmatrix}\mathbf{C}\\ s\mathbf{I}-\mathbf{A}\end{bmatrix} = n \text{ for all } s \end{cases}$ | OD3 | PBH-test : $\begin{cases} \text{No } \mathbf{v}_i \text{ exists such that } \mathbf{Cv}_i = 0 \ \text{ or} \\ \text{rank}\begin{bmatrix}\mathbf{C}\\ z\mathbf{I}-\mathbf{F}\end{bmatrix} = n \text{ for all } z \end{cases}$ |
| OC4 | For a diagonal system with distinct eigenvalues:<br>No zero-columns in **C** | OD4 | For a diagonal system with distinct eigenvalues:<br>No zero-rows in **C** |

As the starting point, take the diagonal transformation of a nondefective continuous time LTI system and for such a system, apply Eq. (3.133) with $\mathbf{P}^{-1} = \mathbf{M}$ (the modal matrix). Carrying this through one obtains for the state transition matrix,

$$e^{\mathbf{A}t} = Me^{\Lambda t}\mathbf{M}^{-1}. \tag{3.328}$$

The complete solution (3.52) of the state equation can now be written

$$\begin{aligned} \mathbf{x}(t) &= e^{\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{Bu}(\tau)d\tau \\ &= \mathbf{M}e^{\Lambda t}\mathbf{M}^{-1}\mathbf{x}_0 + \int_0^t \mathbf{M}e^{\Lambda(t-\tau)}\mathbf{M}^{-1}\mathbf{Bu}(\tau)d\tau. \end{aligned} \tag{3.329}$$

If expressions (3.135) and (3.246) are used here, the zero-input solution becomes

$$\begin{aligned} \mathbf{x}_{u=0}(t) &= \mathbf{M}e^{\Lambda t}\mathbf{M}^{-1}\mathbf{x}_0 \\ &= \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} e^{\lambda_1 t} & 0 & \dots & 0 \\ 0 & e^{\lambda_2 t} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & e^{\lambda_n t} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_n^T \end{bmatrix} \mathbf{x}_0 \\ &= \sum_{m=1}^{n} \mathbf{v}_m e^{\lambda_m t}\mathbf{w}_m^T\mathbf{x}_0. \end{aligned} \tag{3.330}$$

Since $e^{\lambda_m t}$ as well as $\mathbf{w}_m^T\mathbf{x}_0$ are scalars, (3.330) can be written

$$\mathbf{x}_{u=0}(t) = \sum_{m=1}^{n} (\mathbf{w}_m^T\mathbf{x}_0 e^{\lambda_m t})\mathbf{v}_m. \tag{3.331}$$

Following the same lines, the complete solution is

$$\mathbf{x}(t) = \sum_{m=1}^{n} (\mathbf{w}_m^T\mathbf{x}_0 e^{\lambda_m t})\mathbf{v}_m + \sum_{m=1}^{n} \mathbf{v}_m \left( \int_0^t e^{\lambda_m (t-\tau)}\mathbf{w}_m^T\mathbf{Bu}(\tau)d\tau \right). \tag{3.332}$$

Note that the integral contained in the right hand term is also a scalar.

A similar treatment of the discrete time solution (3.98) using (3.134) leads to essentially the same result,

$$\mathbf{x}(k) = \sum_{m=1}^{n} (\mathbf{w}_m^T\mathbf{x}_0\lambda_m^k)\mathbf{v}_m + \sum_{m=1}^{n} \mathbf{v}_m \left( \sum_{i=0}^{k-1} \lambda_m^{k-1-i}\mathbf{w}_m^T\mathbf{Gu}(i) \right), \tag{3.333}$$

the expressions in the parentheses again being scalars.

These particular forms of the solutions are called *modal decompositions*.

It is obvious that the contribution to the solutions originating from the eigenvalue $\lambda_m$ (or rather, from the natural mode $e^{\lambda_m t}$ or $\lambda_m^k$) will lie in the direction of the corresponding right eigenvector $\mathbf{v}_m$.
It is also obvious that if

$$\mathbf{w}_m^T\mathbf{B} \neq 0 \quad \text{or} \quad \mathbf{w}_m^T\mathbf{G} \neq 0 \quad \text{for all } m$$

then all natural modes will be excited by the input. As has been seen, this is the same as saying that the system is controllable/reachable (see controllability/reachability theorems CC3/RD3).

The output becomes (the **D**-matrix is of no importance here and is set to zero)

$$\mathbf{y}(t) = \sum_{m=1}^{n}(\mathbf{w}_m^T\mathbf{x}_0 e^{\lambda_m t})\mathbf{C}\mathbf{v}_m + \sum_{m=1}^{n}\mathbf{C}\mathbf{v}_m\left(\int_0^t e^{\lambda_m(t-\tau)}\mathbf{w}_m^T\mathbf{B}\mathbf{u}(\tau)d\tau\right) \tag{3.334}$$

and

$$\mathbf{y}(k) = \sum_{m=1}^{n}(\mathbf{w}_m^T\mathbf{x}_0\lambda_m^k)\mathbf{C}\mathbf{v}_m + \sum_{m=1}^{n}\mathbf{C}\mathbf{v}_m\left(\sum_{i=0}^{k-1}\lambda_m^{k-1-i}\mathbf{w}_m^T\mathbf{G}\mathbf{u}(i)\right). \tag{3.335}$$

In both terms on the right side of the equal sign it can be seen that if

$$\mathbf{C}\mathbf{v}_m \neq 0 \quad \text{for all} \quad m$$

then all natural modes will be present in the output. According to observability theorem OC3 or OD3, this is equivalent to observability.

### *3.8.11 Controllable/Reachable Subspace Decomposition*

If a system is not controllable/reachable, it is possible by a suitable similarity transformation to *decompose* the system into controllable and noncontrollable parts. The theorem supporting this possibility will be stated here without proof. The proof can be found in Kwakernaak and Sivan (1972).

**Decomposition Theorem 1**

If the system with the matrices **A**, **B** and **C** is not controllable (i.e., **A** is of dimension $n \times n$ and $rank(\mathbf{M}_c) = p < n$) then a similarity transformation can be found given by $\mathbf{z} = \mathbf{Q}^{-1}\mathbf{x}$ such that the transformed matrices have the form,

$$\mathbf{A}_t = \mathbf{Q}^{-1}\mathrm{AQ} = \begin{bmatrix} \mathbf{A}_c & \mathbf{A}_{12} \\ 0 & \mathbf{A}_{nc} \end{bmatrix}, \mathbf{B}_t = \mathbf{Q}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{B}_c \\ 0 \end{bmatrix}, \mathbf{C}_t = \mathbf{C}\mathbf{Q} = [\,\mathbf{C}_c \quad \mathbf{C}_{nc}\,], \tag{3.336}$$

where $dim(\mathbf{A}_c) = p \times p$, $dim(\mathbf{B}_c) = p \times m$, $dim(\mathbf{C}_c) = r \times p$ and where the matrix pair $\{\mathbf{A}_c, \mathbf{B}_c\}$ is controllable,
and where

$$\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \mathbf{C}_c(s\mathbf{I} - \mathbf{A}_c)^{-1}\mathbf{B}_c = \mathbf{G}(s). \tag{3.337}$$

The transformation matrix $\mathbf{Q}$ can be generated as follows:

$\mathbf{M}_c$ has dimension $n \times nm$ and since its rank is $p$, $p$ linearly independent columns can be found among the columns of $\mathbf{M}_c$. Suppose that these columns are $\mathbf{c}_1, \mathbf{c}_2, \ldots \mathbf{c}, \mathbf{c}_p$. Then choose $n - p$ column vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_{n-p}$ such that the $n \times n$ matrix,

$$\mathbf{Q} = [\,\mathbf{c}_1 \quad \mathbf{c}_2 \quad \ldots \quad \mathbf{c}_p \quad \mathbf{v}_1 \quad \mathbf{v}_2 \quad \ldots \quad \mathbf{v}_{n-p}\,], \tag{3.338}$$

becomes nonsingular.

***Example 3.20.* System Controllability Decomposition**

Consider the 2 × 2 system,

$$\mathbf{A} = \begin{bmatrix} -5 & -10 & 10 \\ 2 & -1 & -2 \\ 0 & -4 & 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 4 \\ 1 & 0 \\ 1 & 2 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{3.339}$$

The controllability matrix is readily computed to be

$$\mathbf{M}_c = \begin{bmatrix} 1 & 4 & -5 & 0 & 5 & -20 \\ 1 & 0 & -1 & 4 & -3 & -8 \\ 1 & 2 & -3 & 2 & 1 & -14 \end{bmatrix}.$$

One finds that $rank(\mathbf{M}_c) = 2$ and the system is not controllable.

It is obvious, that the first two columns of $\mathbf{M}_c$ are linearly independent and if $\mathbf{v}_1 = [1 \; 0 \; 0]^T$ is selected, it is found that,

$$\mathbf{Q} = \begin{bmatrix} 1 & 4 & 1 \\ 1 & 0 & 0 \\ 1 & 2 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{Q}^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -5.5 & 0.5 \\ 1 & 1 & -2 \end{bmatrix},$$

and then

$$
\begin{aligned}
\mathbf{A}_t &= \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \left[\begin{array}{cc|c} 1 & 4 & 2 \\ -1 & -1 & -1 \\ \hline 0 & 0 & -3 \end{array}\right], \\
\mathbf{B}_t &= \mathbf{Q}^{-1}\mathbf{B} = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ \hline 0 & 0 \end{array}\right], \\
\mathbf{C}_t &= \mathbf{C}\mathbf{Q} = \left[\begin{array}{cc|c} 1 & 4 & 1 \\ 1 & 0 & 0 \end{array}\right].
\end{aligned}
\tag{3.340}
$$

Now, clearly,

$$
\mathbf{A}_c = \begin{bmatrix} -1 & 4 \\ -1 & -1 \end{bmatrix}, \mathbf{B}_c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{C}_c = \begin{bmatrix} 1 & 4 \\ 1 & 0 \end{bmatrix}. \tag{3.341}
$$

It can also be seen that the noncontrollable part of the system is described by the state equation,

$$
\dot{z}_3 = -3z_3. \tag{3.342}
$$

The state $z_3$ is decoupled from the other states as well as from both inputs and it is therefore clearly not controllable. In contrast to this, it is easy to verify that the system with the matrices $\{\mathbf{A}_c, \mathbf{B}_c\}$ is controllable.

It is also quite easy, although a bit more cumbersome, to verify that the original system (3.339) and the reduced system (3.341) have the same transfer function matrix,

$$
\mathbf{G}(s) = \begin{bmatrix} \dfrac{s-3}{s^2+2s+5} & \dfrac{4(s+2)}{s^2+2s+5} \\ \dfrac{s+1}{s^2+2s+5} & \dfrac{4}{s^2+2s+5} \end{bmatrix}.
$$

Note, that the state vector of the transformed system (3.340) can be found from the expression $\mathbf{z} = \mathbf{Q}^{-1}\mathbf{x}$. With the $\mathbf{Q}$ found above one finds

$$
\begin{aligned}
z_1 &= x_2, \\
z_2 &= -0.5x_2 + 0.5x_3, \\
z_3 &= x_1 + x_2 - 2x_3.
\end{aligned}
$$

As seen from the reduced system (3.341), the state $z_3$ is superfluous in the sense that it is not necessary for the input-output description of the system.

Finally, it is noted that the uncontrollable system has the eigenvalue $\lambda = -3$, which means that this part of the system is stable. Thus it can be concluded, that the system is *stabilizable*, see p. 124. ❐

### 3.8.12 Observable Subspace Decomposition

If a system is not observable, it is possible by a suitable similarity transformation to *decompose* the system into observable and unobservable parts. The proof of this theorem can also in this case be found in Kwakernaak and Sivan (1972).

**Decomposition Theorem 2**

If the system with the matrices **A**, **B** and **C** is not observable (i.e. **A** is of dimension $n \times n$ and $rank(\mathbf{M}_o) = p < n$), then a similarity transformation can be found, given by $\mathbf{z} = \mathbf{Px}$, such that the transformed matrices assume the form,

$$\mathbf{A}_t = \mathbf{PAP}^{-1} = \begin{bmatrix} \mathbf{A}_o & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{no} \end{bmatrix}, \mathbf{B}_t = \mathbf{PB} = \begin{bmatrix} \mathbf{B}_o \\ \mathbf{B}_{no} \end{bmatrix}, \mathbf{C}_t = \mathbf{CP}^{-1} = [\mathbf{C}_o \; \mathbf{0}], \quad (3.343)$$

where $dim(\mathbf{A}_o) = p \times p,\ dim(\mathbf{B}_o) = p \times m, dim(\mathbf{C}_o) = r \times p$,
where the matrix pair $\{\mathbf{A}_o, \mathbf{C}_o\}$ is observable,
and where

$$\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \mathbf{C}_o(s\mathbf{I} - \mathbf{A}_o)^{-1}\mathbf{B}_o = \mathbf{G}(s). \quad (3.344)$$

The transformation matrix **P** can be generated as follows:

$\mathbf{M}_o$ has dimension $nr \times n$ and since its rank is $p$, $p$ linearly independent rows can be found among the columns of $\mathbf{M}_o$. Suppose that these rows are $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_p$, then choose $n - p$ row vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-p}$ such that the $n \times n$ matrix,

$$\mathbf{P} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_p \\ \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_{(n-p)} \end{bmatrix}, \quad (3.345)$$

becomes nonsingular.

***Example 3.21.*** **System Observability Decomposition**

The system,

$$\mathbf{A} = \begin{bmatrix} -1 & -4 & 4 \\ -3 & -4 & 6 \\ -3 & -5 & 7 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 3 & 0 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}, \ \mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 3 & -2 \end{bmatrix}, \tag{3.346}$$

is controllable but not observable, as the $(\mathbf{M}_o) = 2$.

The observability matrix is

$$\mathbf{M}_o = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 3 & -2 \\ -3 & -4 & 6 \\ -4 & -6 & 8 \\ -3 & -2 & 6 \\ -2 & 0 & 4 \end{bmatrix}.$$

The first two rows of $\mathbf{M}_o$ re linearly independent and these rows are selected as the first two rows of the transformation matrix. If one selects $\mathbf{v}_1 = [1\ 0\ 0]$, $\mathbf{P}$ becomes regular:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 3 & -2 \\ 1 & 0 & 0 \end{bmatrix}, \ \mathbf{P}^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1,5 & -0,5 & 0,5 \end{bmatrix}.$$

Then one finds that

$$\begin{aligned} \mathbf{A}_t = \mathbf{PAP^{-1}} &= \left[\begin{array}{cc:c} 5 & -3 & 0 \\ 6 & -4 & 0 \\ \hdashline 2 & -2 & 1 \end{array}\right], \\ \mathbf{B}_t = \mathbf{PB} &= \left[\begin{array}{cc} 1 & 1 \\ 2 & 1 \\ \hdashline 3 & 0 \end{array}\right], \\ \mathbf{C}_t = \mathbf{CP^{-1}} &= \left[\begin{array}{cc:c} 1 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right]. \end{aligned} \tag{3.347}$$

and

$$\mathbf{A}_o = \begin{bmatrix} 5 & -3 \\ 6 & -4 \end{bmatrix}, \ \mathbf{B}_o = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}, \ \mathbf{C}_o = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.348}$$

The second order system (3.348) is controllable as well as observable.

The system (3.348) as well as the original system (3.346) has the transfer function matrix,

$$\mathbf{G}(s) = \begin{bmatrix} \dfrac{1}{s+1} & \dfrac{1}{s-2} \\ \dfrac{2}{s+1} & \dfrac{1}{s-2} \end{bmatrix}. \tag{3.349}$$

All the transfer functions in (3.349) are of first order but nevertheless the system is second order as seen from the controllable and observable state space model (3.348).

The eigenvalues of the original system (3.346) are

$$\lambda_{\mathbf{A}} = \begin{cases} -1 \\ 1 \\ 2 \end{cases}$$

and for the reduced system,

$$\lambda_{\mathbf{A}_0} = \begin{cases} -1 \\ 2 \end{cases} .$$

The remaining unstable eigenvalue, $\lambda = 1$, must belong to the unobservable part of the system and it is therefore *not detectable*. ❐

## 3.9 Canonical Forms

Depending on the controllability/reachability and observability properties of a system, it is possible to construct state space models of a particularly simple structure. These special model forms are called *canonical* or *companion forms*. Below some important canonical forms for SISO systems will be presented. Similar forms can be defined for MIMO systems but they do not have a form which is as simple as those for SISO systems. Their practical use is thus somewhat more limited. The MIMO cases are treated in Kailath (1980).

### *3.9.1 Controller Canonical Form*

In Sect. 2.3 it was demonstrated how a state model could be derived from a SISO transfer function. Using the method of Sect. 2.3, the state space model turned out to be on a special form which was called the phase variable form or the companion form 1, see Eq. (2.46) and (2.47). This form is also known by another name: the *controller canonical form*.

It was seen that a proper transfer function leads to a state space model with a nonzero direct transfer matrix **D**. If the transfer function is strictly proper, the **D**-matrix becomes zero.

If a controllable SISO-system is considered,

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{B}u,\\ y &= \mathbf{Cx} + \mathbf{D}u,\end{aligned} \tag{3.350}$$

it is known that all the $n$ columns of the controllability matrix,

$$\mathbf{M}_c = \begin{bmatrix} \mathbf{B} & \mathbf{AB} & \mathbf{A}^2\mathbf{B} & \ldots & \mathbf{A}^{-1} & \mathbf{B} \end{bmatrix}, \tag{3.351}$$

are linearly independent.

The characteristic polynomial for the system is

$$P_{ch,\mathbf{A}} = det(\lambda \mathbf{I} - \mathbf{A}) = \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0. \tag{3.352}$$

Now a set of $n$ column vectors can be defined as follows:

$$\begin{aligned}
&\mathbf{p}_1 = \mathbf{B},\\
&\mathbf{p}_2 = \mathbf{Ap}_1 + a_{n-1}\mathbf{p}_1 = \mathbf{AB} + a_{n-1}\mathbf{B},\\
&\mathbf{p}_3 = \mathbf{Ap}_2 + a_{n-2}\mathbf{p}_1 = \mathbf{A}^2\mathbf{B} + a_{n-1}\mathbf{AB} + a_{n-2}\mathbf{B},\\
&\quad\vdots\\
&\mathbf{p}_n = \mathbf{Ap}_{n-1} + a_1\mathbf{p}_1 = \mathbf{A}^{n-1}\mathbf{B} + a_{n-1}\mathbf{A}^{n-2}\mathbf{B} + \ldots + a_2\mathbf{AB} + a_1\mathbf{B}.
\end{aligned} \tag{3.353}$$

It can be shown that the **p**-vectors are linearly independent because the columns of $\mathbf{M}_c$ are linearly independent. Consequently, defining the square matrix,

$$\mathbf{P} = [\mathbf{p}_n\ \mathbf{p}_{n-1}\ \mathbf{p}_{n-2}\ \cdots\ \mathbf{p}_1], \tag{3.354}$$

it is known that $\mathbf{P}^{-1}$ exists.

Applying the Cayley Hamilton theorem,

$$\begin{aligned}
\mathbf{Ap}_n &= (\mathbf{A}^n + a_{n-1}\mathbf{A}^{n-1} + \ldots + a_2\mathbf{A}^2 + a_1\mathbf{A} + a_0\mathbf{I})\mathbf{B} - a_0\mathbf{B}\\
&= -a_0\mathbf{B} = -a_0\mathbf{p}_1 = \mathbf{P}[0\ 0\ 0\ \ldots\ 0\ -a_0]^T,\\
\mathbf{Ap}_{n-1} &= \mathbf{p}_n - a_1\mathbf{p}_1 = \mathbf{P}[1\ 0\ 0\ \ldots\ 0 - a_1]^T,\\
\mathbf{Ap}_{n-2} &= \mathbf{p}_{n-1} - a_2\mathbf{p}_1 = \mathbf{P}[0\ 1\ 0\ \ldots\ 0 - a_2]^T,\\
&\ \ \vdots\\
\mathbf{Ap}_1 &= \mathbf{p}_2 - a_{n-1}\mathbf{p}_1 = \mathbf{P}[0\ 0\ 0 \ldots\ 1 - a_{n-1}]^T,
\end{aligned} \tag{3.355}$$

this can be written

$$\mathbf{A}[\mathbf{p}_n\ \mathbf{p}_{n-1}\ \mathbf{p}_{n-2}\ \cdots\ \mathbf{p}_1] \\ = \mathbf{P}\begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ : & : & : & \dots & : & : \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-2} & -a_{n-1} \end{bmatrix} = \mathbf{P}\mathbf{A}_{cc} \tag{3.356}$$

which means that

$$\mathbf{AP} = \mathbf{PA}_{cc} \tag{3.357}$$

or

$$\mathbf{A}_{cc} = \mathbf{P}^{-1}\mathbf{AP}. \tag{3.358}$$

The last result shows that if the similarity transformation $\mathbf{z} = \mathbf{P}^{-1}\mathbf{x}$ is used, the system matrix of the new model representation will be

$$\mathbf{A}_{cc} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ : & : & : & \dots & : & : \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-2} & -a_{n-1} \end{bmatrix}. \tag{3.359}$$

Which gives the immediate result that

$$\mathbf{B} = \mathbf{P}[0\ 0\ 0\ \dots\ 0\ 1]^T \tag{3.360}$$

which gives

$$\mathbf{P}^{-1}\mathbf{B} = \mathbf{B}_{cc} = \begin{bmatrix} 0 \\ 0 \\ : \\ 0 \\ 1 \end{bmatrix}. \tag{3.361}$$

The output matrix becomes

$$\mathbf{C}_{cc} = \mathbf{CP} = [b_0\ b_1\ b_2\ \dots\ b_{n-1}]. \tag{3.362}$$

Finally one find that

$$\mathbf{D}_{cc} = \mathbf{D}. \tag{3.363}$$

The controllability matrix of the controller canonical model is easily computed to be

$$\mathbf{M}_{c,cc} = [\mathbf{B}_{cc} \ \mathbf{A}_{cc}\mathbf{B}_{cc} \ \mathbf{A}_{cc}^2\mathbf{B}_{cc} \ \ldots \ \mathbf{A}_{cc}^{n-1}\mathbf{B}_{cc}]$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 & \ldots \ 0 \ 1 \\ 0 & 0 & 0 & 0 & \ldots \ 1 \ X \\ : & : & : & : & : \ \ : \ \ : \\ 0 & 0 & 0 & 1 & \ldots \ X \ X \\ 0 & 0 & 1 & -a_{n-1} & \ldots \ X \ X \\ 0 & 1 & -a_{n-1} & a_{n-1}^2 - a_{n-2} & \ldots \ X \ X \\ 0 & -a_{n-1} & a_{n-1}^2 - a_{n-2} & -a_{n-1}^3 + 2a_{n-1}a_{n-2} - a_{n-3} & \ldots \ X \ X \end{bmatrix} \tag{3.364}$$

where the $X$-elements are functions of the coefficients $a_i$. The determinant of $\mathbf{M}_{c,cc}$ is always equal to 1 or $-1$ independent of the $a_i$-coefficients and that is, of course, in accordance with the precondition that the system is controllable.

***Example 3.22.* Controller Canonical Form Transformation**

Consider a continuous time SISO system with the matrices,

$$\mathbf{A} = \begin{bmatrix} 1 & 6 & -3 \\ -1 & -1 & 1 \\ -2 & 2 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{C} = [0 \quad 0 \quad 1].$$

The controllability matrix is

$$\mathbf{M}_c = [\mathbf{B} \quad \mathbf{AB} \quad \mathbf{A}^2\mathbf{B}] = \begin{bmatrix} 1 & 4 & -2 \\ 1 & -1 & -3 \\ 1 & 0 & -10 \end{bmatrix}$$

and since the determinant is $det(\mathbf{M}_c) = 36$, the system is clearly controllable.

The characteristic polynomial can be calculated to be

$$P_{ch,\mathbf{A}} = det \begin{bmatrix} \lambda - 1 & -6 & 3 \\ 1 & \lambda + 1 & -1 \\ 2 & -2 & \lambda \end{bmatrix} = \lambda^3 - 3\lambda + 2$$

which means that

$$
\begin{aligned}
a_2 &= 0,\\
a_1 &= -3,\\
a_0 &= 2.
\end{aligned}
$$

The eigenvalues are

$$
\lambda_{\mathbf{A}} = \begin{cases} 1 \\ 1 \\ -2 \end{cases}
$$

so the system is unstable.

The **p**-vectors (3.353) can readily be found to be,

$$
\begin{aligned}
\mathbf{p}_1 &= \mathbf{B} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},\\
\mathbf{p}_2 &= \mathbf{AB} + a_2\mathbf{B} = \mathbf{AB} = \begin{bmatrix} 4 \\ -1 \\ 0 \end{bmatrix},\\
\mathbf{p}_3 &= \mathbf{A}^2\mathbf{B} + a_2\mathbf{AB} + a_1\mathbf{B} = \mathbf{A}^2\mathbf{B} - 3\mathbf{B} = \begin{bmatrix} -5 \\ -6 \\ -13 \end{bmatrix},
\end{aligned}
$$

and the matrix **P** and its inverse are

$$
\mathbf{P} = [\mathbf{p}_3 \quad \mathbf{p}_2 \quad \mathbf{p}_1] = \begin{bmatrix} -5 & 4 & 1 \\ -6 & -1 & 1 \\ -13 & 0 & 1 \end{bmatrix}, \quad \mathbf{P}^{-1} = \begin{bmatrix} 0.02778 & 0.11111 & -0.13889 \\ 0.19444 & -0.22222 & 0.02778 \\ 0.36111 & 1.44444 & -0.80556 \end{bmatrix}.
$$

The similarity transformation leads to the new model representation in the controller canonical form,

$$
\mathbf{A}_{cc} = \mathbf{P}^{-1}\mathbf{AP} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & 3 & 0 \end{bmatrix}, \ \mathbf{B}_{cc} = \mathbf{P}^{-1}\mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \ \mathbf{C}_{cc} = \mathbf{CP} = [\,-13 \quad 0 \quad 1\,].
$$

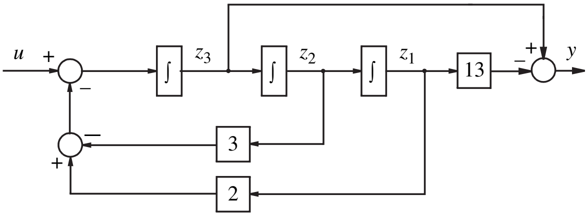


**Fig. 3.27** Block diagram for the controller canonical form

The block diagram for the controller canonical representation is shown on Fig. 3.27.

As expected, the block diagram has the same structure as the block diagram in Fig. 2.14. This kind of block diagram, containing only constants, integrators and summers, is often used as a basis for setting up analog or digital computers for simulation purposes.

It is known that the system on Fig. 3.27 is controllable. The observability matrix is

$$\mathbf{M}_o = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ -2 & 2 & 0 \\ -4 & -14 & 8 \end{bmatrix}$$

and since $det(\mathbf{M}_o) = 36$, the system is also observable. ❐

As it has been pointed out previously, the controller canonical form is very efficient in terms of number of parameters. A strictly proper $n$'th order SISO system will in general have $n^2 + 2n$ parameters in its matrices. The corresponding controller canonical form will at most have $2n$ parameters different from zero or one, precisely the same as in the transfer function.

### 3.9.2 Observer Canonical Form

For an observable SISO system,

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{B}u, \\ y &= \mathbf{Cx} + \mathbf{D}u, \end{aligned} \tag{3.365}$$

with the characteristic equation,

$$P_{ch,\mathbf{A}} = det(\lambda\mathbf{I} - \mathbf{A}) = \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0, \tag{3.366}$$

the following set of linearly independent $n$-dimensional row vectors can be defined:

$$\begin{aligned}
\mathbf{q}_1^T &= \mathbf{C}, \\
\mathbf{q}_2^T &= \mathbf{CA} + a_{n-1}\mathbf{C}, \\
\mathbf{q}_3^T &= \mathbf{CA}^2 + a_{n-1}\mathbf{CA} + a_{n-2}\mathbf{C}, \\
&\vdots \\
\mathbf{q}_n^T &= \mathbf{CA}^{n-1} + a_{n-1}\mathbf{CA}^{n-2} + \ldots + a_2\mathbf{CA} + a_1\mathbf{C}.
\end{aligned} \tag{3.367}$$

Further, define the nonsingular $n \times n$ matrix,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{q}_n^T \\ \mathbf{q}_{n-1}^T \\ \vdots \\ \mathbf{q}_1^T \end{bmatrix}. \tag{3.368}$$

The Q-matrix for the similarity transformation $\mathbf{z} = \mathbf{Qx}$ can be used to obtain the following matrices for the alternative model representation, which is called the *observer canonical form*,

$$\mathbf{A}_{oc} = \mathbf{QAQ}^{-1} = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 & -a_0 \\ 1 & 0 & 0 & \ldots & 0 & -a_1 \\ 0 & 1 & 0 & \ldots & 0 & -a_2 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & -a_{n-2} \\ 0 & 0 & 0 & \ldots & 1 & -a_{n-1} \end{bmatrix}, \tag{3.369}$$

$$\mathbf{B}_{oc} = \mathbf{OB} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{bmatrix}, \tag{3.370}$$

$$\mathbf{C}_{oc} = \mathbf{CQ}^{-1} = [0 \;\; 0 \;\; \ldots \;\; 0 \;\; 1], \tag{3.371}$$

$$\mathbf{D}_{oc} = \mathbf{D}. \tag{3.372}$$

The observability matrix for the observer canonical form is

$$\mathbf{M}_{o,oc} = \begin{bmatrix} \mathbf{C}_{oc} \\ \mathbf{C}_{oc}\mathbf{A}_{oc} \\ \mathbf{C}_{oc}\mathbf{A}_{oc}^2 \\ \vdots \\ \mathbf{C}_{oc}\mathbf{A}_{oc}^{n-1} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 & 1 & -a_{n-1} \\ 0 & 0 & \dots & 0 & 1 & -a_{n-1} & a_{n-1}^2 - a_{n-2} \\ 0 & 0 & \dots & 1 & -a_{n-1} & a_{n-1}^2 - a_{n-2} & -a_{n-1}^3 + 2a_{n-1}a_{n-2} - a_{n-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \dots & X & X & X & X \\ 1 & X & \dots & X & X & X & X \end{bmatrix}. \quad (3.373)$$

Since $det(\mathbf{M}_{o,oc})$ is equal to 1 or $-1$, the system is observable.

The observer canonical form is identical to the companion form 2 from Sect. 2.3.2.

***Example 3.23.* Observer Canonical Form Transformation**

Now an alternative method of achieving the observer canonical form for the system of Eq. (3.365) will be considered.

If one *supposes* that one can obtain the observer canonical form (3.369)–(3.372) by a similarity transformation $\mathbf{z} = \mathbf{Q}\mathbf{x}$, then it is known that the original model (3.365) must be observable. This follows from the fact that observability is invariant under a similarity transformation, see section 3.8.7. So one knows that the observability matrices $\mathbf{M}_o$ and $\mathbf{M}_{o,oc}$, for the original model and the observer canonical form respectively, are both nonsingular. From Eq. (3.316) it is known that

$$\mathbf{M}_{o,oc} = \mathbf{M}_o\mathbf{Q}^{-1} \quad (3.374)$$

or

$$\mathbf{Q} = \mathbf{M}_{o,oc}^{-1}\mathbf{M}_o. \quad (3.375)$$

The system of Example 3.22 has the matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 6 & -3 \\ -1 & -1 & 1 \\ -2 & 2 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{C} = [\,1 \quad 0 \quad 0\,].$$

The coefficients of the characteristic polynomial are

$$\begin{aligned} a_2 &= 0, \\ a_1 &= -3, \\ a_0 &= 2 \end{aligned}$$

and the observability matrix (3.373) of the observer canonical form can be written down directly:

$$\mathbf{M}_{o,oc} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & -a_{n-1} \\ 1 & -a_{n-1} & a_{n-1}^2 - a_{n-2} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix}.$$

The transformation matrix is computed as

$$\mathbf{Q} = \mathbf{M}_{o,oc}^{-1}\mathbf{M}_o = \begin{bmatrix} -4 & -14 & 5 \\ -2 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and the matrices of the observer canonical form become

$$\mathbf{A}_{oc} = \mathbf{Q}\mathbf{A}\mathbf{Q}^{-1} = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 0 & 3 \\ 0 & 1 & 0 \end{bmatrix},$$

$$\mathbf{B}_{oc} = \mathbf{Q}\mathbf{B} = \begin{bmatrix} -13 \\ 0 \\ 1 \end{bmatrix},$$

$$\mathbf{C}_{oc} = \mathbf{C}\mathbf{Q}^{-1} = [\,0 \quad 0 \quad 1\,].$$

If the procedure of Sect. 3.9.2 is used, the same result will of course be obtained. ❐

### 3.9.3 Duality for Canonical Forms

If the matrices in Sect. 3.9.1 and 3.9.2 are compared, it is found that

$$\begin{aligned} \mathbf{A}_{cc} &= \mathbf{A}_{oc}^T, \\ \mathbf{B}_{cc} &= \mathbf{C}_{oc}^T, \\ \mathbf{C}_{cc} &= \mathbf{B}_{oc}^T, \end{aligned} \tag{3.376}$$

and it can be concluded that the controller canonical form and the observer canonical form are dual models according to the definition of Sect. 3.8.9.

### 3.9.4 Pole-zero Cancellation in SISO Systems

In the preceding sections it was shown that a SISO state space model in controller canonical form is always controllable and a model in observer canonical form is always observable. On the other hand, the controller canonical form need not be observable and the observer canonical form need not be controllable.

For a strictly proper transfer function the connection between transfer function and the state space model is

$$\begin{gathered} G(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \\ \text{or} \\ H(z) = \mathbf{C}(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G}. \end{gathered} \tag{3.377}$$

The following argument deals with the continuous time case but is equally valid for the discrete time models.

If $\mathbf{A}$ is nondefective, Eq. (3.377) can be written,

$$\begin{aligned} G(s) &= \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \mathbf{CMM}^{-1}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{MM}^{-1}\mathbf{B} \\ &= \mathbf{CM}(\mathbf{M}^{-1}(s\mathbf{I} - \mathbf{A})\mathbf{M})^{-1}\mathbf{M}^{-1}\mathbf{B} = \mathbf{CM}(s\mathbf{I} - \mathbf{M}^{-1}\mathbf{AM})^{-1}\mathbf{M}^{-1}\mathbf{B} \\ &= \mathbf{CM}(s\mathbf{I} - \Lambda)^{-1}\mathbf{M}^{-1}\mathbf{B}, \end{aligned} \tag{3.378}$$

where $\mathbf{M}$ is the modal matrix for $\mathbf{A}$. Using Eqs. (3.135) and (3.246) one arrives at the result

$$G(s) = \mathbf{C}[\mathbf{v}_1 \quad \mathbf{v}_2 \quad \ldots \quad \mathbf{v}_n](s\mathbf{I} - \Lambda)^{-1}\begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_n^T \end{bmatrix}\mathbf{B} = \sum_{i=1}^{n} \frac{\mathbf{C}\mathbf{v}_i\mathbf{w}_i^T\mathbf{B}}{s - \lambda_i}. \tag{3.379}$$

The pole/eigenvalue $\lambda_i$ will only be present in the transfer function if the products $\mathbf{C}\mathbf{v}_i$ and $\mathbf{w}_i^T\mathbf{B}$ are both nonzero. One knows from the PBH observability and controllability tests that this means that the state space model is observable and controllable. If, for some $i$, one has that $\mathbf{C}\mathbf{V}_i$ or $\mathbf{w}_i^T\mathbf{B} = 0$ (or both), then the pole/eigenvalue $\lambda_i$ will not appear in the transfer function. It is *cancelled* by a zero.

This shows that if a SISO state space model is not controllable or not observable (or neither) then zero/pole cancellation(s) will occur in the transfer function. Note that this rule is *not* valid for MIMO systems.

## 3.10 Realizability

In this section the problem of the construction of state space models from transfer function matrices for MIMO LTI systems is addressed.

The presence of a direct transfer matrix **D** in the state space model is unimportant for the treatment of this subject and therefore it is assumed that the state space models have the simpler form:

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t), \\ &\text{or} \\ \mathbf{x}(k+1) &= \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k), \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k). \end{aligned} \tag{3.380}$$

If one starts by considering the SISO case, one can return to the development in Chap. 2, Sect. 2.3. One can see directly from this section that one can construct (or *realize*) a state space model with $\mathbf{D} = 0$ for any strictly proper SISO transfer function $G(s)$ or $H(z)$, i.e., any transfer function where the numerator polynomial has a smaller degree than the denominator polynomial. In this context the state space model (3.380) is called a *realization* of $G(s)$ or $H(z)$. Note that Sect. 2.3 dealt with continuous time systems only but construction of the companion forms works equally well for discrete time systems. This finding can be extended to MIMO systems.

**Realizability Theorem**

An LTI realization for the transfer function matrices $\mathbf{G}(s)$ or $\mathbf{H}(z)$ can be found if and only if all elements of the transfer function matrix are strictly proper transfer functions. The relation between the matrices of the realization and the transfer function matrices is

$$\begin{gathered} \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \mathbf{G}(s) \\ \text{or} \\ \mathbf{C}(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G} = \mathbf{H}(z). \end{gathered} \tag{3.381}$$

(Do not confuse the continuous time transfer function matrix $\mathbf{G}(s)$ with the discrete time input matrix $\mathbf{G}$).

To prove the ‘only if’ part of the theorem above it can be assumed that $\mathbf{G}(s)$ or $\mathbf{H}(z)$ has the realization (3.380) and then it is known from Sects. 3.2 and 3.5 that (3.381) is true. From the above argument it is also known that the elements of $\mathbf{G}(s)$ or $\mathbf{H}(z)$ are strictly proper transfer functions.

The ‘if’ part can be proved as follows. Suppose that $\mathbf{G}(s)$ or $\mathbf{H}(z)$ have strictly proper transfer functions. If one looks at all the denominator polynomials of say the elements $G_{ij}(s)$ of the $r \times m$ -dimensional $\mathbf{G}(s)$, one can find the least common polynomial containing all factors in these denominator polynomials,

$$d(s) = s^p + d_{p-1}s^{p-1} + \ldots + d_1 s + d_0. \tag{3.382}$$

Then one can write

$$d(s)\mathbf{G}(s) = \mathbf{N}_{p-1}s^{p-1} + \mathbf{N}_{p-2}s^{p-2} + \ldots + \mathbf{N}_1 s + \mathbf{N}_0 \tag{3.383}$$

where $\mathbf{N}_i$ denote constant $r \times m$ -matrices.

It is now claimed that the following state space model is a realization of $\mathbf{G}(s)$:

$$\begin{gathered}
\mathbf{A} = \begin{bmatrix} \mathbf{0}_m & \mathbf{I}_m & \mathbf{0}_m & \ldots & \mathbf{0}_m & \mathbf{0}_m \\ \mathbf{0}_m & \mathbf{0}_m & \mathbf{I}_m & \ldots & \mathbf{0}_m & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_m & \mathbf{0}_m & \mathbf{0}_m & \ldots & \mathbf{0}_m & \mathbf{I}_m \\ -d_0\mathbf{I}_m & -d_1\mathbf{I}_m & -d_2\mathbf{I}_m & \ldots & -d_{p-2}\mathbf{I}_m & -d_{p-1}\mathbf{I}_m \end{bmatrix}, \\
\mathbf{B} = \begin{bmatrix} \mathbf{0}_m \\ \mathbf{0}_m \\ \vdots \\ \mathbf{0}_m \\ \mathbf{I}_m \end{bmatrix}, \quad \mathbf{C} = [\mathbf{N}_0 \ \mathbf{N}_1 \ \ldots \ \mathbf{N}_{p-2} \ \mathbf{N}_{p-1}],
\end{gathered} \tag{3.384}$$

where $\mathbf{0}_m$ and $\mathbf{I}_m$ are the quadratic zero and the identity matrices, both of dimension $m$. As seen from $\mathbf{A}$, the state space model has the dimension $mp$.

Define the transfer function matrix:

$$\mathbf{Z}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{Z}_1(s) \\ \mathbf{Z}_2(s) \\ : \\ \mathbf{Z}_{p-1}(s) \\ \mathbf{Z}_p(s) \end{bmatrix}. \qquad (3.385)$$

The $\mathbf{Z}_i(s)$-matrices have the dimension $m \times m$ and $\mathbf{Z}(s)$ has the dimension $mp \times m$. If multiplying (3.385) by $(s\mathbf{I} - \mathbf{A})$, one finds that

$$s\mathbf{Z} = \mathbf{A}\mathbf{Z} + \mathbf{B} \qquad (3.386)$$

or

$$\begin{bmatrix} s\mathbf{Z}_1 \\ s\mathbf{Z}_2 \\ : \\ s\mathbf{Z}_{p-1} \\ s\mathbf{Z}_p \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_2 \\ \mathbf{Z}_3 \\ : \\ \mathbf{Z}_{p-1} \\ \mathbf{Z}_p \\ -d_0\mathbf{Z}_1 - d_1\mathbf{Z}_2 - \ldots - d_{p-2}\mathbf{Z}_{p-1} - d_{p-1}\mathbf{Z}_p + \mathbf{I}_m \end{bmatrix}. \qquad (3.387)$$

This equation gives two relations:

$$s\mathbf{Z}_i = \mathbf{Z}_{i+1}, \; i = 1, 2, \ldots, p-1 \qquad (3.388)$$

and

$$s\mathbf{Z}_p + d_0\mathbf{Z}_1 + d_1\mathbf{Z}_2 + \ldots + d_{p-1}\mathbf{Z}_p = \mathbf{I}_m. \qquad (3.389)$$

If successively the relations (3.388) are inserted into (3.389), it is found that

$$\mathbf{Z}_1 = \frac{1}{d(s)}\mathbf{I}_m \qquad (3.390)$$

and, using (3.388) again,

$$\mathbf{Z}(s) = \frac{1}{d(s)} \begin{bmatrix} \mathbf{I}_m \\ s\mathbf{I}_m \\ s^2\mathbf{I}_m \\ : \\ s^{p-1}\mathbf{I}_m \end{bmatrix}. \qquad (3.391)$$

Finally premultiplying by **C** one obtains

$$\begin{aligned}\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B} &= \mathbf{C}\mathbf{Z}(s)\\ &= \frac{1}{d(s)}\left[\mathbf{N}_0\ \mathbf{N}_1 s\ \dots\ \mathbf{N}_{p-2}s^{p-2}\ \mathbf{N}_{p-1}s^{p-1}\right] = \mathbf{G}(s).\end{aligned} \tag{3.392}$$

***Example 3.24*. Realizability of a MIMO System**

Consider a MIMO LTI system with the transfer function matrix:

$$\mathbf{G}(s) = \begin{bmatrix} \dfrac{1}{(s+1)^2} & \dfrac{3}{s-2} \\ \dfrac{s+3}{(s+1)(s-2)} & \dfrac{2}{(s-2)^2} \end{bmatrix}. \tag{3.393}$$

The system has two inputs and two outputs, so $r = m = 2$.

The $d$-polynomial must contain all factors of the four denominator polynomials,

$$d(s) = (s+1)^2(s-2)^2 = s^4 - 2s^3 - 3s^2 + 4s + 4.$$

Since $\mathbf{0}_m = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ and $\mathbf{I}_m = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, the matrices **A** and **B** of the state space realization (3.384) are

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -4 & 0 & -4 & 0 & 3 & 0 & 2 & 0 \\ 0 & -4 & 0 & -4 & 0 & 3 & 0 & 2 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.394}$$

Multiplying $\mathbf{G}(s)$ by $d(s)$ yields,

$$d(s)\mathbf{G}(s) = \begin{bmatrix} s^2 - 4s + 4 & 3s^3 - 9s - 6 \\ s^3 + 2s^2 - 5s - 6 & 2s^2 + 4s + 2 \end{bmatrix}$$

which can be written,

$$\begin{aligned} d(s)\mathbf{G}(s) &= \mathbf{N}_3 s^3 + \mathbf{N}_2 s^2 + \mathbf{N}_1 s + \mathbf{N}_0 \\ &= \begin{bmatrix} 0 & 3 \\ 1 & 0 \end{bmatrix} s^3 + \begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix} s^2 + \begin{bmatrix} -4 & -9 \\ -5 & 4 \end{bmatrix} s + \begin{bmatrix} 4 & -6 \\ -6 & 2 \end{bmatrix}. \end{aligned}$$

One can now set up the remaining output matrix,

$$\mathbf{C} = \begin{bmatrix} 4 & -6 & -4 & -9 & 1 & 0 & 0 & 3 \\ -6 & 2 & -5 & 4 & 2 & 2 & 1 & 0 \end{bmatrix}, \tag{3.395}$$

and a state space model corresponding to (3.393) is established. The state space realization is of 8'th order which is a result of the method which has been used.

The order of the transfer functions in (3.393) is one or two, so one might suspect that the order of the state space model is greater than necessary. That is in fact the case. One can however find state space models with the same transfer function matrix but of lower order than 8.

If the state space model is tested for controllability, it is found that the controllability matrix $\mathbf{M}_c$ has the rank 8 and the system is certainly controllable. The observability matrix $\mathbf{M}_o$ has, however, only rank 4, so the model is not observable. As will be seen later the model loses observability because it has superfluous states. ❒

### 3.10.1 Minimality

If one generates a state space realization of a system with a given transfer function matrix $\mathbf{G}(s)$ (or $\mathbf{H}(z)$), one knows that this realization is not unique. One cannot be sure that the realization has the lowest possible order, i.e. that the system matrix $\mathbf{A}$ (or $\mathbf{F}$) of the state space model has the smallest possible dimension. To be able to handle such realizability problems, it is useful to state a formal definition of minimality. The definition can for instance be found in Kailath (1980).

**Minimality Definition**

A *minimal realization* corresponding to $\mathbf{G}(s)$ is a state space model which has the smallest possible system matrix $\mathbf{A}$ for all triples $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ satisfying the relation

$$\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \mathbf{G}(s). \tag{3.396}$$

The definition is similar for discrete time systems:

A minimal realization corresponding to $\mathbf{H}(z)$ is a discrete time state space model which has the smallest possible system matrix $\mathbf{F}$ for all triples $\{\mathbf{F}, \mathbf{G}, \mathbf{C}\}$ satisfying the relation

$$\mathbf{C}(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G} = \mathbf{H}(z). \tag{3.397}$$

It turns out that there is a simple connection between minimality and controllability/observability. In fact one can prove the following theorem.

**Minimality Theorem**

A realization $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ is minimal if and only if it is controllable and observable.

A discrete time realization $\{\mathbf{F}, \mathbf{G}, \mathbf{C}\}$ is minimal if and only if it is reachable and observable.

The continuous time version of the theorem will now be proved. A proof of the discrete time version can be found in Rugh (1996). First assume that $\{\mathbf{A}, \mathbf{B}\}$ are not controllable. If that is the case then by application of the result from Sect. 3.8.11, one could find another realization of smaller order with the same transfer function matrix. But then $\{\mathbf{A}, \mathbf{B}\}$ could not have been be minimal and that proves the 'only if' part of the theorem.

To prove the 'if' part, it is assumed that the $n$'th order realization $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ is controllable and observable and that one can find another controllable and observable realization $\{\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1\}$ with the order $n_1 < n$.

The two realizations have the same transfer function matrix so one knows that

$$\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \mathbf{C}_1(s\mathbf{I} - \mathbf{A}_1)^{-1}\mathbf{B}_1. \tag{3.398}$$

The impulse response (3.56) must therefore also be the same. With $\mathbf{D} = \mathbf{0}$ one finds that

$$\mathbf{g}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B} = \mathbf{C}_1 e^{\mathbf{A}_1 t}\mathbf{B}_1. \tag{3.399}$$

Stepwise differentiation with respect to time yields:

$$\begin{aligned}
\mathbf{C}\mathbf{A}e^{\mathbf{A}t}\mathbf{B} &= \mathbf{C}_1\mathbf{A}_1 e^{\mathbf{A}_1 t}\mathbf{B}_1,\\
\mathbf{C}\mathbf{A}^2 e^{\mathbf{A}t}\mathbf{B} &= \mathbf{C}_1\mathbf{A}_1^2 e^{\mathbf{A}_1 t}\mathbf{B}_1,\\
&\vdots\\
\mathbf{C}\mathbf{A}^n e^{\mathbf{A}t}\mathbf{B} &= \mathbf{C}_1\mathbf{A}_1^n e^{\mathbf{A}_1 t}\mathbf{B}_1.\\
&\vdots
\end{aligned} \tag{3.400}$$

If one evaluates all of these expressions for $t = \mathbf{0}$ one obtains

$$\mathbf{C}\mathbf{A}^i\mathbf{B} = \mathbf{C}_1\mathbf{A}_1^i\mathbf{B}_1 \text{ for all } i. \tag{3.401}$$

The controllability and the observability matrices for $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ are

$$\mathbf{M}_o = \begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \mathbf{C}\mathbf{A}^2 \\ \vdots \\ \mathbf{C}\mathbf{A}^{n-1} \end{bmatrix}_{nr\times n}, \quad \mathbf{M}_c = \begin{bmatrix} \mathbf{B} & \mathbf{A}\mathbf{B} & \mathbf{A}^2\mathbf{B} & \dots & \mathbf{A}^{n-1}\mathbf{B} \end{bmatrix}_{n\times nm}, \tag{3.402}$$

and multiplying the two matrices,

$$\mathbf{M}_o\mathbf{M}_c = \begin{bmatrix} \mathbf{CB} & \mathbf{CAB} & \dots & \mathbf{CA}^{n-1}\mathbf{B} \\ \mathbf{CAB} & \mathbf{CA}^2\mathbf{B} & \dots & \mathbf{CA}^n\mathbf{B} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{CA}^{n-1}\mathbf{B} & \mathbf{CA}^n\mathbf{B} & \dots & \mathbf{CA}^{2n-2}\mathbf{B} \end{bmatrix}_{nr\times nm}. \tag{3.403}$$

Similarly, for $\{\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1\}$ one calculates,

$$\mathbf{M'}_{o,1} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_1\mathbf{A}_1 \\ \mathbf{C}_1\mathbf{A}_1^2 \\ \vdots \\ \mathbf{C}_1\mathbf{A}_1^{n-1} \end{bmatrix}_{nr\times n_1}, \qquad \mathbf{M'}_{c,1} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{A}_1\mathbf{B}_1 & \mathbf{A}_1^2\mathbf{B}_1 & \dots & \mathbf{A}_1^{n-1}\mathbf{B}_1 \end{bmatrix}_{n_1\times nm}, \tag{3.404}$$

although it was assumed that $n_1 < n$. From Eq. (3.401) one can see that calculation of $\mathbf{M}'_{0,1}\mathbf{M}'_{c,1}$ gives exactly the same matrix as in (3.403), so

$$\mathbf{M}_o\mathbf{M}_c = \mathbf{M'}_{o,1}\mathbf{M'}_{c,1}. \tag{3.405}$$

Since $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ is controllable and observable, the Sylvester inequality[†] shows that

$$rank(\mathbf{M}_o) = rank(\mathbf{M}_c) = n \tag{3.406}$$

from which it follows that

$$rank(\mathbf{M}_o\mathbf{M}_c) = n. \tag{3.407}$$

$\{\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1\}$ is also controllable and observable and therefore

$$rank(\mathbf{M}'_{o,1}) = rank(\mathbf{M}'_{c,1}) = n_1 \tag{3.408}$$

and

$$rank(\mathbf{M}'_{o,1}\mathbf{M}'_{c,1}) = n_1. \tag{3.409}$$

[†] The Sylvester inequality: If the matrix A has dimension $m \times n$ and B has dimension $n \times p$, then $rank(\mathbf{A}) + rank(\mathbf{B}) - n \leq rank(\mathbf{AB}) \leq min(rank(\mathbf{A}), rank(\mathbf{B}))$

But then, from (3.405) it follows that $n_1 = n$, which clearly contradicts the assumption that $n_1 < n$. In other words, if $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ is controllable and observable, no realization of lower order exists and the theorem is proved.

***Example 3.25.*** **Minimality Analysis of a MIMO System**

By application of the minimality theorem above, a suspicion concerning the realization in Example 3.24 is confirmed. The state space model (3.394)–(3.395) is controllable but not observable and therefore it is not minimal. The observability matrix has dimension 16 × 8 and rank $p = 4$. To carry out the observable subspace decomposition of Sect. 3.8.12, one must pick the same number of linearly independent rows from $\mathbf{M}_o$. It turns out that its first four rows are linearly independent and one can set up the transformation matrix $\mathbf{P}$ as follows:

$$\mathbf{P} = \left[\begin{array}{cccccccc} 4 & -6 & -4 & -9 & 1 & 0 & 0 & 3 \\ -6 & 2 & -5 & 4 & 2 & 2 & 1 & 0 \\ 0 & -12 & 4 & -18 & -4 & 0 & 1 & 6 \\ -4 & 0 & -10 & 2 & -2 & 4 & 4 & 2 \\ \hline 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right].$$

The upper four rows are the rows from $\mathbf{M}_o$ and the rows under the dotted line were selected in such a way that $\mathbf{P}$ becomes nonsingular.

The similarity transformation (3.343) yields the matrices of the minimal realization,

$$\dot{\mathbf{z}} = \mathbf{A}_o\mathbf{z} + \mathbf{B}_o\mathbf{u},$$
$$\mathbf{y} = \mathbf{C}_o\mathbf{z},$$

where

$$\mathbf{A}_o = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 2.375 & 3.735 & 1.375 & -1.6875 \\ 0.75 & 2.75 & 0.75 & 0625 \end{bmatrix}, \mathbf{B}_o = \begin{bmatrix} 0 & 3 \\ 1 & 0 \\ 1 & 6 \\ 4 & 2 \end{bmatrix}, \mathbf{C}_o = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

It can be verified that the reduced system $\{\mathbf{A}_o, \mathbf{B}_o, \mathbf{C}_o\}$ is controllable and observable and therefore a minimal realization. It can also be verified that the system has the transfer function matrix (3.393). The state variables of the minimal model can be found using the transformation $\mathbf{z} = \mathbf{P}\mathbf{x}$.

The system $\{\mathbf{A}_o, \mathbf{B}_o, \mathbf{C}_o\}$ has the eigenvalues

$$\lambda_{\mathbf{A_0}} = \begin{cases} -1 \\ -1 \\ 2 \\ 2 \end{cases} .$$

The 8'th order state space model (3.394) has the 8 eigenvalues,

$$\lambda_{\mathbf{A}} = \begin{cases} -1 \\ -1 \\ -1 \\ -1 \\ 2 \\ 2 \\ 2 \\ 2 \end{cases} .$$

Since the unobservable part of (3.394)/ (3.395) has unstable eigenvalues, it is not detectable. ❐

***Example 3.26*. Analysis of a Hydraulic Servo Cylinder**

A simplified model of a symmetric control cylinder of a hydraulic position servo shown on Fig. 3.28 can be modelled as follows.

Expressions for the two volume flows can be written,

$$q_1 = A_c\dot{x} + \frac{V}{\beta}\dot{p}_1 + C_l(p_1 - p_2), \tag{3.410}$$

$$q_2 = A_c\dot{x} - \frac{V}{\beta}\dot{p}_2 + C_l(p_1 - p_2), \tag{3.411}$$

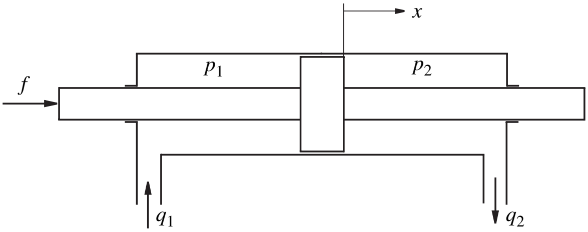


**Fig. 3.28** Hydraulic servo cylinder

where $A_c$ is the cylinder area, $V$ is the oil volume for each of the cylinder chambers, $\beta$ is the bulk modulus (the stiffness coefficient) of the oil, $p_1$ and $p_2$ are the pressures in the cylinder chambers and $C_l$ is a leakage coefficient. $x$ is the position of the piston and $f$ is an external load force.

If the cylinder is fed by a perfect, symmetric servo valve, the volume flows can be considered equal and if the valve is linear and very fast compared to the dynamics of the rest of the system, the simple relationship,

$$q_1 = q_2 = ku, \tag{3.412}$$

can be assumed where $u$ is the input voltage to the servo valve and $k$ is a proportionality constant.

The model is completed by using Newton's second law for the total mass $M$ of piston and piston rods:

$$M\ddot{x} = f + A_c(p_1 - p_2) - \mathbf{C}_f\dot{x}. \tag{3.413}$$

The last term of Eq. (3.413) is the viscous friction between cylinder and piston.

The set of equations (3.410)–(3.413) is the basis for the block diagram on Fig. 3.29. Defining the state variables as indicated on the block diagram, the state equations can be written down by direct inspection . With $\mathbf{x} = [x \quad \dot{x} \quad p_1 \quad p_2]^T$ one finds

$$\begin{aligned}
\dot{x}_1 &= x_2,\\
\dot{x}_2 &= \frac{1}{M}(-C_f x_2 + A_c(x_3 - x_4) + f),\\
\dot{x}_3 &= \frac{\beta}{V}(-A_c x_2 - C_l(x_3 - x_4) + ku),\\
\dot{x}_4 &= \frac{\beta}{V}(A_c x_2 + C_l(x_3 - x_4) - ku),\\
y &= x_1,
\end{aligned} \tag{3.414}$$

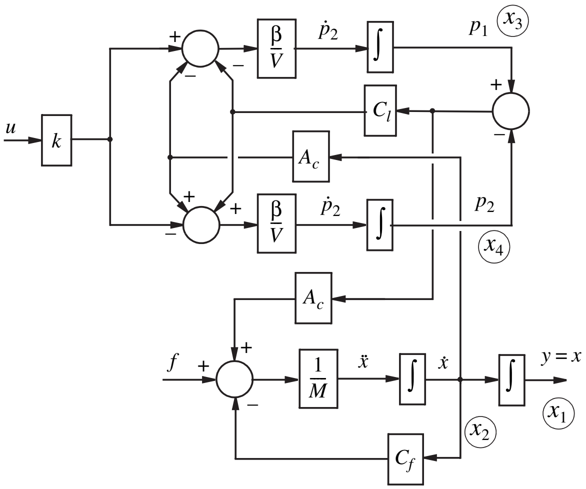


**Fig. 3.29** Block diagram of hydraulic servo cylinder

or in matrix-vector form

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{C_f}{M} & \frac{A_c}{M} & -\frac{A_c}{M} \\ 0 & -\frac{A_c\beta}{V} & -\frac{C_l\beta}{V} & \frac{C_l\beta}{V} \\ 0 & \frac{A_c\beta}{V} & \frac{C_l\beta}{V} & -\frac{C_l\beta}{V} \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ \frac{k\beta}{V} \\ -\frac{k\beta}{V} \end{bmatrix}u + \begin{bmatrix} 0 \\ \frac{1}{M} \\ 0 \\ 0 \end{bmatrix}f, \tag{3.415}$$

$$y = [1 \quad 0 \quad 0 \quad 0]\mathbf{x}.$$

Insert the following data in the equations:

$$\begin{aligned} A_c &= 150\,\mathrm{cm}^2, \\ V &= 3000\,\mathrm{cm}^3, \\ M &= 500\,\mathrm{kg} = 0.5 \cdot 1000\,\mathrm{kg}, \\ \beta &= 7000\,\mathrm{bar} \\ C_1 &= 1\,\frac{\mathrm{cm}^3}{\mathrm{bar}\cdot\mathrm{sec}}, \\ C_f &= 0.1\,\frac{10\mathrm{N}\cdot\mathrm{sec}}{\mathrm{cm}}, \\ k &= 20\,\frac{\mathrm{cm}^3}{\mathrm{sec}\cdot\mathrm{volt}}. \end{aligned}$$

Note that the units cm$^3$ 1000 kg and bar are used, instead of the SI-units: meters, kilograms and Pascals. The reason for this is that the alternative units lead to matrices which are better conditioned for numerical computations than the SI-units.

With the data above the system matrices are:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -0.2 & 300 & -300 \\ 0 & -350 & -2.333 & 2.333 \\ 0 & 350 & 2.333 & -2.333 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 46.67 \\ -46.67 \end{bmatrix}, \quad \mathbf{B}_v = \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix},$$

$$\mathbf{C} = [1 \quad 0 \quad 0 \quad 0].$$

The eigenvalues of **A** are

$$\lambda = \begin{cases} 0 \\ 0 \\ -2.433 \pm j458.3 \end{cases}.$$

If the controllability and the observability matrices are found, it will be discovered that the system is neither controllable nor observable. The controllability matrix can be computed to be

$$\mathbf{M}_c = \begin{bmatrix} 0 & 0 & 2.8 \cdot 10^4 & -1.363 \cdot 10^5 \\ 0 & 2.8 \cdot 10^4 & -1.363 \cdot 10^5 & -5.88 \cdot 10^9 \\ 46.67 & -217.8 & -9.8 \cdot 10^6 & 9.342 \cdot 10^7 \\ -46.67 & 217.8 & 9.8 \cdot 10^6 & 9.342 \cdot 10^7 \end{bmatrix},$$
$$det(\mathbf{M}_c) = 0.$$

One can also find that $rank(\mathbf{M}_c) = 3$, which means that 3 columns of $\mathbf{M}_c$ are linearly independent. A closer look at $\mathbf{M}_c$ shows that the first 3 columns are linearly independent and a suitable transformation matrix for a controllable subspace transformation (see Sect. 3.8.11),

$$\mathbf{z} = \mathbf{Q}^{-1}\mathbf{x}, \tag{3.416}$$

could be

$$\mathbf{Q} = \begin{bmatrix} 0 & 0 & 2.8 \cdot 10^4 & 0 \\ 0 & 2.8 \cdot 10^4 & -1.363 \cdot 10^5 & 0 \\ 46.67 & -217.8 & -9.8 \cdot 10^6 & 1 \\ -46.67 & 217.8 & 9.8 \cdot 10^6 & 1 \end{bmatrix}$$

which gives

$$\mathbf{Q}^{-1} = \begin{bmatrix} 7.5 & 1.666 \cdot 10^{-4} & 1.071 \cdot 10^{-2} & -1.071 \cdot 10^{-2} \\ 1.738 \cdot 10^{-4} & 3.571 \cdot 10^{-5} & 0 & 0 \\ 3.571 \cdot 10^{-5} & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix}.$$

The matrices of the transformed system will be:

$$\mathbf{A}_t = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \left[\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1 & 0 & -2.1 \cdot 10^5 & 0 \\ 0 & 1 & -4.866 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right],$$

$$\mathbf{B}_t = \mathbf{Q}^{-1}\mathbf{B} = \left[\begin{array}{c} 1 \\ 0 \\ 0 \\ \hline 0 \end{array}\right], \quad \mathbf{C}_t = \mathbf{C}\mathbf{Q} = \left[\begin{array}{c|c} 002.8 \cdot 10^4 & 0 \end{array}\right],$$

and the controllable part of the system is described by the matrices

$$\mathbf{A}_c = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & -2.1 \cdot 10^5 \\ 0 & 1 & -4.866 \end{bmatrix},$$

$$\mathbf{B}_c = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{C}_c = \begin{bmatrix} 0 & 0 & 2.8 \cdot 10^4 \end{bmatrix}.$$

The system $\{\mathbf{A}_c, \mathbf{B}_c, \mathbf{C}_c\}$ is not only controllable, it is also observable. The states of the transformed system can be found from Eq. (3.416):

$$\mathbf{z} = \left[\begin{array}{c} \mathbf{z}_c \\ \hline \mathbf{z}_{nc} \end{array}\right] = \left[\begin{array}{c} 7.5x_1 + 1.666 \cdot 10^{-4}x_2 + 1.071 \cdot 10^{-2}(x_3 - x_4) \\ 1.071 \cdot 10^{-2}x_1 + 3.571 \cdot 10^{-5}x_2 \\ 3.571 \cdot 10^{-5}x_1 \\ \hline 0.5(x_1 + x_2). \end{array}\right]$$

The controllable (and observable) system has the 3 states in the state vector $\mathbf{z}_c$ and these states are of course linear combinations of the original states. But it is also seen that the cylinder chamber pressures $x_3$ and $x_4$ ($p_1$ and $p_2$ respectively) are no longer present individually but only as the pressure *difference*, $x_3 - x_4$. It can thus be concluded that the two pressures can not be controlled (or observed) individually.

One of the 4 states of the original system is superfluous. Looking at the differential equations or block diagram on Fig. 3.29, this is not a surprising result. As a matter of fact, the two chamber pressures only occur as the difference $p_1 - p_2$ and it would be natural to choose this pressure difference as a state instead of the two individual pressures. ❐

## 3.11 Summary

This chapter has dealt with the problems of investigating the characteristics of dynamical systems. The treatment has included the following main issues:

1. Solution of the state equations.
2. Similarity transformations.
3. Stability.
4. Controllability and observability.
5. Realizability and minimality.

**Regarding 1**

Since an analytical solution to the general state equations (2.8) and (2.135) usually cannot be found, the treatment was restricted to the linear equations (3.8) and (3.88). It turns out however that an analytical solution is difficult to achieve even in this case. For the time invariant equations (3.37) and (3.96), the situation is more favourable. In this case it is easy to obtain a solution which is immediately useful. It is also possible to apply Laplace and Z-transformation to these equations and define a generalization of the transfer function and the notion of impulse response known from the classical discussion of SISO systems. It should be noted that the importance of the analytical solutions is primarily connected to further analysis issues such as stability and controllability. Determination of a specific response for a more or less complicated system, linear or nonlinear, is much more easily found by computer simulation.

**Regarding 2**

In contrast to the transfer function system formulation, the state space model is not unique. The state vector can be chosen in indefinitely many ways and certain choices provide one with specific advantages. Given one state space model, one can easily change it to another by a similarity transformation where one basically selects a new set of state variables by using the expression (3.125). It is important to note that the new model shares all the important properties of the original one. A particularly simple state model is the model with a diagonal system matrix. In such a diagonal system all states are completely decoupled from each other and this makes most analysis easier. All system models with nondefective system matrices can be diagonalized with a similarity transformation.

**Regarding 3**

One of the most important characteristics of dynamic systems is stability. It can be a quite difficult task to determine whether a general nonlinear system models is stable or not. Even for linear time varying models this question is not trivial. On the otherhand, for LTI models there is a simple stability criteria based on the position of the system eigenvalues in the complex plane. This is the case no matter which of the several possible stability definitions one wants to use. When using these criteria, it must be remembered that the linear time invariant model is always an approximation to a more comprehensive and correct description of the system. In most cases, the LTI model has emerged by a linearization of a nonlinear system around a stationary operating point. Consequently, the stability analysis based on this linear model can only (at most) tell about the stability in a close vicinity of this stationary state. Nevertheless, the linear model stability criteria are very useful for practical analysis purposes.

**Regarding 4**

When designing controllers for dynamic systems, it is very important, prior to the design process, to be able to determine which options are available for the design. The notions of controllability and observability provide tools for analysis in this area. Loosely expressed, a system is said to be controllable if it is possible to obtain entry into the system via the input variables and influence all states individually. Similarly, the system is observable if one can obtain information about all the states by viewing the system through the output. There is a variety of criteria for determination of controllability and observability. The most important ones are expressed in the controllability theorems CC2/RD2 (Eqs. (3.230) and (3.285)) and in the observability theorems OC2/OD2 (Eqs. (3.301) and (3.321)).

A system is said to be stabilizable if any non-controllable state is stable. A system is detectable if any non-observable state is stable. For systems which are not controllable or not observable, it is possible to carry out a similarity transformation which reveals the non-controllable or non-observable states. This allows one to determine if the system is stabilizable or detectable respectively.

**Regarding 5**

Realizability is concerned with the problem of formulating state space models from a transfer function matrix. It turns out that this is always possible if the individual transfer functions are all strictly proper. Moreover, the direct transfer matrix **D** will be zero in this case. Another problem connected to state space models and especially models derived from a transfer function matrix, is minimality. One says that a state space model equivalent to a given transfer function matrix is minimal if it has the smallest possible number of states. It was shown, that the system is minimal if and only if it is controllable as well as observable.

## 3.12 Notes

### *3.12.1 Linear Systems Theory*

The main mathematical background for this chapter is what is now called linear systems theory. This theory, though presented as a whole here, has emerged piecemeal from the work of many different investigators both mathematicians and physicists, over a period of nearly 400 years.

This first main contribution to linear systems theory was published by René du Perron Descartes in 1637. Descartes was a French mathematician and philosopher and one of the first modern scientists. He originated what is now called analytic or coordinate geometry. The first exposition of the elements of Cartesian coordinate geometry (*Geometry*) was published as an appendix to a natural philosophical work (*Discourse on the Method of Reasoning Well and Seeking Truth in the Sciences*). *Discourse* was an attempt to collect together Descartes' thoughts on the underlying order in the physical world and a great departure from the traditions of his time. While *Discourse* was revolutionary, *Geometry* was an evolution of the mathematics known at that time: its goal was the unification of geometry and algebra. The natural philosophical methods and principles of *Discourse* proved to be very popular and were not generally displaced until the time of Newton, some 30 years later.

While it is generally acknowledged that the invention of differential calculus is due to Newton, Pierre de Fermat was responsible for an approach to finding the tangent to a curve, effectively differentiation, in about 1630 as well as extentions to the coordinate geometry of Descartes. These results were first published after his death in around 1682 and they make Fermat the first worker to actually differentiate a function.

Isaac Newton was the originator of differential and integral calculus but his results were first published in 1687, about 10 years after their actual discovery. This was three years after the publication of similar work by Gottfried Wilhelm Leibniz in 1684. Newton is responsible for the notation for the time derivative (or fluxion at that time) which is $\dot{x}$. Leibniz's notation was intuitively much more suggestive for he wrote the same quantity as $\frac{dx}{dt}$. It is also Leibniz who introduced the notation s for sum (or integral) or, in modern terms, $\int x\,dt$. The work of Newton and Leibniz emphasized the inverse relationship of differentiation and integration and set the stage for the exposition of the basic laws of mechanics, Newton's Laws.

Further important developments in the area of differential equations and their application to physical problems are due in particular to three members of the Bernoulli family, Jakob, Johann and later Daniel Bernoulli, in the period from about 1690 to 1770. The Bernoullis learned calculus from the work of Leibniz and were in constant contact with him and thus adopted his notation. In part, because of the intuitive quality of Leibniz's notation, they made significant contributions to the theory and applications of differential equations and the calculus of variations.

The active period of the Bernoulli family overlapped that of Leonhard Euler who is responsible for the use of symbols $\pi$ and $e$ and for discovering the Euler identity $e^{i\theta} = \sin(\theta) + i\cos(\theta)$ and relating its properties. Euler is also credited with important contributions to the theory of analysis (advanced calculus), differential equations and the calculus of variations.

When Euler left Berlin for St. Petersburg in 1766, it was arranged that his position as director for the mathematical division of the Berlin Academy be filled by Joseph Louis Lagrange. Lagrange introduced the method of variation of parameters to solve differential equations, extentions of the calculus of variations in a modern form and an influential book on the mathematical unification of general mechanics. An assistant to Lagrange during one of his later appointments was Pierre Simon de Laplace who, apart from important contributions to celestial mechanics, the theory of partial differential equations and the theory of probability, is responsible for the Laplace transform.

Karl Friedrich Gauss (1777–1855) was a singularly gifted mathematician who was also interested in physics. It is Gauss who first introduced the modern requirements and standards for rigorous mathematical proofs. His production of new mathematics was so profound and fundamental that it would be difficult to relate here his actual contribution to linear systems theory. It will only be noted here that Gauss was the first to plot complex numbers as points in a two dimensional plane. That is the representation where the complex number is represented with as the abscissa on the real axis and is the ordinate on the imaginary axis. On this background it is not surprizing that Gauss is also responsible for the analytic geometry of complex numbers and functions and the rigorous foundation for complex variable theory. The possibility of drawing a picture established complex variable theory as a valid branch of mathematics whereas earlier it had been viewed with some scepticism by other mathematicians. This work was published in 1831. Another of Gauss's discoveries, the Gaussian probability distribution function, will be treated in a later chapter of this book.

Vector analysis is a degenerate form of the mathematics of quaternions, developed most completely by William Rowan Hamilton and published in 1843. Quaternions are four-fold numbers which have some of the qualities of vectors for multipicative algebraic operations. In spite of Hamilton's best efforts, they were never widely used because they are too complex for easy understanding and application. The vector analysis which is currently used is a simplification of Hamilton's work due to an American physicist, Josiah Willard Giggs, originating from about 1880.

The invention of, as well as the mechanics of the manipulation of matrices is due to an Englishman, Authur Cayley in many publications from about 1863 to 1883. The proofs of many of his results are due to his friend and co-worker, James Joseph Sylvester. In fact the word matrix is due to Sylvester and was first mentioned in a publication in 1848. Many of the important connections between matrices and determinates and the theory of invariants is also due to these workers.

Linear systems theory requires the solution of systems of coupled linear differential equations with constant coefficients. A simple method for doing this is Laplace Transform, Operator or Heavyside Calculus and is due to the efforts of Oliver Heavyside, an innovative but unconventional English electrical engineer from about 1891. This work was not initially appreciated because of its apparent lack of rigor and because of its odd notation. Now this calculus is widely recognized and used to solve physical and engineering problems involving systems of linear differential equations. It has also been given a respectible mathematical foundation using function theory, though with some difficulty, by Carson and Bromwich and others. Heavyside also made contributions to the development and practical applications of James Clerk Maxwell's electromagnetic theory. The Heavyside layer in the atmosphere is named after the same engineer.

To complete the picture of the linear systems theory now currently used, Jean Baptiste Joseph Fourier should be mentioned here because of his unifying influence on the time and frequency domain formulations of the theory. In its original form, Fourier's book from 1822 showed that any periodic function can be represented as an infinite sum of sine and cosine functions of varying amplitudes. Such a collection of sine and cosine functions is called the spectrum of the signal. It is possible to extend such formulations to aperiodic functions using Fourier transforms. These concepts are often used in modern signal analysis and in many practical devices such as frequency synthesizers and spectrum analyzers.

From the statements above, it is clear that the vital elements of linear systems theory were all more or less available from 1900. It has been further developed and simplified into its current convenient form through numerous applications of it by many workers. These applications have been primarily to electrical and electronic circuits but also to control systems, especially in the last 50 years. The control applications are extremely wide ranging and stretch from microscopic solid state electric motors to aircraft, ships and even buildings (stabilization against earthquakes).

## 3.13 Problems

### *Problem 3.1*

Given the time varying linear system:

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 \\ \frac{1}{t} \end{bmatrix}u(t),\ \mathbf{x}(t_0) = \begin{bmatrix} x_{10} \\ x_{20} \end{bmatrix}. \tag{3.417}$$

a. Find the state transition matrix $\phi(t, t_0)$ (Hint: Solve the two state equations for the homogeneous equation directly).
b. Check the result by application of equation (3.21).
c. Find the complete solution of (3.417) for $u(t) = 1 \quad \text{for} \quad t \geq t_0$.
d. Check the solution of part c. by inserting into (3.417).

### *Problem 3.2*

One has the continuous time system:

$$\dot{\mathbf{x}} = \begin{bmatrix} -3 & 2 \\ 1 & -2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 \\ b \end{bmatrix}u,\ y = [1 \quad 0]\mathbf{x}.$$

a. Find eigenvalues, resolvent matrix, transfer function, state transition matrix and impulse response.

b. Determine the parameter $b$ such that the input only influences the natural mode corresponding to the eigenvalue farthest away from the origin.
c. Choose the initial state $\mathbf{x}_0 = [1 \quad -1]^T$ and the input $u(t) = 2e^t, \quad t \geq 0$ and determine $b$ such that $\lim_{t\to\infty} y(t) = 0$.
   Find $y(t)$ for this value of $b$.

***Problem 3.3***

A linear 3. order system has the system matrix:

$$\mathbf{A} = \begin{bmatrix} -4 & \frac{1}{2} & 0 \\ 0 & -1 & 8 \\ 0 & 0 & -3 \end{bmatrix}.$$

a. Determine the resolvent matrix $\Phi(s)$ and the state transition matrix $\phi(t)$.
b. Determine the state vector $\mathbf{x}(t)$ for $t \geq 0$ when $u(t) = 0$ and $\mathbf{x}_0 = [0 \quad 1 \quad 0]^T$.

***Problem 3.4***

Given the discrete time system:

$$\mathbf{x}(k+1) = \begin{bmatrix} 0 & 1 \\ -\frac{1}{8} & \frac{3}{4} \end{bmatrix}\mathbf{x}(k) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u(k), \;\; y(k) = \begin{bmatrix} -\frac{1}{8} & -\frac{1}{4} \end{bmatrix}\mathbf{x}(k).$$

a. Set $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $u(k) = 1$ for all $k$.
   Find $y(0)$, $y(1)$ and $y(2)$.
b. Determine the eigenvalues, the natural modes and the resolvent matrix.
c. Find an analytical expression for the state transition matrix $\mathbf{F}^k$.
d. Find the transfer function $H(z) = y(z)/u(z)$.
e. Find an analytical expression for the unit pulse response $h(k)$.
f. For $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $u(k) = 1$ for all $k$ , find an analytical expression for $y(k)$ (hint: apply Eq. (3.102)).

   Compute $y(1)$ and $y(2)$ and compare with the results from question a.

***Problem 3.5***

Consider the system with the **A**-matrix from Problem 3.3:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \;\; \mathbf{x}_0 = [0 \quad 1 \quad 0]^T.$$

a. Find the eigenvalues and a set of eigenvectors for **A**.
b. Carry out the diagonal transformation of **A**.
c. Use the diagonal transformation and Eq. (3.65) to determine the state vector $\mathbf{x}$(t).

### Problem 3.6

A linear system is given by its state space model:

$$\dot{\mathbf{x}} = \begin{bmatrix} -4 & -3 \\ 1 & 0 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 3 \\ 0 \end{bmatrix}u, \ y = \begin{bmatrix} 0 & 1 \end{bmatrix}\mathbf{x}.$$

a. Find the eigenvalues and the eigenvectors of the system.
b. Determine the diagonal transformed system.
c. Determine the state transition matrix of the diagonal system.
d. Determine the state transition matrix of the original system (Hint: use (3.133)).

### Problem 3.7

A perfect ball rolls in the vertical plane on the surfaces shown on Fig. 3.30.

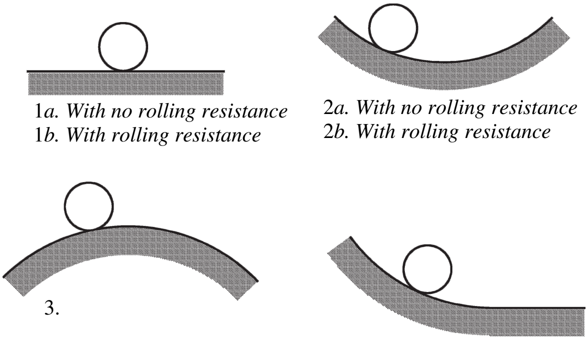


**Fig. 3.30** Rolling ball on surface in the vertical plane

1*a. With no rolling resistance*
1*b. With rolling resistance*

2*a. With no rolling resistance*
2*b. With rolling resistance*

3.

4. *With rolling resistance*

a. Use stability definitions 1 and 2 in the beginning of Sect. 3.7 to characterize the stability properties in all 6 cases.

Justify the results.

### Problem 3.8

Given the continuous time state space model,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t),$$

with the following system matrices

1. $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -3 & 2 \end{bmatrix}$
2. $\mathbf{A} = \begin{bmatrix} -2 & -3 & 5 \\ 3 & 2 & -5 \\ 2 & 1 & -3 \end{bmatrix}$
3. $\mathbf{A} = \begin{bmatrix} -2 & -3 & 5 \\ 6 & 15 & -21 \\ 5 & 14 & -19 \end{bmatrix}$
4. $\mathbf{A} = \begin{bmatrix} -2 & -3 & 5 \\ 10 & 20 & -20 \\ 9 & 19 & -18 \end{bmatrix}$

a. In all four cases, find the eigenvalues and mark them on a drawing of the complex plane.
b. Characterize the four system's stability properties.

### *Problem 3.9*

One has the discrete time state space model,

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}u(k),$$

with the following system matrices,

$$1.\ \ \mathbf{F} = \begin{bmatrix} \frac{1}{2} & \frac{1}{8} \\ -\frac{1}{2} & 1 \end{bmatrix} \qquad 2.\ \ \mathbf{F} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 1 & -1 & 1 \end{bmatrix}$$

$$3.\ \ \mathbf{F} = \begin{bmatrix} -1 & 2 & -1 \\ -1.5 & 2.5 & -1 \\ 3 & -3 & 2 \end{bmatrix} \qquad 4.\ \ \mathbf{F} = \begin{bmatrix} 1.5 & -0.5 & 0.25 \\ 1 & 0 & 0.25 \\ 0.5 & -0.5 & 0.75 \end{bmatrix}$$

a. In all four cases, find the eigenvalues and mark them on a drawing of the complex plane with the unit circle.
b. Characterize the four system's stability properties.

### *Problem 3.10*

Consider the system:

$$\mathbf{x}(k+1) = \begin{bmatrix} 0 & 1 \\ -1 & \frac{5}{2} \end{bmatrix}\mathbf{x}(k) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u(k), \quad y(k) = [-2 \quad 1]\mathbf{x}(k).$$

a. Find the system's eigenvalues and natural modes.
   Is the system asymptotically stable?
b. Find the transfer function of the system.
   Is the system BIBO-stable?

### *Problem 3.11*

Given the following continuous time LTI-system:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u, \quad y = [a \quad -1]\mathbf{x}.$$

a. Is the system internally stable?
b. Determine *a* such that the system is BIBO-stable.

***Problem 3.12***

One has the system:

$$\dot{\mathbf{x}} = \begin{bmatrix} -3 & 4 \\ -2 & 3 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\mathbf{u}, \quad \mathbf{y} = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix}\mathbf{x}.$$

a. Is the system controllable?
   Is it observable?

Now the system is changes in such a way that

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 2 \\ 1 & -2 \end{bmatrix}.$$

b. How does this change influence the controllability and the observability?

***Problem 3.13***

Consider the system:

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & -2 & -3 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}\mathbf{u}, \quad \mathbf{y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix}\mathbf{x}.$$

a. Compute the eigenvalues and the corresponding eigenvectors for $A$.
b. Use a similarity transformation to achieve the diagonal form of the system.
c. Draw block diagrams of the original as well as of the diagonal system.
d. Determine the transfer function matrix for both systems.
e. Find the left eigenvectors for the system matrix $A$ (hint: use Eq. (3.246),

   i.e., $\begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_n^T \end{bmatrix} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \ldots \quad \mathbf{v}_n]^{-1}$.

f. Use Eqs. (3.230) and (3.301) to determine controllability and observability of the system.
g. Repeat f. using the PBH-test.

***Problem 3.14***

Given the following system:

$$\dot{\mathbf{x}} = \begin{bmatrix} -2 & -3 & 5 \\ 4 & 5 & -5 \\ 3 & 4 & -3 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}u, \; y = [-2 \quad -5 \quad 5]\mathbf{x}.$$

a. Find the characteristic polynomial of the system.
b. If possible, find the controller canonical form and the observer canonical form for the system.

c. Determine the transfer function
d. Comment on the system's controllability and observability in relation to the properties of the transfer function.

### *Problem 3.15*

For a linear system a model is given in the form of the following differential equation:

$$\frac{d^3y}{dt^3}+\frac{d^2y}{dt^2}-4\frac{dy}{dt}-4y=\frac{du}{dt}-2u.$$

a. Show that the system model can be written as the state space model:

$$\dot{\mathbf{x}}=\begin{bmatrix}0&1&0\\0&0&1\\4&4&-1\end{bmatrix}\mathbf{x}+\begin{bmatrix}0\\0\\1\end{bmatrix}u,\ y=[-2\quad 1\quad 0]\mathbf{x}. \tag{3.418}$$

b. Is the system internally stable?
   Is the system minimal?
c. Find a minimal state model for the system.
d. Is the system (3.418) stabilizable?
   Is the system (3.418) detectable?

### *Problem 3.16*

A block diagram of a MIMO LTI system is seen on Fig. 3.31.

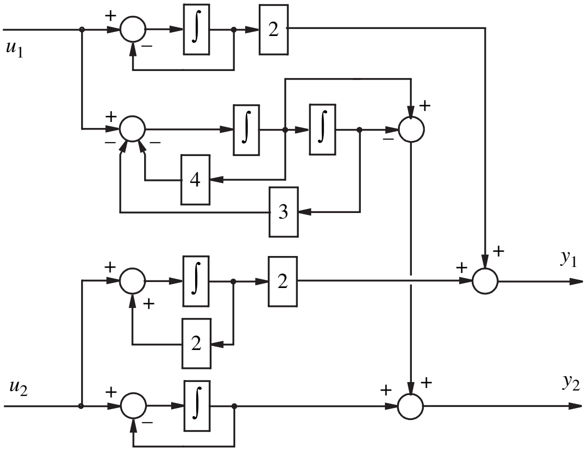


**Fig. 3.31** *Block diagram of a MIMO system*

a. Determine the transfer function matrix $\mathbf{G}(s)$ of the system:

$$\begin{bmatrix} Y_1(s) \\ Y_2(s) \end{bmatrix} = \mathbf{G}(s) \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix}.$$

b. Choose the natural state variables and derive a state space model.
c. Is the state model minimal?
d. Derive a state space model using the method in Sect. 3.10 (the Eq. (3.384)). Is this system minimal?
e. Find a minimal system by controllable subspace decomposition.
f. Is the system in part b. stabilizable and detectable?

# Chapter 4
# Linear Control System Design

**Abstract** In this chapter a review of the design of multivariable feedback controllers for linear systems will be considered. This review treats mainly deterministic control objects with deterministic disturbances. After giving an overview of the type of linear systems to be treated, this chapter will handle the basic control system design method known as pole or eigenvalue placement. First systems where measurements of all the states are available will be treated. For cases when such complete state measurements are not available the concept of deterministic observers to estimate the states which are not measured directly will be introduced. It will also be shown that it is often possible to design reduced order observers where only the unmeasured states are estimated.

## 4.1 Control System Design

Before going into the specific task of designing linear control systems it is necessary to set the ground rules for the treatment. This can be done by presenting an overall picture of the components and configuration of the system which is to be considered.

A control system is a dynamic system which is designed to operate in a prescribed manner without external interference, in spite of unavoidable effects (disturbances) which impede its proper operation. The main purpose of this book is to present methods to analyze and synthesize such systems. A second purpose is to present methods to model disturbances and design control systems for minimum disturbance sensitivity. This requires a tabulation of the main elements of such systems and a presentation of their general configuration.

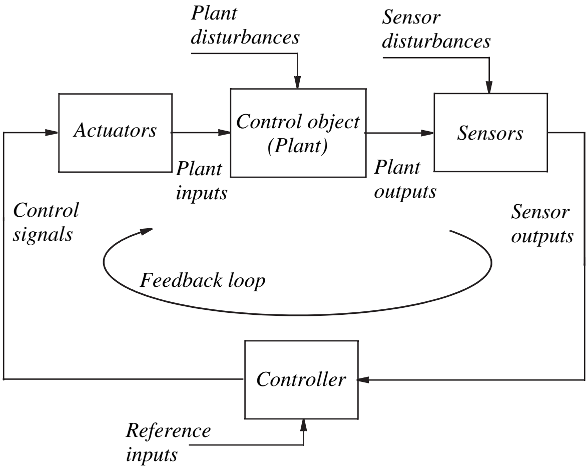
The main components of a control system are

1. The plant or control object.
2. The actuators or drivers for the plant.
3. The sensors which measure the current operating point of the plant.
4. The controller which drives the plant in accordance with the overall control objective given the sensor measurements.

E. Hendricks et al., *Linear Systems Control*, DOI: 10.1007/978-3-540-78486-9_4,

© Springer-Verlag Berlin Heidelberg 2008

**Fig. 4.1** Block diagram of a typical control system. The plant or control object is shown in the upper part of the figure while the controller and feedback loop is shown at the bottom



A block diagram of a typical control system is presented on Fig. 4.1. Note, that the actuators and sensors are usually considered to be external to the control object itself. However it is often necessary that the dynamics of these components are taken into account in the design of the overall feedback control system.

The plants or control objects which may be controlled in this way can be of many different types: mechanical, electrical, fluid dynamic, economic, biological, etc. or combinations of such plants. The only limitation to the nature of the plant (as far as this book is concerned) is that it be described in terms of a coupled set of differential or difference equations. Actuators are devices which are coupled to the control inputs of the plant to supply the energy necessary to effectuate the control commands of the controller. Sensors are devices for measuring the outputs and/or states of the plant. This general description can be used on many types of systems. The controller is in general a dynamic system which on the basis of the measurements provided by the sensors gives an input to the actuators which drive the control object in such a way as to accomplish the desired control objective.

The main feature of control system theory is feedback. This means use of the sensor measurements to derive a signal or signals which are used to drive the actuators of the control object to accomplish a given control task. Such a feedback (loop) is shown on Fig. 4.1 and it is in general external to the control object itself. This mechanism is used to increase the speed or bandwidth of the control object, to increase control accuracy at one or many operating points or to achieve some other desirable control effect.

Another important feature of feedback control systems is an external input which is inserted into the controller in order to provide information as to what the desired control point or trajectory is. This input is shown on the bottom of Fig. 4.1 and is commonly called the reference or command input. Often this input takes the form of a desired value for one or more of the outputs or states of the control object. It may be either constant or variable.

One of the main reasons to use control systems is to suppress disturbances of the control object or minimize the effect of noisy or inaccurate measurements of the state of the plant. Disturbances which are inherent in the plant are often called state disturbances, biases or noise. These are indicated on Fig. 4.1. Such disturbances are often generated physically in the plant but may also include changes in the characteristics of the control object itself. These modelling error disturbances are just as important as those which come from other sources because they make it difficult to know the nature of the plant itself. Thus they destroy the basis for proper modelling and accurate design. Such disturbances are called plant parameter variations or modelling errors. When measurements are made on the states of a system, it is often the case that disturbances are introduced in the measurement mechanism or sensor itself. This is shown on Fig. 4.1 as the measurement disturbance or noise source. In this chapter the presence of the state disturbance will be considered indirectly and it will be assumed to be of an unmodelled 'deterministic' type. Disturbances cannot in general be manipulated by the control system designer so that the designer can at best suppress or minimize the effects of disturbances on a given control system. This nearly always involves a design compromise between carrying out the desired control task while at the same time suppressing system disturbances.

Many examples of control systems are readily visible in the world at the present time. These include control systems for chemical process plants, hard disk drives, automotive engines and vehicles, aircraft autopilots, space vehicle attitude and navigation, investment and economic management systems. In fact the availability of inexpensive semiconductor chips as central processing units (CPUs) and analog signal processors ensure that control systems will be built into a very large percentage of the more complex industrial products and services which will be offered in the future.

Currently most of the control systems which have been produced are for physical or chemical systems. As an easily understood physical example one can take an aircraft autopilot. In such systems the control objective is to control the speed, altitude and attitude of the aircraft as accurately as possible. A second main objective is to provide for automatic path following as well, but only the first objective will be considered here. The plant is of course the aircraft itself: it is a dynamic system because the control surfaces and engine of the aircraft, once activated, can only react with time constants and/or time delays. Increasing for example the throttle control of a jet engine will cause an increase in thrust only on time scales of say 10–30 s.

The actuator for the engine is the fuel injection nozzle and fuel control valve. For attitude control the actuator is the control surface positioning mechanism. This mechanism is often an electric or a hydraulic motor. Sensors on aircraft include gyroscopes to sense the angle of the aircraft with respect to the horizon and rate gyros to sense rate of rotation. Air speed sensors are often pitot tubes. Currently the sensor signals are fed into a controller which is a digital computer. In former times the controller was a electrical/mechanical analog computer. One important disturbance which prevents an aircraft from following a desired path is the air movement around it. This can take the form of side wind, updrafts and downdrafts. When the aircraft deviates from its desired flight path because of external air movements, the job of the controller is to maintain a certain attitude and speed as closely as possible. Additionally it can be stated that aircraft autopilots are currently so advanced that the pilots' main task in many cases (even during landing and takeoff) is reduced to simply monitoring the aircraft's attitude and path controllers.

The consideration of control systems for other types of systems than physical or chemical has also reached an advanced level in some areas. One of the most interesting of these areas is control and/or planning applied to economic systems. Consider the case of the inventory control. A sales organization wants to control the size of its inventory so that it can always fill its orders. Its inventory on any given day is equal to the size of its stock plus the orders which are given to the organization's supplier the previous day, minus its daily sales. From experience the manager knows that his sales vary around a certain value every day. The question is: what sort of control must be applied in order to make certain that there is always something to sell? The sensor here is the counting up of the stock every day. The actuator is the giving of the order. Dynamically the system is dominated by its inherent time delays. Possible state disturbances in the system are the sales level which changes and the variability of production quantity and quality. Measurement disturbance could be possible mistakes in counting up the stock or difficulties in keeping track of large numbers of different commodities. This is a typical management problem but other interesting problems are for example optimal investment strategies and commodity pricing.

### 4.1.1 Controller Operating Modes

Control systems can obviously be used for many different purposes but in general they operate in two basic control modes:

1. As regulators, intended to operate around a single set or operating point in state space.
2. As trackers, intended to follow a certain trajectory in state space.

For linear systems, which do not change their dynamic characteristics with their operating point in state space, there is very little difference between these two modes of operation. However, as most control objects are nonlinear, there may be significant control problems involved in operating in these two different modes.

When operating in the regulator mode a control system has the goal of keeping the control object at a certain location (or set point) in state space. This location is specified by defining a constant value of a single state or output variable, multiple states or outputs or some linear combination of them. This is probably the most common operating mode for a control system.

When a control system operates in the tracking mode, it is the intention that the state vector describe a certain path (trajectory) in state space. Usually this is accomplished by changing the reference input to the control system according to some predetermined pattern in order to cause it to move as desired. This is the second most common operating mode for a control system.

### *Example 4.1.* Regulator for the Idle Speed of a Spark Ignition Engine

Modern gasoline driven vehicles spend on the order of 30% of their operating life at idle when operating in cities. It is also true that drivers in large measure judge the quality of their vehicles by the evenness of their engines in idle speed. This means that it is important for vehicle manufacturers to ensure that their engines idle at a constant speed (usually 600–1200 rpm) and smoothly, despite load disturbances. These disturbances are mainly due to secondary engine loads (lighting, generators, pumps for steering assistance, electric windows, air conditioning, etc.).

Mostly because of emission restrictions, modern vehicles are nearly all provided with electronic engine actuators, sensors as well as three-way catalyst exhaust systems. In fact the main reason for the advanced engine controls is emissions legislation. Recently it has become clear for reasons of cost that most vehicles will be provided with drive-by-wire throttle bodies. In such systems the accelerator pedal is connected to a potentiometer and the throttle plate operated with an electric motor position control system. This makes it possible to eliminate several other actuators for other engine functions than idle speed control and thus reduce the overall cost of the engine control system. Thus idle speed control is accomplished by controlling the position of the throttle plate and hence the air flow to the engine. A secondary engine speed control is via the spark timing. The throttle plate is a butterfly valve placed just before the intake manifold or plenum of the engine. This plenum is then connected by runners to the engine ports on the engine head itself. The actuator for the throttle plate is often a DC electric motor with or without some gear reduction. The sensor for the engine speed is a magnetic pick up which is coupled to the engine crank shaft.

Figure 4.2 is a block diagram for an idle speed control system. Here it is shown that the idle speed control system measures the crank shaft speed, $n$, and on this basis adjusts the port air mass flow, $m_{ap}$, via the throttle plate angle $\alpha$ and the engine spark timing. In general for physical reasons the throttle angle control is used for slow speed control while spark timing is used only for fast, momentary speed control. Also shown on the block diagram only for the sake of completeness is the air/fuel ratio control which is assumed in this example to be ideal (the ratio of the port air mass flow to the fuel mass flow is fixed at $m_{ap}/m_{fi} = 14.7$. This ratio is the stochiometric air/fuel ratio (AFR) necessary for low emission operation of modern catalyst equipped engines. The idle speed control subsystem has to work in conjunction with the engine air/fuel ratio control loop. ❒

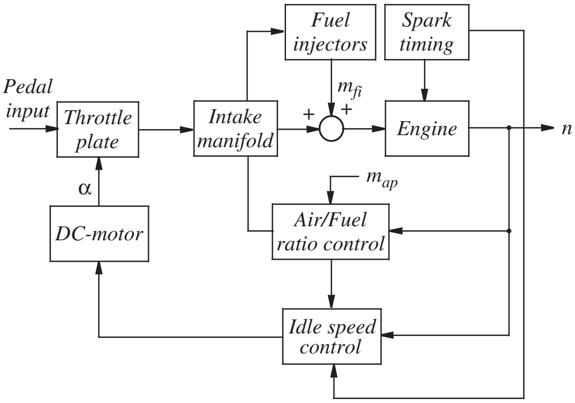


**Fig. 4.2** Block diagram for an idle speed control system for a port injected gasoline engine

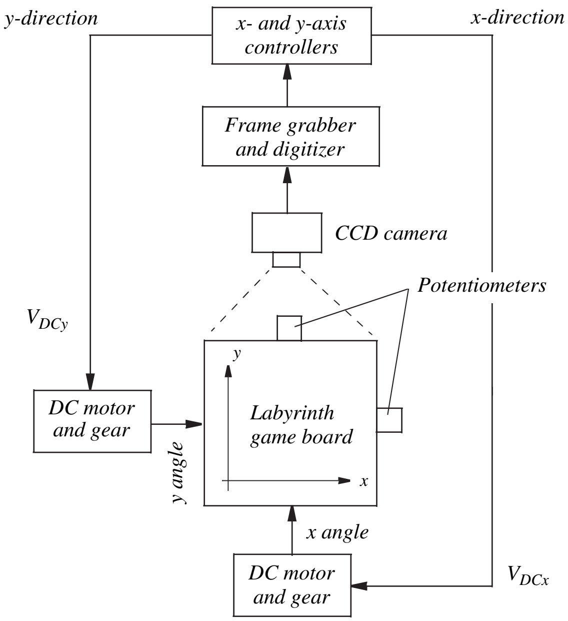
### *Example 4.2.* Optical Tracking of a Labyrinth Game

An interesting tracking control study with numerous possible industrial applications has been carried out at Department of Automation at the Technical University of Denmark on a labyrinth game using a television sensor. The control object for this study is an x–y game labyrinth equipped with actuators (small DC-motors) for the $x$ and $y$ directions. The object of the game is to use the $x$ and $y$ controls to tip the game board in either direction to move a small ball around a predetermined path marked on the game board. To sense the position of the ball on the game board, a small CCD television camera is used together with a frame grabber. To make this exercise more difficult, holes are cut in the game board close to the target path on the game board.

Figure 4.3 is a block diagram of the experimental set up. The television camera acts as a two dimensional optical sensor for the position of the ball on the labyrinth game board. To make its pictures understandable for the controller microprocessor a frame grabber is used to digitize the picture. The camera picture also provides information about the target track along which the ball must be moved. This is the path which must be tracked by the control system. The tracking controller generates $x$ and $y$ angular inputs to the tipping actuators built into the labyrinth game box. These are control voltages $V_{DCx}$ and $V_{DCy}$ which drive the DC-motors. The main state disturbances for the ball are the irregularities in the surface of the labyrinth game board and the ball itself. Measurement noise enters the system via the digitizing of the television camera picture and consists of pixel noise and scattered light from the surroundings.

To start a game the ball is placed in its starting position. The controller and its associated software then locates the outline of the game board and the target track on it. It also locates the optical center of the ball. To move the ball along the target trajectory, the reference for the ball center is moved continuously along it slowly enough for the ball to follow. In other words the ball is set to track a certain changing position or state trajectory on the game board. ❐

**Fig. 4.3** Path tracking controller of a labyrinth game



## 4.2 Full State Feedback for Linear Systems

In this section the control object under consideration will be described by the LTI state equation known from Chap. 2. For a continuous time system a vector state equation can be used:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{B}_v\mathbf{v}(t). \tag{4.1}$$

The disturbance $\mathbf{v}(t)$ is assumed to be of a deterministic nature. Measurements are made on this system which can be either the states themselves or linear combinations of them:

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{w}(t). \tag{4.2}$$

The vector $\mathbf{w}(t)$ represents the measurement noise, but it will not be considered explicitly in this chapter.

In order to establish *linear feedback* around the system above, a linear feedback law can be implemented which can be written:

$$\mathbf{u}(t) = -\mathbf{K}\mathbf{x}(t) + \mathbf{r}(t), \tag{4.3}$$

where $\mathbf{K}$ is a *feedback matrix* (or a *gain matrix*) of dimension $n \times m$. $\mathbf{r}(t)$ is the reference input vector to the system. It has the same dimension as the input vector $\mathbf{u}(t)$.

As all of the states are measured, the resulting feedback system is called a *full state feedback* system. Usually it is intended that the output of the control object should follow the reference input in some sense. A short name for the state feedback controller which is sometimes used is *state controller*. A block diagram of the *closed loop system* is seen on Fig. 4.4.

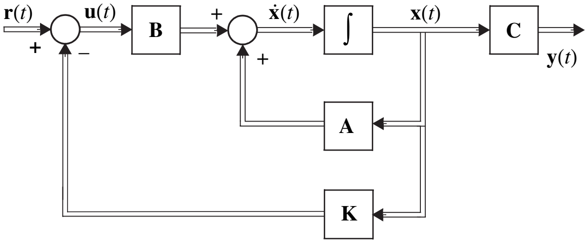


**Fig. 4.4** Closed loop system with full state feedback

Inserting Eqs. (4.3) into (4.1) and ignoring the state disturbance for the moment, an equation for the closed loop system can be derived:

$$\dot{\mathbf{x}}(t) = (\mathbf{A} - \mathbf{BK})\mathbf{x}(t) + \mathbf{Br}(t). \tag{4.4}$$

Equation (4.4) is the state equation of the closed loop system with the linear state feedback. This system is asymptotically stable if and only if the system matrix,

$$\mathbf{A_K} = \mathbf{A} - \mathbf{BK}, \tag{4.5}$$

has all its eigenvalues in the left half plane. As will be seen later, it is possible–under mild conditions–to place the eigenvalues of $\mathbf{A_K}$ *arbitrarily* when full state feedback is used. The eigenvalues are determined as the solutions to the equation:

$$det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{BK}) = 0. \tag{4.6}$$

This method of design is often called pole placement and is a common method for initial system design, Jacobs (1993), Friedland (1987). Before going into a more complete development of the eigenvalue placement method, it is useful to look at an introductory example based on the DC-motor of Example 2.3.

***Example 4.3*. Pole Placement Regulator of a DC Motor**

Consider an example where a DC-motor is used for angular or linear position control. A practical example of the former control type is the throttle position control subsystem which forms part of the idle speed control system in Example 4.1. An example of a controller for linear movement is that for the $x$ and $y$ coordinates of an x–y plotter. For the sake of simplicity an angular position control is considered here.

The states are the angular position and velocity respectively,

$$x_1 = \theta,$$

$$x_2 = \dot{\theta}.$$

Assume that the armature inductance is negligible so that an adequate state equation is Eq. (2.28),

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & -\dfrac{b_b R + K_a k_e}{JR} \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ \dfrac{K_a}{JR} \end{bmatrix} u, \qquad (4.7)$$
$$y = [1 \quad 0]\mathbf{x}.$$

To obtain some insight into the meaning of the matrix elements, the transfer function for the motor can be found. First the individual equations are written down,

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -\frac{b_b R + K_a k_e}{JR} x_2 + \frac{K_a}{JR} u, \qquad (4.8) \\ y &= x_1. \end{aligned}$$

Then these equations are Laplace transformed:

$$\begin{aligned} s x_1(s) &= x_2(s), \\ s x_2(s) &= -\frac{b_b R + K_a k_e}{JR} x_2(s) + \frac{K_a}{JR} u(s), \qquad (4.9) \\ y(s) &= x_1(s). \end{aligned}$$

Eliminating the state variables gives directly the relation between $u$ and $y$,

$$\frac{y(s)}{u(s)} = \frac{\dfrac{K_a}{b_b R + K_a k_e}}{s\left(\dfrac{JR}{b_b R + K_a k_e} s + 1\right)} = \frac{K_{vm}}{s(\tau_m s + 1)}, \qquad (4.10)$$

where $\tau_m$ is the motor time constant and $K_{vm}$ the static motor gain. In many cases it is true that,

$$b_b R \ll K_a k_e,$$

and therefore,

$$\tau_m \cong \frac{JR}{K_a k_e} \text{ and } K_{vm} \cong \frac{1}{k_e}.$$

Substituting these quantities into the state Equation in (4.7) one finds

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & -\dfrac{1}{\tau_m} \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ \dfrac{K_{vm}}{\tau_m} \end{bmatrix} u. \tag{4.11}$$

It is desireable to place the eigenvalues of the closed loop system in predetermined positions. For instance one can choose,

$$\lambda_{cl} = -\alpha \pm j\beta,$$

where $\alpha$ and $\beta$ are positive numbers. This means that it is desired that the closed loop characteristic polynomial to be

$$P_{ch}(\lambda) = (\lambda + \alpha + j\beta)(\lambda + \alpha + -j\beta) = \lambda^2 + 2\alpha\lambda + \alpha^2 + \beta^2. \tag{4.12}$$

This is also the characteristic polynomial of the closed loop system matrix $\mathbf{A_K}$,

$$P_{ch}(\lambda) = det(\lambda\mathbf{I} - \mathbf{A_K}) = det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{BK}), \tag{4.13}$$

with

$$\mathbf{K} = [\,k_1 \quad k_2\,]$$

and with the matrices from Eq. (4.11), one has

$$\begin{aligned} P_{ch}(\lambda) &= det\left( \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 0 & -\dfrac{1}{\tau_m} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \dfrac{k_1 K_{vm}}{\tau_m} & \dfrac{k_2 K_{vm}}{\tau_m} \end{bmatrix} \right) \\ &= \lambda^2 + \frac{1 + k_2 K_{vm}}{\tau_m}\lambda + \frac{k_1 K_{vm}}{\tau_m}. \end{aligned} \tag{4.14}$$

Equating (4.12) and (4.14) leads directly to expressions for the determination of $k_1$ and $k_2$,

$$\begin{aligned} \frac{1 + k_2 K_{vm}}{\tau_m} &= 2\alpha, \\ \frac{k_1 K_{vm}}{\tau_m} &= \alpha^2 + \beta^2, \end{aligned} \tag{4.15}$$

which gives

$$
\begin{aligned}
k_1 &= \frac{\tau_m}{K_{vm}}(\alpha^2 + \beta^2), \\
k_2 &= \frac{\tau_m}{K_{vm}}\left(2\alpha - \frac{1}{\tau_m}\right).
\end{aligned}
\tag{4.16}
$$

From classical system theory it is known that the reciprocal of the time constant is the cutoff (or break frequency) of the first order system,

$$\frac{1}{\tau_m} = \omega_b.$$

For the characteristic polynomial (4.12) of the second order system it is also known that

$$\omega_n = \sqrt{\alpha^2 + \beta^2} \text{ and } \zeta\omega_n = \alpha.$$

If the natural frequency $\omega_n$ is to be, say 5 times the cutoff frequency and the damping ratio to be $\zeta = \sqrt{2}/2$, then the parameters must be

$$\alpha = \beta = \frac{5\sqrt{2}}{2\tau_m}$$

which gives the gains

$$
\begin{aligned}
k_1 &= \frac{25}{K_{vm}\tau_m}, \\
k_2 &= \frac{5\sqrt{2} - 1}{K_{vm}}.
\end{aligned}
\tag{4.17}
$$

A block diagram of the closed loop system is shown on Fig. 4.5. Now it is obvious that what has been achieved here is nothing but a reinvention of the classical position servomechanism with internal tachometer feedback. But, as will become clear presently, this achievement has much wider perspectives.

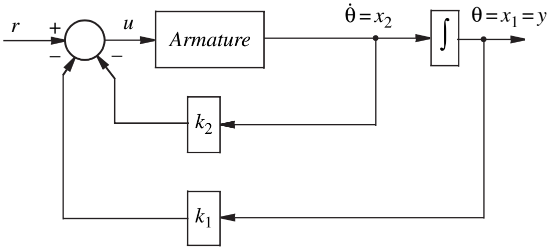


**Fig. 4.5** DC-motor with state feedback

❐

### *Example 4.4.* Pole Placement for a Mass, Spring, Damper System

A large number of dynamic systems can be described as second order systems with more or less damping. These include suspensions for transport vehicles, various hydraulic components, simple seismometers, simple rate gyros, loudspeakers and many other technical devices. It is also often true that the effective response of complex higher order systems is dominated by a single complex pole pair, again, more or less damped. Thus it is relevant to look at such systems in order to evaluate the effect of full state feedback. A simple system which can be used as an example is a forced mass-spring-damper system.

The differential equation describing such a system can be written,

$$m\ddot{x} = -kx - b\dot{x} + F, \tag{4.18}$$

where $m$ is the mass, $x$ is its position, $k$ is the spring constant, $b$ is the damping coefficient and $F$ is an external driving force. A sketch of the system is shown on Fig. 4.6.

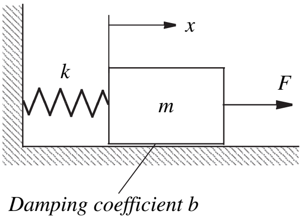


**Fig. 4.6** A driven mass-spring-damper system

Usually for mechanical systems the position and the velocity are selected as state variables,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix},$$

and the state equations can readily be formulated as

$$\dot{x}_1 = x_2,$$

$$\dot{x}_2 = -\frac{k}{m}x_1 - \frac{b}{m}x_2 + \frac{1}{m}F.$$

If one lets the driving acceleration be the input to the system (i.e., $u = F/m$) the following dynamic and input matrices result:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Deriving the transfer function the same way as in Example 4.3 the natural frequency and the damping ratio for the system are found to be

$$\omega_n = \sqrt{\frac{k}{m}},$$

$$\zeta = \frac{b}{2\sqrt{mk}}.$$

If it is desired to design a state controller with measurement of and feedback from the two states, one can proceed as in Example 4.3. Choosing the closed loop natural frequency $\omega_{ncl}$ and damping ratio $\zeta_{cl}$, the closed loop characteristic polynomial becomes

$$P_{ch}(\lambda) = \lambda^2 + 2\zeta_{cl}\,\omega_{ncl}\lambda + \omega_{ncl}^2.$$

The closed loop characteristic polynomial is

$$P_{ch}(\lambda) = det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{BK}) = \lambda^2 + (k_2 + 2\zeta\omega_n)\lambda + k_1 + \omega_n^2,$$

where the feedback gain matrix which has been used is

$$\mathbf{K} = [k_1 \quad k_2].$$

Comparison the two expressions for $P_{ch}(\lambda)$ gives the gains,

$$\begin{aligned} k_1 &= \omega_{ncl}^2 - \omega_n^2, \\ k_2 &= 2(\zeta_{cl}\omega_{ncl} - \zeta\omega_n). \end{aligned} \tag{4.19}$$

Notice that the treatment above allows the possibility that the original damping ratio $\zeta$ of the system can be either positive or negative. A negative damping may be hard to imagine in this case but second order systems with negative damping do exist. Moreover, the natural frequencies may have any value. If it is desired that the closed loop system to be slower than the original system ($\omega_{ncl}^2 < \omega_n^2$) then the gain $k_1$ becomes negative. So it is always possible to stabilize the system no matter how badly damped or how fast it is. The only requirement is that an adequate driving force can be generated.

Now a numerical example will be given which is representative of a heavy industrial system. Assume the following data

$$m = 1500\,\text{kg}, \quad k = 2000\,\frac{\text{N}}{\text{m}}, \quad b = 1\,\frac{\text{N}\cdot\text{sec}}{\text{m}}.$$

This leads to the matrices,

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1.333 & -0.1167 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The system's eigenvalues, natural frequency and damping ratio are

$$\lambda = -0.0583 \pm j1.153, \quad \omega_n = 1.155 \frac{\text{rad}}{\text{sec}}, \quad \zeta = 0.0505.$$

This is a system with very low damping and an objective for the control system design will be to increase the damping. If the values

$$\omega_{ncl} = 2 \frac{\text{rad}}{\text{sec}}, \quad \zeta_{cl} = 0.6,$$

are selected, the feedback gains may be calculated from Eq. (4.19),

$$k_1 = 2.666,$$

$$k_2 = 2.283.$$

Figure 4.7 shows the closed loop system. The result of a simulation of the system with and without feedback is shown on Fig. 4.8. The input signal to the system with feedback is a unit step and for the uncontrolled system it is a step of height 1/3.

The force time function for the feedback case is seen on Fig. 4.9. The responses show that the dynamic behaviour of the system has been improved considerably by the feedback. In spite of the fast response of the controlled system, the force necessary for the movement is not excessive.

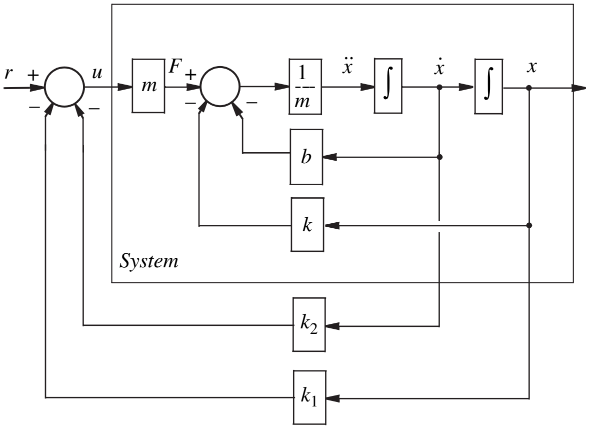


**Fig. 4.7** Mass-spring-damper system with state feedback

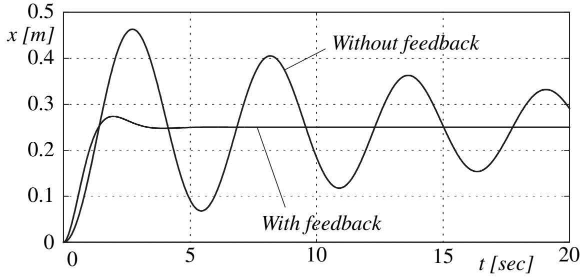


**Fig. 4.8** Response of the system on Fig. 4.7 with and without state feedback

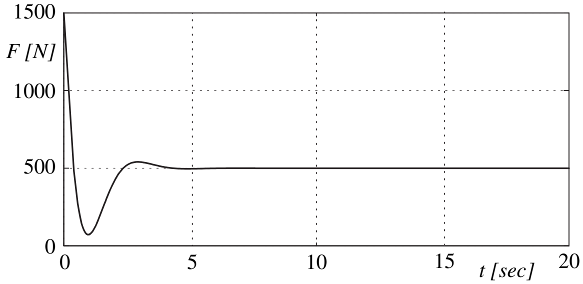


**Fig. 4.9** Force for the system with state feedback

❐

The last example shows that stable or unstable systems can be stabilized using full state feedback. In fact the eigenvalues have been placed arbitrarily in the complex plane in order to satisfy a predetermined performance requirement. This has been done for a low order system but it can also be done for a higher order system though this may require more sophisticated design procedures that those used here. Notice however that the simple method presented above gives no clue as to where the eigenvalues of the closed loop system might be placed for optimal results. This requires a more complex treatment which will be presented in Chap. 5.

Though the remarks and examples above indicate that full state feedback can be used to obtain any given eigenvalue placement, some cautionary remarks are in order with respect to what can be accomplished on general systems. One of the reasons for caution is that the discussion above has not considered the question of system zeros. As will be detailed later, system zeros cannot be moved using this type of feedback (for SISO systems) and this means that they may set a fundamental limit to system performance. Zeros have an important influence on the transient response of a system. In particular a zero can cause a large overshoot even in a stable system. Nonminimum phase zeros (those in the right half plane) are particularly worrying in this respect. It must be remembered that right half plane zeros cannot be cancelled with right half plane poles.

## 4.3 State Feedback for SISO Systems

### *4.3.1 Controller Design Based on the Controller Canonical Form*

The design procedure in Sect. 4.2 can be systematized in several ways. For SISO systems the most obvious general approach is based on the controller canonical form of the state equation. In Sect 3.9 it was seen that a controllable system can always be transformed into the controller canonical form using the similarity transformation,

$$\mathbf{z} = \mathbf{P}^{-1}\mathbf{x}, \tag{4.20}$$

where **P** can be found from Eq. (3.354). If the state equation is given as a controller canonical form, as will be the case if the state model is derived directly from a transfer function, then one already knows beforehand, that the system is controllable.

The controller canonical form is:

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{A}_{cc}\mathbf{z} + \mathbf{B}_{cc}u, \\ y &= \mathbf{C}_{cc}\mathbf{z}, \end{aligned} \tag{4.21}$$

where

$$\mathbf{A}_{cc} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ : & : & : & \ldots & : & : \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \ldots & -a_{n-2} & -a_{n-1} \end{bmatrix}, \quad \mathbf{B}_{cc} = \begin{bmatrix} 0 \\ 0 \\ : \\ 0 \\ 1 \end{bmatrix}, \tag{4.22}$$

$$\mathbf{C}_{cc} = [\, b_0 \quad b_1 \quad b_2 \ldots \quad b_{n-1} \,].$$

With the full state feedback,

$$u = -\mathbf{K}_{cc}\mathbf{z} + r, \tag{4.23}$$

and the feedback gain matrix,

$$\mathbf{K}_{cc} = [\, k_1' \quad k_2' \ldots k_n' \,], \tag{4.24}$$

the following closed loop system is obtained,

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{A}_{\mathbf{K}_{cc}}\mathbf{z} + \mathbf{B}_{cc}r, \\ y &= \mathbf{C}_{cc}\mathbf{z}, \end{aligned} \tag{4.25}$$

where the system matrix becomes

$$\mathbf{A}_{\mathbf{K}_{cc}} = \mathbf{A}_{cc} - \mathbf{B}_{cc}\mathbf{K}_{cc}$$
$$= \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ : & : & : & \dots & : & : \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -a_0 - k_1' & -a_1 - k_2' & -a_2 - k_3' & \dots & -a_{n-2} - k_{n-1}' & -a_{n-1} - k_n' \end{bmatrix}. \tag{4.26}$$

Since the input matrix is still the same as before, it is seen immediately that the closed loop system is *also* in the controller canonical form. This means that the elements in the bottom row of $\mathbf{A}_{\mathbf{K}_{cc}}$ are the coefficients of the characteristic polynomial of the closed loop system.

Now if it is desired that the eigenvalues of the closed loop system are to be placed in specific positions in the complex plane,

$$\lambda_{cl} = \lambda_{cl1}, \lambda_{cl2}, \dots, \lambda_{cln}, \tag{4.27}$$

the closed loop characteristic polynomial can be written

$$P_{ch,\, \mathbf{A}_{\mathbf{K}_{cc}}} = \prod_{i=1}^{n}(\lambda - \lambda_{cli}) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \dots + \alpha_1\lambda + \alpha_0. \tag{4.28}$$

Comparing (4.26) and (4.28) allows one to set up a very simple set of equations for determination of the feedback gains

$$\begin{aligned} \alpha_0 &= a_0 + k_1', \\ \alpha_1 &= a_1 + k_2', \\ &\vdots \\ \alpha_{n-1} &= a_{n-1} + k_n', \end{aligned} \tag{4.29}$$

which implies that

$$\begin{aligned} k_1' &= \alpha_0 - a_0, \\ k_2' &= \alpha_1 - a_1, \\ &\vdots \\ k_n' &= \alpha_{n-1} - a_{n-1}. \end{aligned} \tag{4.30}$$

Substituting (4.20) into (4.23) the gain matrix for the original system can be found:

$$
\begin{aligned}
u &= -\mathbf{K}_{cc}\mathbf{P}^{-1}\mathbf{x} + r = -\mathbf{K}\mathbf{x} + r \\
&\Rightarrow \mathbf{K} = \mathbf{K}_{cc}\mathbf{P}^{-1}.
\end{aligned}
\tag{4.31}
$$

If the system (4.21) is the original system and not the result of a similarity transformation this last step is of course irrelevant.

## 4.3.2 Ackermann's Formula

It is possible to avoid the similarity transformation prior to application of the design procedure above. To show how this can be achieved the treatment will be specialized for a third order system for the sake of simplicity. The result can immediately be extended to systems of any order.

The controllability matrix for the third order system,

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u, \tag{4.32}$$

is

$$\mathbf{M}_c = \begin{bmatrix} \mathbf{B} & \mathbf{A}\mathbf{B} & \mathbf{A}^2\mathbf{B} \end{bmatrix}. \tag{4.33}$$

If the controller is transformed into canonical form (using $\mathbf{z} = \mathbf{P}^{-1}\mathbf{x}$) one finds

$$\dot{\mathbf{z}} = \mathbf{A}_{cc}\mathbf{z} + \mathbf{B}_{cc}u, \tag{4.34}$$

with the controllability matrix (see Eq. (3.256)),

$$\mathbf{M}_{c,cc} = \begin{bmatrix} \mathbf{B}_{cc} & \mathbf{A}_{cc}\mathbf{B}_{cc} & \mathbf{A}_{cc}^2\mathbf{B}_{cc} \end{bmatrix} = \mathbf{P}^{-1}\mathbf{M}_c, \tag{4.35}$$

which shows that

$$\mathbf{P}^{-1} = \mathbf{M}_{c,cc}\mathbf{M}_c^{-1}. \tag{4.36}$$

The matrices $\mathbf{A}$ and $\mathbf{A}_{cc}$ have the same characteristic polynomial,

$$P_{ch,\mathbf{A}}(\lambda) = \lambda^3 + a_2\lambda^2 + a_1\lambda + a_0 = P_{ch,\mathbf{A}_{cc}}(\lambda), \tag{4.37}$$

which means that

$$\mathbf{A}_{cc} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix}, \quad \mathbf{B}_{cc} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \text{ and } \mathbf{M}_{c,\,cc} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & -a_2 \\ 1 & -a_2 & a_2^2 - a_1 \end{bmatrix}. \qquad (4.38)$$

As before the coefficients of the closed loop characteristic polynomial are selected:

$$P_{ch,\,\mathbf{A_K}}(\lambda) = \lambda^3 + \alpha_2\lambda^2 + \alpha_1\lambda + \alpha_0. \qquad (4.39)$$

In this polynomial substitute $\mathbf{A}_{cc}$ for $\lambda$ and obtain the matrix polynomial:

$$P_{ch,\mathbf{A_K}}(\mathbf{A}_{cc}) = \mathbf{A}_{cc}^3 + \alpha_2\mathbf{A}_{cc}^2 + \alpha_1\mathbf{A}_{cc} + \alpha_0\mathbf{I}. \qquad (4.40)$$

From the Cayley-Hamilton theorem it is known that

$$\mathbf{A}_{cc}^3 + a_2\mathbf{A}_{cc}^2 + a_1\mathbf{A}_{cc} + a_0\mathbf{I} = \mathbf{0}. \qquad (4.41)$$

Subtracting (4.41) from (4.40) yields

$$P_{ch,\mathbf{A_K}}(\mathbf{A}_{cc}) = (\alpha_2 - a_2)\mathbf{A}_{cc}^2 + (\alpha_1 - a_1)\mathbf{A}_{cc} + (\alpha_0 - a_0)\mathbf{I}. \qquad (4.42)$$

Since

$$\mathbf{A}_{cc}^2 = \begin{bmatrix} 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \\ a_0a_2 & -a_0 + a_1a_2 & -a_1 + a_2^2 \end{bmatrix} \qquad (4.43)$$

Equation (4.42) becomes

$$P_{ch,\mathbf{A_K}}(\mathbf{A}_{cc}) = \begin{bmatrix} \alpha_0 - a_0 & \alpha_1 - a_1 & \alpha_2 - a_2 \\ X & X & X \\ X & X & X \end{bmatrix}, \qquad (4.44)$$

where the elements marked $X$ are functions of the coefficients $a_i$ and $\alpha_i$.

From Eq. (4.30) one has that

$$\mathbf{K}_{cc} = [\,\alpha_0 - a_0 \quad \alpha_1 - a_1 \quad \alpha_2 - a_2\,] \qquad (4.45)$$

which means that one can write

$$\begin{aligned} \mathbf{K}_{cc} &= [\,1 \quad 0 \quad 0\,]P_{ch,\mathbf{A_K}}(\mathbf{A}_{cc}) = [\,1 \quad 0 \quad 0\,]P_{ch,\mathbf{A_K}}(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) \\ &= [\,1 \quad 0 \quad 0\,]\mathbf{P}^{-1}P_{ch,\mathbf{A_K}}(\mathbf{A})\mathbf{P} \end{aligned}. \qquad (4.46)$$

Equations (4.31), (4.46) and (4.36) lead to

$$\begin{aligned}\mathbf{K} &= \mathbf{K}_{cc}\mathbf{P}^{-1} = [1 \quad 0 \quad 0]\mathbf{P}^{-1}P_{ch,\mathbf{A_K}}(\mathbf{A}) \\ &= [1 \quad 0 \quad 0]\mathbf{M}_{c,cc}\mathbf{M}_c^{-1}P_{ch,\mathbf{A_K}}(\mathbf{A}) = [0 \quad 0 \quad 1]\mathbf{M}_c^{-1}P_{ch,\mathbf{A_K}}(\mathbf{A}).\end{aligned} \tag{4.47}$$

As mentioned earlier, there is no difficulty in extending this method to systems of arbitrary order $n$ and the complete *Ackermann's formula* for the gain is therefore

$$\mathbf{K} = [0 \quad 0 \quad \dots \quad 0 \quad 1]\mathbf{M}_c^{-1}P_{ch,\mathbf{A_K}}(\mathbf{A}), \tag{4.48}$$

where the row vector to the right of the equal sign has the length $n$.

### 4.3.3 Conditions for Eigenvalue Assignment

Equation (4.48) shows that $\mathbf{K}$ can be found if $\mathbf{M}_c$ is nonsingular, i.e., if the system is controllable. This was also the case for the first design procedure in this section and it is concluded that controllability is a *sufficient* condition for arbitrary eigenvalue placement.

Considering an uncontrollable system the controllable subspace decomposition in Sect. 3.8.11 can be applied. A similarity transformation $\mathbf{z} = \mathbf{Q}^{-1}\mathbf{x}$ is used where $\mathbf{Q}$ is found from (3.338). If it is assumed that a gain matrix $\mathbf{K}$ has been found for the system (4.1) then it is clear that

$$\mathbf{K}_t = \mathbf{KQ}, \tag{4.49}$$

where $\mathbf{K}_t$ is the gain matrix for the transformed system.
The eigenvalues for the closed loop system are the solutions of the equation

$$det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{BK}) = 0. \tag{4.50}$$

Pre- and post-multiplying with the determinants of $\mathbf{Q}_{-1}$ and $\mathbf{Q}$ and carry through the following calculation:

$$\begin{aligned}det(\mathbf{Q}^{-1}) \cdot det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{BK}) \cdot det(\mathbf{Q}) &= det(\mathbf{Q}^{-1}(\lambda\mathbf{I} - \mathbf{A} + \mathbf{BK})\mathbf{Q}) \\ = det(\lambda\mathbf{I} - \mathbf{Q}^{-1}\mathbf{AQ} + \mathbf{Q}^{-1}\mathbf{BKQ}) &= det(\lambda\mathbf{I} - \mathbf{A}_t + \mathbf{B}_t\mathbf{K}_t) = 0.\end{aligned} \tag{4.51}$$

If the system's controllability matrix has the rank $p$ the gain matrix can be partitioned,

$$\mathbf{K}_t = [\mathbf{K}_{t1} \quad \mathbf{K}_{t2}], \tag{4.52}$$

such that $\mathbf{K}_{t1}$ has the length $p$ and $\mathbf{K}_{t2}$ has length $n - p$. Now insert the partitioned matrices into the last expression in Eq. (4.51) and see that

$$\begin{aligned} det(\lambda\mathbf{I} - \mathbf{A}_t + \mathbf{B}_t\mathbf{K}_t) &= det\left(\lambda\mathbf{I} - \begin{bmatrix} \mathbf{A}_c & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{nc} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_c \\ \mathbf{0} \end{bmatrix} [\,\mathbf{K}_{t1} \quad \mathbf{K}_{t2}\,]\right) \\ &= det\left(\left[\begin{array}{c:c} \lambda\mathbf{I}_p - \mathbf{A}_c + \mathbf{B}_c\mathbf{K}_{t1} & -\mathbf{A}_{12} + \mathbf{B}_c\mathbf{K}_{t2} \\ \hdashline \mathbf{0} & \lambda\mathbf{I}_{n-p} - \mathbf{A}_{nc} \end{array}\right]\right) \\ &= det(\lambda\mathbf{I}_p - \mathbf{A}_c + \mathbf{B}_c\mathbf{K}_{t1}) \cdot det(\lambda\mathbf{I}_{n-p} - \mathbf{A}_{nc}) = 0. \end{aligned} \tag{4.53}$$

The last expression shows that the closed loop eigenvalues consist of the $p$ eigenvalues which can be influenced by the gain matrix $\mathbf{K}_{t1}$ and the remaining $n - p$ eigenvalues of the uncontrollable system matrix $\mathbf{A}_{nc}$. These eigenvalues cannot be influenced by the feedback gains and will therefore remain in their original positions. Thus it is concluded that controllability is also a *necessary* condition for arbitrary eigenvalue placement.

Another important conclusion can be drawn from the result above. It is clear that all the eigenvalues in the controllable subspace can be assigned specific values, even if the controllable subsystem is not stable. If the eigenvalues of the uncontrollable system are also in the left half plane then the closed loop system can be made stable by a proper choice of $\mathbf{K}_{t1}$. So it can be seen that it is quite reasonable to call such a system *stabilizable*. See p. 124.

### System Zeros

If a full state feedback controller for the system in controller canonical form is designed, it will be noticed that the output matrix is not changed by the feedback. Since the coefficients of the numerator polynomial of the transfer function are given entirely by the $\mathbf{C}_{cc}$ matrix and since the transfer function is unique, it is obvious that the zeros of the system will remain unchanged by the state feedback. See also the remarks in the end of Sect. 4.2, p. 207.

### Discrete Time Systems

It should be pointed out that the design procedures detailed earlier in this section are also valid for discrete time systems. The only difference is the actual positions which are normally selected for the closed loop system eigenvalues.

If the system equations are

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k), \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k), \end{aligned} \tag{4.54}$$

the linear feedback law will be

$$\mathbf{u}(k) = -\mathbf{Kx}(k) + \mathbf{r}(k) \tag{4.55}$$

resulting in the closed loop equations

$$\begin{aligned} \mathbf{x}(k+1) &= (\mathbf{F} - \mathbf{GK})\mathbf{x}(k) + \mathbf{Gr}(k), \\ \mathbf{y}(k) &= \mathbf{Cx}(k). \end{aligned} \tag{4.56}$$

The closed loop system matrix is

$$\mathbf{F_K} = \mathbf{F} - \mathbf{GK}. \tag{4.57}$$

The eigenvalues are determined by the equation,

$$det(\lambda\mathbf{I} - \mathbf{F} + \mathbf{GK}) = 0. \tag{4.58}$$

The conditions for eigenvalue assignment stated above are equally valid for discrete time systems.

The relations between the eigenvalues in the continuous and discrete time domains are given by Eq. (3.87):

$$\lambda_\mathbf{F} = e^{\lambda_\mathbf{A} T} \Leftrightarrow \lambda_\mathbf{A} = \frac{1}{T}\log\lambda_\mathbf{F}. \tag{4.59}$$

Thus if a continuous time eigenvalue pair is given by

$$\lambda_\mathbf{A} = a \pm jb \tag{4.60}$$

the corresponding discrete time eigenvalue pair will be

$$\lambda_\mathbf{F} = e^{aT}(\cos bT \pm j\sin bT). \tag{4.61}$$

In general the controller designer will attempt to give the closed loop system well damped eigenvalues with a sufficiently high natural frequency, or if real eigenvalues are selected, time constants of a reasonable magnitude will be chosen. These requirements can be illustrated by the eigenvalue placements on Fig. 4.10. If the eigenvalues are placed in the shaded region, they will have a certain minimum damping and a certain minimum natural frequency (or a maximum time constant). For the discrete time system the corresponding region will be as shown on Fig. 4.11 (see also appendix D),

It is important to note that one should not try to place the eigenvalues too far to the left or too close to the origin in the continuous and discrete time cases respectively (see the dashed curves on the figures). This will result in large gain

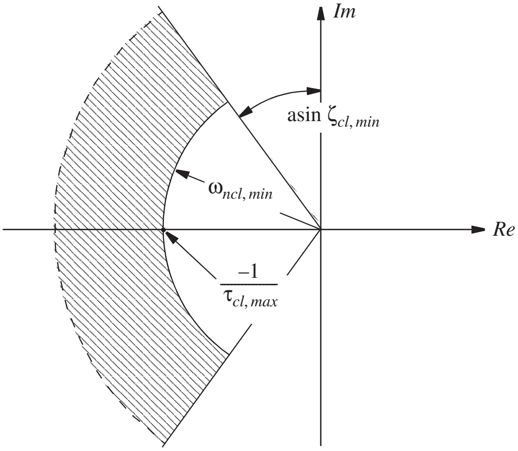


**Fig. 4.10** Eigenvalue placement for continuous time systems

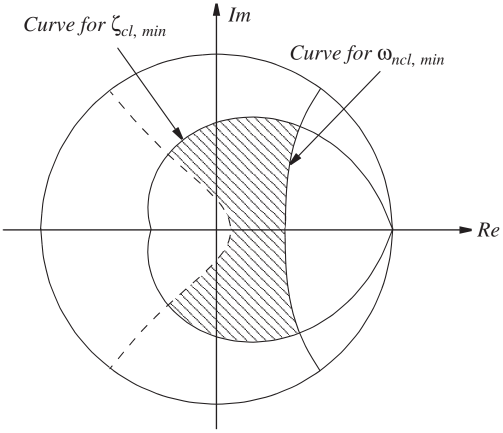


**Fig. 4.11** Eigenvalue placement for discrete time systems

values in the **K**-matrix and therefore also to excessively large input signals and high noise sensitivity.

***Example 4.5*. Continuous Control via the Controller Canonical Form**

The continuous third order system,

$$\dot{\mathbf{x}} = \begin{bmatrix} -0.14 & 0.33 & -0.33 \\ 0.1 & -0.28 & 0 \\ 0 & 1.7 & -0.77 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ -0.025 \end{bmatrix} u, \ y = [2 \quad 0 \quad 0]\mathbf{x}, \qquad (4.62)$$

will now be investigated. This system has the eigenvalues

$$\lambda = \begin{cases} -0.8986 \\ -0.1457 \pm j\,0.2157 \end{cases}.$$

which correspond to

$$\tau = 1.11 \text{ sec},$$
$$\omega_n = 0.26\,\text{sec}^{-1},\ \zeta = 0.56$$

The determinant of the controllability matrix for the system of Eq. (4.62) is nonzero, so the system is controllable. It is desired to design a state controller such that the closed loop system has real eigenvalues corresponding to time constants of about 1.5 s. Therefore choose the closed loop eigenvalues,

$$\lambda_{cl} = \begin{cases} -0.67 \\ -0.67, \\ -0.67 \end{cases} \tag{4.63}$$

which means that the closed loop characteristic polynomial will be

$$P_{ch}(\lambda) = \lambda^3 + 2.01\lambda^2 + 1.3467\lambda + 0.3008.$$

First the design method involving the controller canonical form will be used. The necessary similarity transformation matrix is found from Eq. (3.353). One has

$$\mathbf{P} = [\,\mathbf{p}_3\ \mathbf{p}_2\ \mathbf{p}_1\,] = \begin{bmatrix} 2.31 & 8.25 & 0 \\ 0.825 & 0 & 0 \\ -0.155 & -10.5 & -25 \end{bmatrix} \times 10^{-3}$$

and

$$\mathbf{P}^{-1} = \begin{bmatrix} 0 & 1212 & 0 \\ 121.1 & -339.4 & 0 \\ -50.9 & 135 & -40 \end{bmatrix}$$

and consequently

$$A_{cc} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -0.06087 & -0.3296 & -1.19 \end{bmatrix}, \quad \mathbf{B}_{cc} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

$$\mathbf{C}_{cc} = [\,0.00462 \quad 0.0165 \quad 0\,].$$

This is all that is required to determine the feedback gains from (4.30),

$$\mathbf{K}_{cc} = [k_1' \; k_2' \; k_3'],$$

where

$$\begin{aligned} k_1' &= \alpha_0 - a_0 = 0.3008 - 0.06087 = 0.2399, \\ k_2' &= \alpha_1 - a_1 = 1.3467 - 0.3296 = 1.10171, \\ k_3' &= \alpha_2 - a_2 = 2.01 - 1.19 = 0.82. \end{aligned}$$

The feedback matrix for the control canonical form has been found here. To transform back to the original system one has to use Eq. (4.31),

$$\mathbf{K} = \mathbf{K}_{cc}\mathbf{P}^{-1} = [81.54 \quad 56.26 \quad -32.8].$$

Using this gain matrix the eigenvalues of the closed loop system are found to be

$$det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{BK}) = 0 \Longrightarrow \lambda_{cl} = \begin{cases} -0.687 \pm j\,0.029 \\ -0.637 \end{cases}.$$

This is not precisely what was specified and the reason is small round off errors in the calculation above. The deviations are of no significant importance though. The system designed will have properties which can hardly be distinguished from the system with the specified eigenvalues, Eq. (4.63).

A simulation of the closed loop system,

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BK})\mathbf{x} + \mathbf{B}r, \;\; y = \mathbf{C}\mathbf{x}, \tag{4.64}$$

can now be carried out. If $r$ is a unit step, the responses on Fig. 4.12 are found. The initial state is a zero-vector. The output has an overshoot of approximately 12 percent, a feature which cannot be explained from the real eigenvalues. If the transfer function of the System (4.62) is calculated, one finds

$$\mathrm{G}(s) = \frac{y(s)}{u(s)} == \frac{0.0165s + 0.004262}{s^3 + 1.19s^2 + 3.296s + 0.06087}.$$

The system has the zero,

$$z = -0.28,$$

and since the state feedback does not influence the zeros, the same zero will be present in the closed loop system. This zero is close enough to the eigenvalues to be responsible for the overshoot on Fig. 4.12. ❒

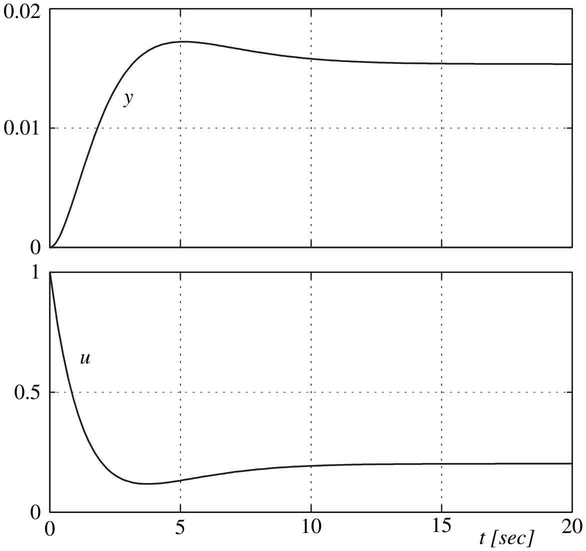


**Fig. 4.12** Output and input signals for unit step

***Example 4.6.*** **Discrete Control via the Controller Canonical Form**

Now a discrete time controller for the system in Example 4.5 will be designed. The procedure starts with the discretization of the continuous system using the Formulas (3.83)

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}u(k),\ \ y(k) = \mathbf{C}\mathbf{x}(k).$$

First an appropriate sample period must be selected, one which is approximately one fifth of the system time constant:

$$T = 0.2\,\text{sec}.$$

Using MATLAB's Control System Toolbox it is found that

$$\mathbf{F} = e^{\mathbf{A}T} = \begin{bmatrix} 0.9730 & 0.05293 & -0.06031 \\ 0.01918 & 0.9461 & -0.00061 \\ 0.003142 & 0.3063 & 0.8572 \end{bmatrix}$$

and

$$\mathbf{G} = \int_0^T e^{\mathbf{A}t}\mathbf{B}dt = \begin{bmatrix} 1.554 \cdot 10^{-4} \\ 1.0369 \cdot 10^{-6} \\ -4.635 \cdot 10^{-3} \end{bmatrix}.$$

Using the same closed loop eigenvalues as in the previous example, the discrete time eigenvalues will be,

$$\lambda_{cld} = e^{\lambda_{cl} T} = \begin{cases} 0.87459 \\ 0.87459 \,, \\ 0.87459 \end{cases}$$

which means that the closed loop characteristic polynomial will be

$$P_{ch,\mathbf{F_K}}(\lambda) = \lambda^3 - 2.6238\lambda^2 + 2.2947\lambda - 0.66898.$$

This time applying Ackermann's formula, the matrix polynomial is easily found by substituting **F** for λ in this polynomial,

$$P_{ch,\mathbf{F_K}}(\mathbf{F}) = \mathbf{F}^3 - 2.6238\mathbf{F}^2 + 2.2947\mathbf{F} - 0.66898\mathbf{I}.$$

The controllability matrix is found to be

$$\mathbf{M} = \left[\mathbf{G}\ \mathbf{FG}\ \mathbf{F}^2\mathbf{G}\right] = \begin{bmatrix} 155.39 & 430.72 & 658.96 \\ 1.0369 & 6.7881 & 17.107 \\ -4633.9 & -3971.4 & -3400.9 \end{bmatrix} \cdot 10^{-6}.$$

Some the elements of $\mathbf{M}_c$ are very small and the determinant of $\mathbf{M}_c$ is therefore also small:

$$det(\mathbf{M}_c) = -7.641 \cdot 10^{-12}.$$

Ackermann's formula requires the inverse of $\mathbf{M}_c$ and consequently the very small values for $det(\mathbf{M}_c)$ can cause numerical problems when the gain matrix is calculated. However, in this case the formula works properly and the gain matrix obtained is

$$\mathbf{K} = [0 \quad 0 \quad 1]\mathbf{M}_c^{-1} P_{ch,\,\mathbf{F_K}}(\mathbf{F}) = [74.634 \quad 47.719 \quad -30.393].$$

Testing the closed loop eigenvalues yields the result,

$$det(\lambda\mathbf{I} - \mathbf{F} + \mathbf{GK}) = 0 \Longrightarrow \lambda_{cl} = \begin{cases} 0.87827 \pm j\,0.002375 \\ 0.87724 \end{cases} .$$

This is close to the desired placement and it indicates that the matrix inversion did not cause significant numerical problems.

Problems may arise, however, if $det(\mathbf{M}_c)$ becomes too small. If, for instance, a sample period is selected which is ten times smaller, $T = T_1 = 0.02\,\text{s}$. Repeating the entire procedure, one will end up with a controllability matrix $\mathbf{M}_c1$ with

$$det(\mathbf{M}_{c1}) = -1.051 \cdot 10^{-17}.$$

The corresponding gain matrix for the same closed loop eigenvalues turns out to be,

$$\mathbf{K}_1 = [\,12974 \quad 265900 \quad -667.14\,],$$

or up to a factor 5000 larger than before. When the controllability matrix comes close to being singular (the system is close to 'lose controllability'), the controller must then work much harder to maintain the desired control. Gains of this size may give rise to severe accuracy and noise problems if they are implemented in a real system.

A discrete time simulation of the unit step response for the closed loop system with the gain matrix **K**,

$$\mathbf{x}(k+1) = (\mathbf{F} - \mathbf{GK})\mathbf{x}(k) + \mathbf{G}r(k), \; y(k) = \mathbf{C}\mathbf{x}(k),$$

is easily carried out using the state equation directly as a recursive formula. For a zero initial state vector the result on Fig. 4.13 is seen. The responses are almost exactly equal to the responses of the continuous system at the sampling instants which is also expected. It might be noted that the stationary value of the discrete time output differs slightly from that of the continuous system. This is caused by the fact that no attempt has been made to ensure that the *stationary gain* for the closed loop system should have a specific value. So far

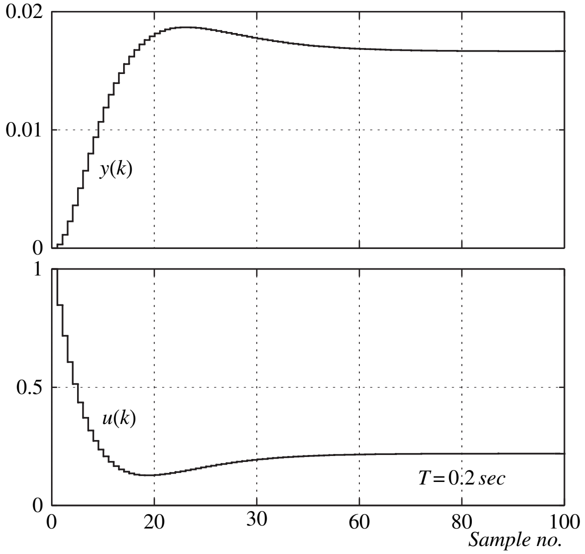


**Fig. 4.13** Output and input signals for unit step

the only problem considered has been of giving the systems certain *dynamic* properties. ❐

***Example 4.7*. Control of an Uncontrollable but Stabilizable System**

A state controller for the uncontrollable system below will now be designed:

$$\dot{\mathbf{x}} = \begin{bmatrix} -5 & -10 & 10 \\ 2 & -1 & -2 \\ 0 & -4 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 4 \\ 0 \\ 2 \end{bmatrix}u, \quad y = [1 \quad 0 \quad 0]\mathbf{x}.$$

The eigenvalues are

$$\lambda_{\mathbf{A}} = \begin{cases} -1 \pm j2 \\ -3 \end{cases}$$

and although the system is uncontrollable, it is stabilizable because all eigenvalues are in the left half plane.

The controllability matrix and its determinant are readily found:

$$\mathbf{M}_c \begin{bmatrix} 4 & 0 & -20 \\ 0 & 4 & -8 \\ 2 & 2 & -14 \end{bmatrix}, \quad det(\mathbf{M}_c) = 0.$$

The rank of $\mathbf{M}_c$ is 2 and it is immediately seen that the two first columns are linearly independent.

The matrix required for the controllability subspace decomposition can be for instance (see Eq. (3.338)),

$$\mathbf{Q} = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 4 & 1 \\ 2 & 2 & 0 \end{bmatrix},$$

which has

$$det(\mathbf{Q}) = -16 \quad \text{and} \quad \mathbf{Q}^{-1} = \begin{bmatrix} 0.125 & -0.125 & 0.25 \\ -0125 & 0.125 & 0.25 \\ 0.5 & 0.5 & -1 \end{bmatrix}.$$

The similarity transformation $\mathbf{z} = \mathbf{Q}^{-1}\mathbf{x}$ gives the matrices

$$\mathbf{A}_t = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \left[\begin{array}{cc:c} 0 & -5 & -3 \\ 1 & -2 & 1 \\ \hdashline 0 & 0 & -3 \end{array}\right]], \mathbf{B}_t = \mathbf{Q}^{-1}\mathbf{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

$$\mathbf{C}_t = \mathbf{C}\mathbf{Q} = \left[\begin{array}{cc:c} 4 & 0 & 0 \end{array}\right],$$

and the controllable part of the system is described by the matrices,

$$\mathbf{A}_c \begin{bmatrix} 0 & -5 \\ 1 & -2 \end{bmatrix}, \quad \mathbf{B}_c = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{C}_c = [4 \quad 0],$$

with the nonsingular controllability matrix,

$$\mathbf{M}_{c,c} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Since this system is controllable the two eigenvalues can be assigned arbitrary values by the two-dimensional gain matrix $\mathbf{K}_{t1}$ (see Eqs. (4.52) and (4.53)). If the eigenvalues and characteristic polynomial are chosen as

$$\lambda_{\mathbf{A}_{\mathbf{K}_{t1}}} = \begin{cases} -4 \\ \\ -5 \end{cases} \Longrightarrow \mathbf{P}_{ch,\,t1} = \lambda^2 + 9\lambda + 20,$$

the gain matrix can be found from Ackermann's formula,

$$\mathbf{K}_{t1} = [0\ 1]\mathbf{M}_{c,c}^{-1}P_{ch,\,t1}(\mathbf{A}_c) = [7\ 1].$$

The third element in the total gain matrix $\mathbf{K}_t$ has no influence on the closed system eigenvalues and it can be assigned any value, for instance zero,

$$\mathbf{K}_t = [\mathbf{K}_{t1} \quad \mathbf{K}_{t2}] = [7 \quad 1 \quad 0].$$

Re-transformation to the original basis gives finally

$$\mathbf{K} = \mathbf{K}_t\mathbf{Q}^{-1} = [0.75 \quad -0.75 \quad 2].$$

Note that with the above choice of **Q**, all states are fed back. But in contrast to the controllable case, the **K**-matrix is *not* unique. As a matter of fact it is dependent on the specific choice of the columns of **Q** which can easily be verified.

One can test the closed loop eigenvalues as usual by solving

$$det(\lambda \mathbf{I} - \mathbf{A} + \mathbf{BK}) = 0$$

and obtain

$$\lambda_{cl} = \begin{cases} -4 \\ -5, \\ -3 \end{cases}$$

which is precisely as expected. The uncontrollable eigenvalue $\lambda = -3$ is kept unchanged and the two other eigenvalues are as assigned. ❐

***Example 4.8*. Control of the Labyrinth Game**

It is an interesting exercise to derive a position regulator for the labyrinth game presented in Example 4.2. This can be accomplished by considering one degree of freedom at a time, for example the $x$ direction. Figure 4.14 shows a schematic drawing of the forces working on the ball on the game board in the $x$ direction. Applying Newton's Second Law to the system, assuming that the angle $\theta$ is small and the acceleration $\ddot{\theta}$ is negligible, gives

$$\begin{aligned} m\ddot{x} &= mg\theta - F_f + m\dot{\theta}^2 x, \\ I\dot{\omega} &= rF_f. \end{aligned}$$

If it is assumed that the ball rolls on the board, it must be true that

$$\ddot{x} = r\dot{\omega}$$

Inserting the moment of inertia for the ball,

$$I = \frac{2}{5}mr^2,$$

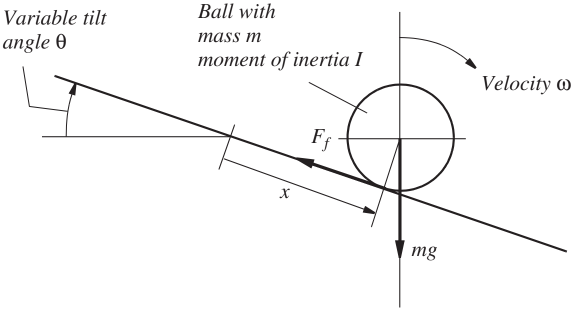


**Fig. 4.14** Schematic drawing of the labyrinth game board

and eliminating $F_f$ and $\omega$ from the three equations above leads to the result:

$$\ddot{x} = \frac{5}{7}(g\theta + x\dot{\theta}^2). \tag{4.65}$$

The Model (4.65) is nonlinear and it has two inputs, $\theta$ and $\dot{\theta}$.
If the position and velocity vectors are defined as

$$\mathbf{x} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad \mathbf{u} = \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

the model can be described with the state equation,

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) = \begin{bmatrix} x_2 \\ \frac{5}{7} g u_1 + \frac{5}{7} x_1 u_2^2 \end{bmatrix}.$$

Linearizing the model around the stationary state $\mathbf{x}_0 = 0$, $\mathbf{u}_0 = 0$, one has

$$\dot{\Delta}\mathbf{x} = \mathbf{A}\Delta\mathbf{x} + \mathbf{B}'\Delta\mathbf{u}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B}' \begin{bmatrix} 0 & 0 \\ \frac{5}{7} g & 0 \end{bmatrix}.$$

Since the last column of **B** is zero, the system is reduced to a single input system with the input matrix

$$\mathbf{B} = \begin{bmatrix} 0 \\ \frac{5}{7} g \end{bmatrix}.$$

The **A**-matrix shows that the linear system approximating the nonlinear one is nothing but a double integrator: the acceleration of the ball is simply proportional to the board tilt angle.

It is desired to design a discrete time controller for the system and since the controller involves video information, the sample period is given by the CCIR video scan rate. This is 25 Hz. So choose $T = 0.04$ s. In this simple case the discretization can be carried out by hand using the Formulas (3.83):

$$\mathbf{F} = e^{AT} = \mathscr{L}^{-1}\{(s\mathbf{I}-\mathbf{A})^{-1})\}_{t=T} = \mathscr{L}^{-1}\left\{\begin{bmatrix} s & -1 \\ 0 & s \end{bmatrix}^{-1}\right\}_{t=T}$$

$$= \mathscr{L}^{-1}\left\{\begin{bmatrix} \frac{1}{s} & \frac{1}{s^2} \\ 0 & \frac{1}{s} \end{bmatrix}\right\}_{t=T} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix},$$

$$\mathbf{G} = \int_0^T e^{\mathbf{A}t}\mathbf{B}dt = \int_0^T \begin{bmatrix} \alpha t \\ \alpha \end{bmatrix} dt = \begin{bmatrix} \frac{\alpha}{2}T^2 \\ \alpha T \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix},$$

where $\alpha = 5g/7$.
Inserting the values for $T$ and $g$ finally gives the model,

$$\Delta\mathbf{x}(k+1) = \begin{bmatrix} 1 & 0.04 \\ 0 & 1 \end{bmatrix}\Delta\mathbf{x} + \begin{bmatrix} 5.606\cdot 10^{-3} \\ 0.2803 \end{bmatrix}\Delta u.$$

The controllability matrix is

$$\mathbf{M}_c = [\mathbf{G}\ \mathbf{FG}] = \begin{bmatrix} g_1 & g_1 + Tg_2 \\ g_2 & g_2 \end{bmatrix} = \begin{bmatrix} 5.606\cdot 10^{-3} & 0.01681 \\ 0.2803 & 0.2803 \end{bmatrix}, \qquad (4.66)$$

$det(\mathbf{M}_c) \neq 0$ and the system is controllable.

The closed loop system is rquired to have a natural frequency of approximately 1 Hz ($\omega_{ncl} = 6.28$ rad/s) and the damping ratio $\zeta_{cl} = 0.5$. The continuous eigenvalues providing these properties are

$$\lambda_c = -3.14 \pm j5.44.$$

The corresponding discrete time eigenvalues are found from Eq. (4.61),

$$\lambda_d = 0.8611 \pm j0.1904,$$

which implies that the characteristic polynomial is

$$P_{ch}(\lambda) = \lambda^2 - 1.7222\lambda + 0.7778.$$

The gain matrix is readily found using Ackermann's formula,

$$\mathbf{K} = [0\ 1]\mathbf{M}_c^{-1}P_{ch}(\mathbf{F}) = [4.959\ 0.8919].$$

A closed loop step response is shown on Fig. 4.15 ❐

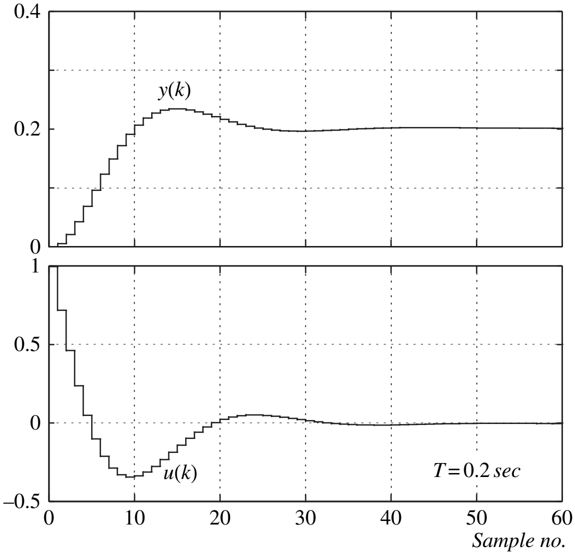


**Fig. 4.15** Output and control signal for unit step input

## 4.4 State Feedback for MIMO Systems

The design methods in Sect. 4.3 are simple but unfortunately only valid for systems with one input. For MIMO systems (or rather for systems with multiple inputs, MI systems, since the output is not involved in the control) the situation is somewhat different and more complicated.

A more comprehensive treatment of MIMO state feedback will be left until Chap. 5 and in this section the treatment will be confined to a quick overview of a couple of available methods for MIMO feedback design. Only the continuous time case will be treated but the methods work equally well for discrete time systems.

The main principle in the control is still the linear state feedback according to the control law:

$$\mathbf{u} = -\mathbf{K}\mathbf{x} + \mathbf{r}. \tag{4.67}$$

Inserting this control signal in the state equation gives, as seen before, the closed loop state equation,

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}\mathbf{K})\mathbf{x} + \mathbf{B}\mathbf{r} = \mathbf{A}_{\mathbf{K}}\mathbf{x} + \mathbf{B}\mathbf{r}. \tag{4.68}$$

Eigenvalue assignment means that a gain vector **K** must be found such that the eigenvalues assume preselected values:

$$\lambda_{\mathbf{A_K}} = \begin{cases} \lambda_1 \\ \lambda_2 \\ : \\ \lambda_n \end{cases} . \tag{4.69}$$

It can be shown, that the conditions for eigenvalue assignment for SISO systems discussed in Sect. 4.3.3 are also valid in the MIMO case:

1. The closed loop eigenvalues for the System (4.68) can be placed arbitrarily if and only if the system is controllable.
2. The System (4.68) can be stabilized if and only if the uncontrollable states are stable (the system is stabilizable).

### *4.4.1 Eigenstructure Assignment for MIMO Systems*

Whereas the eigenvalue placement can be accomplished *uniquely* (i.e., for a unique **K**) for any controllable SISO system, it turns out that further options exist for MIMO systems. Besides assigning the *eigenvalues*, it is also possible to make some demands on the *eigenvectors*. This is important, because the eigenvectors determine the system's response as seen from the modal decomposition in Sect. 3.8.10.

A controller design involving eigenvalues as well as eigenvectors is called *eigenstructure assignment*. The closed loop eigenvectors are defined by the relation

$$(\mathbf{A} - \mathbf{BK})\mathbf{v}_i = \lambda_i \mathbf{v}_i. \tag{4.70}$$

This equation can be rearranged to obtain

$$[\lambda_i \mathbf{I} - \mathbf{A} \quad \mathbf{B}] \begin{bmatrix} \mathbf{v}_i \\ \mathbf{K}\mathbf{v}_i \end{bmatrix} = \mathbf{0}. \tag{4.71}$$

For each $\lambda_i$ (4.71) is a homogeneous set of linear equations. The set consists of $n$ equations with $n+m$ unknowns. It is known from the controllability theorem CC3 that the left hand matrix (the coefficient matrix) has full rank $n$ if and only if the open loop system is controllable. Then the theory of linear equations requires that there will be $m$ linearly independent solutions for each of the chosen $\lambda_i$. This indicates again, that in the SISO case there will be only one solution since $m=1$.

Equation (4.71) also shows that one cannot pick eigenvalues and eigenvectors completely arbitrarily. The solution vector,

$$\mathbf{p}_i = \begin{bmatrix} \mathbf{v}_i \\ \mathbf{K}\mathbf{v}_i \end{bmatrix} = \begin{bmatrix} \mathbf{v}_i \\ \mathbf{q}_i \end{bmatrix}, \tag{4.72}$$

must lie in the nullspace N of the coefficient matrix $\mathbf{T}_i = [\lambda_i \mathbf{I} - \mathbf{A} \quad \mathbf{B}]$.

The nullspace of $T_i$ is spanned by a set of $m$ linearly independent vectors $\mathbf{n}_{i,1}, \mathbf{n}_{i,2}, \ldots, \mathbf{n}_{i,m}$ each of dimension $n+ m$. The solution vector $\mathbf{p}_i$ is in the nullspace if it can be composed of a linear combination of the vectors $\mathbf{n}_{i,1}, \mathbf{n}_{i,2}, \ldots, \mathbf{n}_{i,m}$. According to the definition in (4.72) one has that $\mathbf{K}\mathbf{v}_i = \mathbf{q}_i$ from which

$$\mathbf{K}[\mathbf{v}_1 \quad \mathbf{v}_2 \quad \ldots \quad \mathbf{v}_n] = [\mathbf{q}_1 \quad \mathbf{q}_2 \quad \ldots \quad \mathbf{q}_n] \tag{4.73}$$

or

$$\mathbf{K}\mathbf{V} = \mathbf{Q}. \tag{4.74}$$

If all the eigenvectors are selected to be linearly independent then the gain matrix below can be found,

$$\mathbf{K} = \mathbf{Q}\mathbf{V}^{-1}. \tag{4.75}$$

**K** must be a real valued matrix but this is assured if the eigenvalues as well as the eigenvectors are selected as complex conjugate pairs.

Thus to complete the design task the $n$ eigenvalues $\lambda_i$ must be chosen, $n$ vectors $\mathbf{p}_i$ determined satisfying (4.71) and then **K** computed from (4.75). The procedure may seem simple, but it can be quite laborious for higher order systems.

***Example 4.9.* Eigenstructure Control of a MISO System**

A design for an eigenstructure state controller for the MISO-system,

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 0 & 1 \\ -5 & 1 & -3 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{u}, \quad y = [0 \quad 0 \quad 1]\mathbf{x}, \tag{4.76}$$

is to be made here. The eigenvalues of the system matrix are

$$\lambda_{\mathbf{A}} = \begin{cases} -1 \\ 1 \pm j2 \end{cases}.$$

The system has a complex conjugate eigenvalue pair in the right half plane and it is therefore unstable.

The system is to be stabilized with the closed loop eigenvalues,

$$\lambda_{\mathbf{A_K}} = \begin{cases} -2 \\ -3 \\ -4 \end{cases} . \qquad (4.77)$$

The matrix $\mathbf{T}_i$ is

$$\mathbf{T}_i = [\lambda_i \mathbf{I} - \mathbf{A} \quad \mathbf{B}] = \begin{bmatrix} \lambda_i & 0 & -1 & 1 & 0 \\ 5 & \lambda_1 - 1 & 3 & 0 & 0 \\ 0 & -1 & \lambda_i & 0 & 1 \end{bmatrix} .$$

Now the vectors spanning the nullspace have to be found. This can be done in several ways but an easy one is selected here: the MATLAB function `null`.

The following results for the three closed loop eigenvalues are found:

$\lambda_1 = -2,$

$$\mathbf{T}_1 = \begin{bmatrix} -2 & 0 & -1 & 1 & 0 \\ 5 & -3 & 3 & 0 & 0 \\ 0 & -1 & -2 & 0 & 1 \end{bmatrix}, \mathbf{n}_{1,1} = \begin{bmatrix} 0.6923 \\ 0.7692 \\ -0.3846 \\ 1 \\ 0 \end{bmatrix}, \mathbf{n}_{1,2} = \begin{bmatrix} -0.2308 \\ 0.0769 \\ 0.4615 \\ 0 \\ 1 \end{bmatrix} .$$

$\lambda_2 = -3,$

$$\mathbf{T}_2 = \begin{bmatrix} -3 & 0 & -1 & 1 & 0 \\ 5 & -4 & 3 & - & - \\ 0 & -1 & -3 & - & 1 \end{bmatrix}, \mathbf{n}_{2,1} = \begin{bmatrix} 0.375 \\ 0.375 \\ -0.125 \\ 1 \\ 0 \end{bmatrix}, \mathbf{n}_{2,2} = \begin{bmatrix} -0.1 \\ 0.1 \\ 0.3 \\ 0 \\ 1 \end{bmatrix} .$$

$\lambda_3 = -4,$

$$\mathbf{T}_3 = \begin{bmatrix} -4 & 0 & -1 & 1 & 0 \\ 5 & -5 & 3 & 0 & 0 \\ 0 & -1 & -4 & 0 & 1 \end{bmatrix}, \mathbf{n}_{3,1} = \begin{bmatrix} 0.2644 \\ 0.2299 \\ -0.0575 \\ 1 \\ 0 \end{bmatrix}, \mathbf{n}_{3,2} = \begin{bmatrix} -0.0575 \\ 0.0805 \\ 0.2299 \\ 0 \\ 1 \end{bmatrix} .$$

The next step is to assign the **p**-vectors and hereby the eigenvectors of the closed loop system. The three **p**-vectors can be constructed as arbitrary linear combinations of the respective **n**-vector pairs,

$$\mathbf{p}_i = \alpha_{i,1}\mathbf{n}_{i,1} + \alpha_{i,2}\mathbf{n}_{i,2}.$$

The three uppermost entries of the $\mathbf{p}_i$-vector constitute the eigenvector of the closed loop system. If the l'th entry is selected to be zero then the corresponding eigenvalue, $\lambda_i$ (and natural mode $e^{\lambda_i t}$), will not appear in the l'th state variable. This follows from Eq. (3.332).

Following this scheme it is easy to obtain the following **p**-vectors:

$$\mathbf{p}_1 = \begin{bmatrix} 0 \\ -0.2308 \\ -0.2308 \\ -0.2308 \\ -0.6923 \end{bmatrix}, \quad \mathbf{p}_2 = \begin{bmatrix} 0.074 \\ 0 \\ -0.125 \\ 0.1 \\ -0.375 \end{bmatrix}, \quad \mathbf{p}_3 = \begin{bmatrix} 0.0575 \\ 0.0575 \\ 0 \\ 0.2299 \\ 0.0575 \end{bmatrix}.$$

Now the matrices **V** and **Q** can be found by inspection,

$$\mathbf{V} = \begin{bmatrix} 0 & 0.075 & 0.0575 \\ -0.2308 & 0 & 0.0575 \\ -0.2308 & -0.125 & 0 \end{bmatrix},$$

$$\mathbf{Q} = \begin{bmatrix} -0.2308 & 0.1 & 0.2299 \\ -0.6923 & -0.375 & 0.0575 \end{bmatrix},$$

and finally the gain matrix is

$$\mathbf{K} = \mathbf{Q}\mathbf{V}^{-1}\begin{bmatrix} 5.5 & -1.5 & 2.5 \\ 2.5 & -1.5 & 4.5 \end{bmatrix}.$$

Testing the closed loop eigenvalues assures that the requirement (4.77) is met. The unit step responses for the closed loop system are shown on Fig. 4.16. The two curves are the output $y(t)$ for a unit step on each of the two reference inputs. ❐

**Robust Eigenstructure Assignment**

The freedom to select eigenvectors in multi-input state feedback can be used to increase the *robustness* of the control system. This subject will not be pursued at this point but it is only mentioned that robustness against deterioration of the

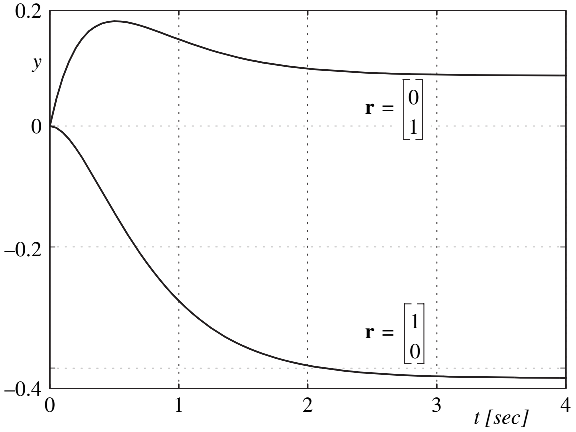


**Fig. 4.16** Unit step response for the MISO system

system performance caused by imprecise or incomplete modelling or by deviations between the linearization state and the actual operating state is an important issue. It can be shown, that if the closed loop eigenvectors $\mathbf{v}_i$ are chosen orthogonal (or as close to orthogonality as possible) then sensitivity of the eigenvalue location subject to parameter changes in the model will be minimized.

Orthogonality algorithms have been developed and one of them is implemented in the MATLAB function `place`. This algorithm is based on a paper by Kautsky et al. (1985).

### *Example 4.10.* Eigenstructure Control of a MISO System with Matlab

Returning to the system from Example 4.9 and using the MATLAB `place` function with the same eigenvalues as before a different gain matrix can be found. The function is invoked by the command,

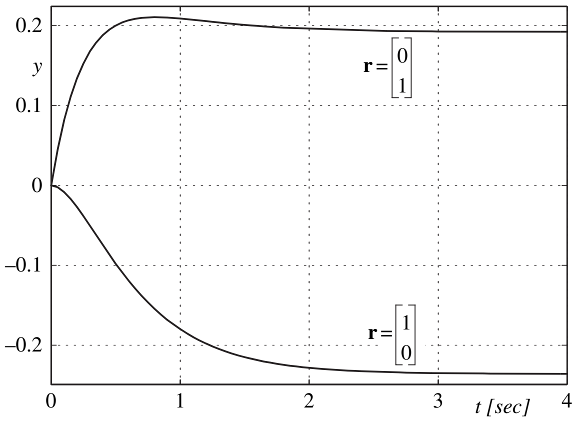
$$\mathtt{K} = \mathtt{place}(\mathtt{A}, \mathtt{B}, [-\mathtt{2} \ \ -\mathtt{3} \ \ -\mathtt{4}])$$

and the result is the gain matrix,

$$\mathbf{K} = \begin{bmatrix} 5.785 & -2.080 & 2.681 \\ 2.016 & -0.5358 & 4.215 \end{bmatrix}.$$

A simulation similar to the one in Example 4.9 leads to the result shown on Fig. 4.17. Apart from the stationary values, the responses resemble those of Example 4.9. It is noted though that the overshoot for the step input to input channel no. 2 is much smaller here than on Fig. 4.16. In this respect the eigenvector assignment is more favorable with the MATLAB `place` function than with the scheme used in Example 4.9. ❐

**Fig. 4.17** Unit step response for the MISO system (with the MATLAB `place` function)



## *4.4.2 Dead Beat Regulators*

Usually a continuous time controller has a discrete time equivalent and vice versa. However there is an exception to this rule. One can design discrete time controllers which *do not* have a continuous counterpart and which are very fast.

The reason for this is that the transformation which is used to discretize continuous systems has the particular characteristic that continuous time eigenvalues with $Re(\lambda_c) \cong -\infty$ are mapped into discrete time eigenvalues at the origin. If a state controller is designed for all closed loop eigenvalues placed at the origin, the characteristic equation will be very simple,

$$P_{ch,\mathbf{F}_\mathbf{K}}(\lambda) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \ldots + \alpha_1\lambda + \alpha_0 = \lambda^n. \tag{4.78}$$

All the α-coefficients are zero.

Cayley-Hamilton’s theorem states that every quadratic matrix satisfies its own characteristic equation or

$$\mathbf{F}_\mathbf{K}^n + \alpha_{n-1}\mathbf{F}_\mathbf{K}^{n-1} + \ldots + \alpha_1\mathbf{F}_\mathbf{K} + \alpha_0\mathbf{I} = \mathbf{0}. \tag{4.79}$$

But, since all the coefficients are zero, this leads to

$$\mathbf{F}_\mathbf{K}^n = \mathbf{0} \tag{4.80}$$

which obviously means that

$$\mathbf{F}_\mathbf{K}^p = \mathbf{0} \quad \text{for} \quad p \geq n. \tag{4.81}$$

If the asymptotically stable closed loop system,

$$\mathbf{x}(k+1) = \mathbf{F_K}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k), \tag{4.82}$$

is released in the initial state $\mathbf{x}_0$ and $\mathbf{u}(\mathrm{k})=0$ for all $k$ then a response will be found such that (see Eq. (3.98))

$$\mathbf{x}(k) = \mathbf{F}_\mathbf{k}^k\mathbf{x}_0. \tag{4.83}$$

According to (4.80) this will then achieve the result that

$$\mathbf{x}(k) = \mathbf{0} \quad \text{for} \quad k \geq n. \tag{4.84}$$

In other words, the system comes to rest in at most $n$ sample periods. Systems with this behaviour are called *dead-beat* or *finite settling time* systems.

If one lets $\mathbf{u}(k) = \mathbf{u}_0$ (a constant vector) then the zero-state output can also be found from Eq. (3.98),

$$\mathbf{x}(k) = \sum_{i=0}^{k-1} \mathbf{F}_\mathbf{K}^{k-1-i}\mathbf{G}\mathbf{u}_0 = (\mathbf{F}_\mathbf{k}^{k-1} + \mathbf{F}_\mathbf{K}^{k-2} + \ldots + \mathbf{F_K} + \mathbf{I})\mathbf{G}\mathbf{u}_0. \tag{4.85}$$

This shows that because of (4.80), the output will be constant for $k \geq n$. So the system also has a dead-beat behaviour for a constant nonzero input.

With a dead-beat regulator the system can apparently be made as fast as desired. It is just a question of choosing a sufficiently small sample period. But of course there are limits. The cost of the very rapid response is very large feedback gains which implies large actuator drive signals and high noise sensitivity. It also implies a large sensitivity to unmodelled dynamics. This in turn can lead to poor transient response and in the worst case, instability.

For a deadbeat regulator the only design parameter is the sampling time. This means that for any given plant several different sample periods will have to be tried in order to find one which will yield the desired performance of that system. This is especially true in the presence of modelling error or neglected dynamics. In any case it must be remembered that the input signals will usually be large compared to those in a more common eigenvalue placement regulator. For these reasons the practical implementation of a dead-beat controller can be somewhat problematic.

***Example 4.11.* Deadbeat Control of the Labyrinth Game**

If a dead-beat controller is designed for the labyrinth game board from Example 4.8, Ackermann's formula can be used with (4.66) because this system is SISO.

The characteristic matrix polynomial is

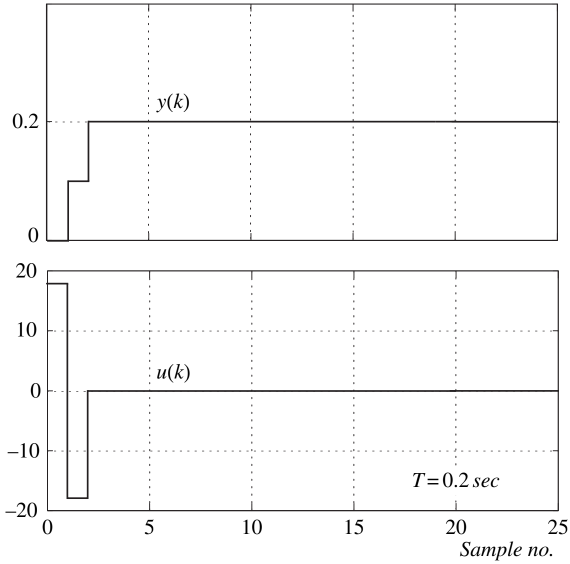
$$p_{ch}(\mathbf{F}) = \mathbf{F}^2 = \begin{bmatrix} 1 & 2T \\ 0 & 1 \end{bmatrix}.$$

One finds a gain matrix:

$$\mathbf{K} = [0 \quad 1]\mathbf{M}_c^{-1}\mathbf{F}^2 = \left[\frac{1}{\alpha T^2} \quad \frac{3}{2\alpha T}\right] = [89.2 \quad 5.352].$$

A step response with $y$ reaching the same final value as in Example 4.8 is shown on Fig. 4.18. Comparing this with Fig. 4.15, one can clearly see that the settling time has been reduced from approximately 50 times the sample period (10 s) to exactly 2 (0.4 s). The price is that the maximum value of the control signal is increased by a factor of 18. ❐

**Fig 4.18** Output and control signal for step input for the labyrinth game



## 4.5 Integral Controllers

One of the drawbacks of state variable feedback is that the output of the system is not directly involved in the control. This means that the controller design does not provide the possibility of assigning a predetermined stationary relationship between the inputs and the outputs.

Consider the closed loop system with state feedback law:

$$\dot{\mathbf{x}} = \mathbf{A_K}\mathbf{x} + \mathbf{B}\mathbf{r}. \tag{4.86}$$

If is is assumed that the system is in a stationary state then the time derivative of the state vector is zero,

$$\mathbf{0} = \mathbf{A_K x}_0 + \mathbf{Br}_0. \tag{4.87}$$

The system matrix $\mathbf{A_K}$ will probably not have any eigenvalues at the origin so it will be nonsingular and the stationary state can be found by solving Equation (4.87) for $\mathbf{x}_0$:

$$\mathbf{x}_0 = -\mathbf{A_K}^{-1}\mathbf{Br}_0. \tag{4.88}$$

The stationary output with this state will be

$$\mathbf{y}_0\mathbf{Cx}_0 = -\mathbf{CA_K}^{-1}\mathbf{Br}_0. \tag{4.89}$$

If the system has the same number of inputs and outputs ($m = r$), the right hand matrix will be square and the necessary reference vector for a given output can finally be determined to be

$$\mathbf{r}_0 = -(\mathbf{CA_K}^{-1}\mathbf{B})^{-1}\mathbf{y}_0, \tag{4.90}$$

provided of course that the matrix $\mathbf{CA_K}^{-1}\mathbf{B}$ is also nonsingular.

It should be kept in mind however that the basis for the state controller design is a linear model which is usually the result of a linearization of a nonlinear system model at some stationary operating point. This means that even in cases where the reference vector (4.90) can be found, it is only useful if the system's state is exactly the stationary state on which the linearization was based. Another difficulty arises if the system is subject to disturbances since the 'plain' state controller does not take a nonzero **v**-vector into account. Fortunately, as in classical control, all of the problems mentioned here can be more or less overcome by a single means: system augmentation with integrators.

Integration can be included in state feedback control in several ways but the following scheme is the most straightforward. Figure 4.19 shows the overall system. The state is fed back as before, but in addition the output is also measured and fed back in an outer loop to a new primary summation point. The output of this summing point is the *system error vector*. $\mathbf{e}(t)$. The output vector and the reference vector must have the same dimension, $r$, but since a new gain matrix $\mathbf{K}_i$ has been inserted, this need not be the same as the number of inputs to the system, i.e., the dimension $m$ of the input vector $u$. As usual, the system can only be in a stationary state if the outputs $\mathbf{x}_\mathrm{i}$ of the integrators are constant. This can only be the case if the error is zero since

$$\mathbf{e} = \dot{\mathbf{x}}_i. \tag{4.91}$$

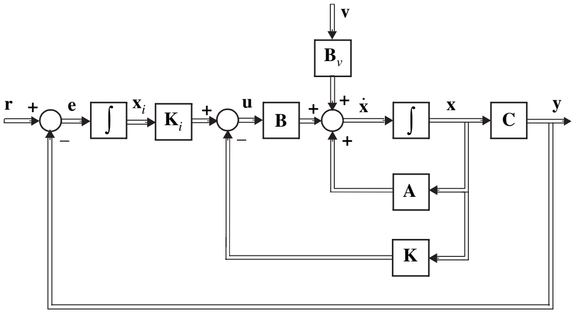


**Fig. 4.19** State feedback with integration

This is the case *no matter which constant values* of **r** and **v** are imposed on the system.

By adding integrators the order of the system is by increased the number of integrators used. Therefore it is natural to define a new *augmented* state vector,

$$\mathbf{x}_a = \begin{bmatrix} \mathbf{x} \\ \mathbf{x}_i \end{bmatrix}. \tag{4.92}$$

The new state vector will have the dimension $n_a = n + r$, where $r$ is the number of integrator states. The equations governing the augmented system are

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{B}_v\mathbf{v}, \\ \mathbf{u} &= -\mathbf{K}\mathbf{x} + \mathbf{K}_i\mathbf{x}_i, \\ \dot{\mathbf{x}}_i &= -\mathbf{C}\mathbf{x} + \mathbf{r}, \\ \mathbf{y} &= \mathbf{C}\mathbf{x}. \end{aligned} \tag{4.93}$$

This can be written in matrix form as

$$\begin{aligned} \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{x}}_i \end{bmatrix} &= \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ -\mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{x}_i \end{bmatrix} + \begin{bmatrix} \mathbf{B} \\ \mathbf{0} \end{bmatrix} \mathbf{u} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_r \end{bmatrix} \mathbf{r} + \begin{bmatrix} \mathbf{B}_v \\ \mathbf{0} \end{bmatrix} \mathbf{v}, \\ \mathbf{y} &= [\mathbf{C} \quad \mathbf{0}] \begin{bmatrix} \mathbf{x} \\ \mathbf{x}_i \end{bmatrix}, \end{aligned} \tag{4.94}$$

or using the augmented state vector,

$$\begin{aligned} \dot{\mathbf{x}}_a &= \mathbf{A}_1\mathbf{x}_a + \mathbf{B}_1\mathbf{u} + \mathbf{B}_r\mathbf{r} + \mathbf{B}_{v1}\mathbf{v}, \\ \mathbf{y} &= \mathbf{C}_1\mathbf{x}_a. \end{aligned} \tag{4.95}$$

$\mathbf{I}_r$ is the $r$-dimensional identity matrix and

$$\mathbf{A}_1 = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ -\mathbf{C} & \mathbf{0} \end{bmatrix}, \ \mathbf{B}_1 = \begin{bmatrix} \mathbf{B} \\ \mathbf{0} \end{bmatrix}, \ \mathbf{B}_r = \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_r \end{bmatrix}, \ \mathbf{B}_{v1} = \begin{bmatrix} \mathbf{B}_v \\ \mathbf{0} \end{bmatrix}, \ \mathbf{C}_1 = [\mathbf{C} \quad \mathbf{0}]. \tag{4.96}$$

The input equation can be written,

$$\mathbf{u} = -[\mathbf{K} \quad -\mathbf{K}_i] \begin{bmatrix} \mathbf{x} \\ \mathbf{x}_i \end{bmatrix} = -\mathbf{K}_1 \mathbf{x}_a, \tag{4.97}$$

where

$$\mathbf{K}_1 = [\mathbf{K} \quad -\mathbf{K}_i]. \tag{4.98}$$

Inserting this into Eq. (4.95) leads to

$$\dot{\mathbf{x}}_a = (\mathbf{A}_1 - \mathbf{B}_1\mathbf{K}_1)\mathbf{x}_a + \mathbf{B}_r\mathbf{r} + \mathbf{B}_{v1}\mathbf{v}. \tag{4.99}$$

Apart from the fact that a disturbance term has been added, this equation is precisely the same as Eq. (4.4), with the augmented matrices replacing the original ones. As before, if it is desired to design the state controller by eigenvalue or eigenstructure assignment, then the gain matrix $\mathbf{K}_1$ must be found such that the solutions to the equation,

$$det(\lambda\mathbf{I} - \mathbf{A}_1 + \mathbf{B}_1\mathbf{K}_1) = det\left(\lambda\mathbf{I} - \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{K} & \mathbf{B}\mathbf{K}_i \\ -\mathbf{C} & \mathbf{0} \end{bmatrix}\right) = 0, \tag{4.100}$$

are the desired eigenvalues.

### Discrete Time Systems

The integrating state controller can also be implemented in discrete time systems. The continuous time integrator on Fig. 4.19 is an $r$-dimensional dynamic system whose state equation can be written, see Eq. (4.91),

$$\dot{\mathbf{x}}_i = \mathbf{0}\mathbf{x}_i + \mathbf{e}, \tag{4.101}$$

where the 'system matrix' is a zero matrix with $r$ eigenvalues which are all zero. The discrete equivalent to this system is a system with $r$ eigenvalues which are all one. Such a system has the discrete state equation,

$$\mathbf{x}_i(k+1) = \mathbf{I}_r\mathbf{x}_i(k) + \mathbf{e}(k) = \mathbf{x}_i(k) + \mathbf{e}(k), \tag{4.102}$$

where the system matrix is the $r$-dimensional identity matrix.

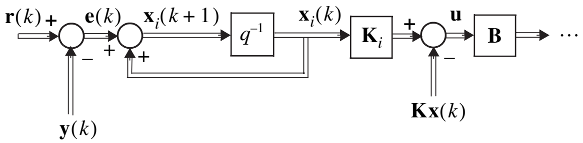


**Fig. 4.20** Discrete time state feedback with integration

The discrete time integrator can be inserted as shown on Fig. 4.20. The set of equations for the overall system is now

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{Fx}(k) + \mathbf{Gu}(k) + \mathbf{G}_v\mathbf{v}(k),\\ \mathbf{u}(k) &= -\mathbf{Kx}(k) + \mathbf{K}_i\mathbf{x}_i(k),\\ \mathbf{x}_i(k+1) &= \mathbf{x}_i(k) - \mathbf{Cx}(k) + \mathbf{r}(k),\\ \mathbf{y}(k) &= \mathbf{Cx}(k).\end{aligned} \tag{4.103}$$

Defining again the augmented state vector,

$$\mathbf{x}_a(k) = \begin{bmatrix}\mathbf{x}(k)\\ \mathbf{x}_i(k)\end{bmatrix}, \tag{4.104}$$

(4.103) can be written

$$\begin{aligned}\begin{bmatrix}\mathbf{x}(k+1)\\ \mathbf{x}_i(k+1)\end{bmatrix} &= \begin{bmatrix}\mathbf{F} & \mathbf{0}\\ -\mathbf{C} & \mathbf{I}_r\end{bmatrix}\begin{bmatrix}\mathbf{x}(k)\\ \mathbf{x}_i(k)\end{bmatrix} + \begin{bmatrix}\mathbf{G}\\ \mathbf{0}\end{bmatrix}\mathbf{u}(k) + \begin{bmatrix}\mathbf{0}\\ \mathbf{I}_r\end{bmatrix}\mathbf{r}(k) + \begin{bmatrix}\mathbf{G}_v\\ \mathbf{0}\end{bmatrix}\mathbf{v}(k),\\ \mathbf{y}(k) &= [\,\mathbf{C} \quad \mathbf{0}\,]\begin{bmatrix}\mathbf{x}(k)\\ \mathbf{x}_i(k)\end{bmatrix},\end{aligned} \tag{4.105}$$

or

$$\begin{aligned}\mathbf{x}_a(k+1) &= \mathbf{F}_1\mathbf{x}_a(k) + \mathbf{G}_1\mathbf{u}(k) + \mathbf{G}_r\mathbf{r}(k) + \mathbf{G}_{v1}\mathbf{v}(k),\\ \mathbf{y}(k) &= \mathbf{C}_1\mathbf{x}_a(k),\end{aligned} \tag{4.106}$$

where

$$\mathbf{F}_1 = \begin{bmatrix}\mathbf{F} & \mathbf{0}\\ -\mathbf{C} & \mathbf{I}_r\end{bmatrix},\ \mathbf{G}_1 = \begin{bmatrix}\mathbf{G}\\ \mathbf{0}\end{bmatrix},\ \mathbf{G}_r = \begin{bmatrix}\mathbf{0}\\ \mathbf{I}_r\end{bmatrix},\ \mathbf{G}_{v1} = \begin{bmatrix}\mathbf{G}_v\\ \mathbf{0}\end{bmatrix},\ \mathbf{C}_1 = [\,\mathbf{C} \quad \mathbf{0}\,]. \tag{4.107}$$

Inserting the feedback law

$$\mathbf{u}(k) = -[\,\mathbf{K} \quad -\mathbf{K}_i\,]\begin{bmatrix}\mathbf{x}(k)\\ \mathbf{x}_i(k)\end{bmatrix} = -\mathbf{K}_1\mathbf{x}_a(k) \tag{4.108}$$

where

$$\mathbf{K}_1 = [\mathbf{K} \quad -\mathbf{K}_i] \tag{4.109}$$

gives finally,

$$\mathbf{x}_a(k+1) = (\mathbf{F}_1 - \mathbf{G}_1\mathbf{K}_1)\mathbf{x}_a(k) + \mathbf{G}_r\mathbf{r}(k) + \mathbf{G}_{v1}\mathbf{v}(k), \tag{4.110}$$

which is quite similar to Eq. (4.99).

The result of the investigation in this section is that addition of integrators does not give rise to new problems. By augmenting the state vectors and the matrices, the integral control case can be formulated as a standard state feedback control problem. This problem can then be solved in the standard way using eigenvalue and/or eigenstructure assignment. This also means that the necessary and sufficient condition for full eigenvalue placement is that the matrix-pairs ($\mathbf{A}_1$, $\mathbf{B}_1$) or ($\mathbf{F}_1$, $\mathbf{G}_1$) are controllable.

It is known from classical control that insertion of integral control (usually in the shape of a PI or a PID controller) may give dynamic problems caused by the negative phase shift of the integrator. Problems of this sort do not arise when using state feedback. As long as the augmented system is controllable, the closed loop system dynamics can be designed as desired in spite of the presence of the single (or multiple) integrator(s).

***Example 4.12*. Integral Control of a Third Order System**

In this example the third order system from Examples 4.5 and 4.6 will again be considered. The continuous system is augmented with an extra integrator for the state $x_i$,

$$\dot{\mathbf{x}}_a = \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{x}_i \end{bmatrix} = \begin{bmatrix} -0.14 & 0.33 & -0.33 & 0 \\ 0.1 & -0.28 & 0 & 0 \\ 0 & 1.7 & -0.77 & 0 \\ -2 & 0 & 0 & 0 \end{bmatrix} \mathbf{x}_a + \begin{bmatrix} 0 \\ 0 \\ -0.025 \\ 0 \end{bmatrix} u + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} r,$$

$$y = [2 \quad 0 \quad 0 \quad 0]\mathbf{x}_a.$$

The augmented system is controllable. The gain matrix will be found using Ackermann's formula and the four eigenvalues are selected as in Example 4.5,

$$\lambda_{cl} = \begin{cases} -0.67 \\ -0.67 \\ -0.67 \\ -0.67 \end{cases}.$$

Using MATLAB's Control System Toolbox and invoking the Ackermann design function by the command,

$$\texttt{K1} = \texttt{acker(A1}, \texttt{B1}, \texttt{[-0.67 -0.67 -0.67 -0.67])},$$

a gain matrix can be obtained which is

$$\mathbf{K}_1 = [\mathbf{K} \quad -k_i] = [210.7 \quad -88.95 \quad -59.6 \quad -43.62].$$

Carrying out a simulation based on a model with the block diagram on Fig. 4.19 with $r = 0.1$ and with $\mathbf{B}_v = \mathbf{0}$, the output $y(t)$ and the input $u(t)$ can be determined just as in Example 4.5. The responses are plotted on Fig. 4.21. The desired final value is now achieved exactly for $t \rightarrow \infty$. The control signal is somewhat larger than in Example 4.5 and the reason is that the gains, especially the first entry of $\mathbf{K}_1$, are quite large. As a matter of fact, one may be demanding too much of this system. It requires a large control effort to achieve the eigenvalues selected here. If they were moved closer to the origin, the gains would decrease and consequently the settling time would increase. Figure 4.22 shows the four states,

$$\mathbf{x}_a = \begin{bmatrix} x_{a1} \\ x_{a2} \\ x_{a3} \\ x_{a4} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_i \end{bmatrix},$$

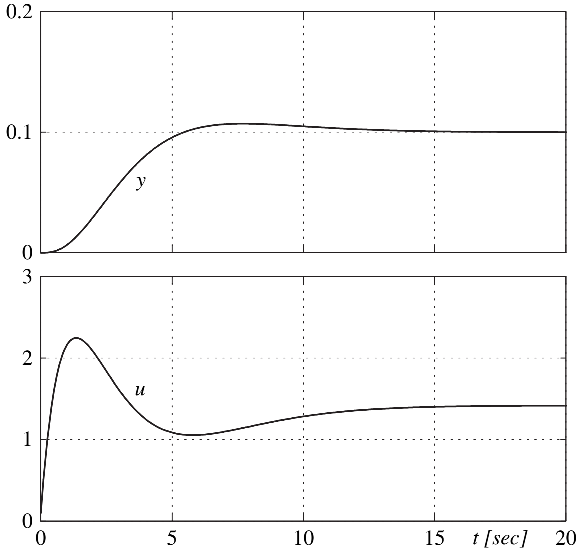


**Fig. 4.21** Responses for third order system with integrator

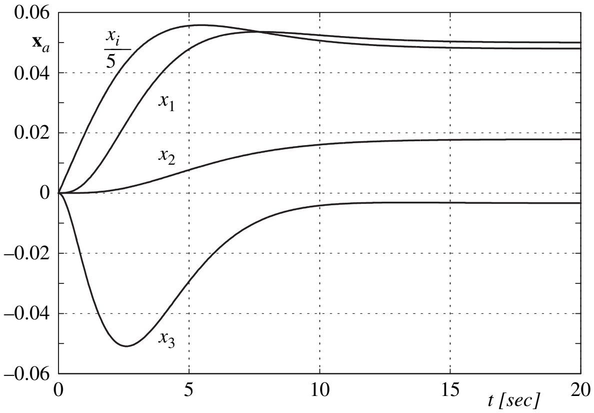


**Fig. 4.22** States for third order system with integrator

during the settling of the system. Note that the plot shows $x_i/5$.

For the discrete time case the matrices from Example 4.6 will be used. The state vector is augmented to find

$$\mathbf{x}_a(k+1) = \begin{bmatrix} 0.9730 & 0.05293 & -0.06031 & 0 \\ 0.01918 & 0.9461 & -0.00061 & 0 \\ 0.003142 & 0.3063 & 0.8572 & 0 \\ -2 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}_a(k)$$

$$+ \begin{bmatrix} 1.554 \cdot 10^{-4} \\ 1.0369 \cdot 10^{-6} \\ -4.634 \cdot 10^{-3} \\ 0 \end{bmatrix} u(k) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} r(k).$$

Using Ackermann's formula again, one can find $\mathbf{K}_1$ for the desired eigenvalue placement. If the continuous time eigenvalues above are translated into the discrete time values they are the same as in Example 4.6.

In order to reduce the gains, one can alternatively compute the gain matrix for a discrete eigenvalue set at a greater distance from the origin. One finds

Case 1

$$\lambda_{cl} = \begin{cases} -0.67 \\ -0.67 \\ -0.67 \\ -0.67 \end{cases} \Rightarrow \lambda_{cld} = \begin{cases} 0.8746 \\ 0.8746 \\ 0.8746 \\ 0.8746 \end{cases} \Rightarrow \mathbf{K}_1 = [193.78\ -77.55\ -53.49\ -7.529],$$

Case 2

$$\lambda_{cl} = \begin{cases} -0.35 \\ -0.35 \\ -0.35 \\ -0.35 \end{cases} \Rightarrow \lambda_{cld} = \begin{cases} 0.9324 \\ 0.9324 \\ 0.9324 \\ 0.9324 \end{cases} \Rightarrow \mathbf{K}_1 = [37.99 \ -38.36 \ -8.809 \ -0.6358].$$

The 'slower' eigenvalues result in gain values which are up to a factor 12 smaller than in the 'fast' case. This could be important in preventing noise problems.

The response to a step of height 0.1 is seen on Fig. 4.23 and 4.24. In both cases it can be seen that the integrator fulfills its required function. The stationary value of the output is exactly 0.1.

Case 1 resembles the continuous time system on Fig. 4.21 as would be expected. Case 2 has a longer settling time but it has no overshoot and the control signal does not have the peak observed in Case 1. ❐

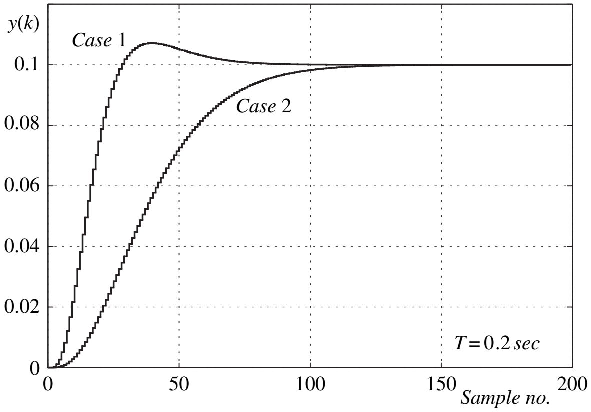


**Fig. 4.23** Responses for discrete time systems with integrator

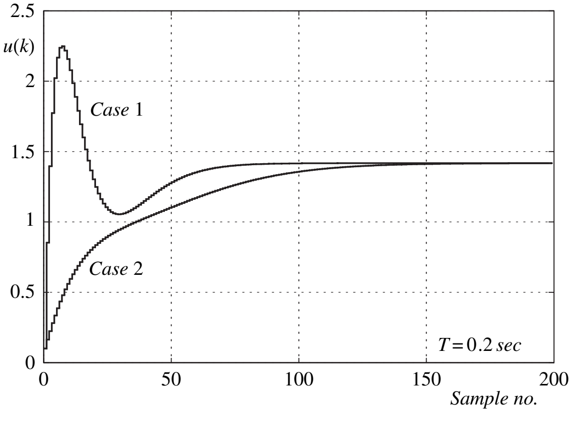


**Fig. 4.24** Control signals for discrete time systems with integrator

***Example 4.13.*** **Water Tank Process Control with Multiple Integrators**

The tank system in Example 2.9 has the state vector, input vector and disturbance vectors,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} H_1 \\ H_2 \\ T_1 \\ T_2 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} A_\nu \\ T_w \\ T_c \end{bmatrix}.$$

The inputs to the system are the two control valve voltages.

The system was linearized at the stationary operating point given by

$$\mathbf{x}_0 = \begin{bmatrix} 2.03 \\ 1.519 \\ 45 \\ 45 \end{bmatrix}, \quad \mathbf{u}_0 = \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \quad \mathbf{v}_0 = \begin{bmatrix} 0.0122 \\ 60 \\ 30 \end{bmatrix},$$

and the dynamic, input, disturbance and output matrices found were:

$$\mathbf{A} = \begin{bmatrix} -0.0499 & 0.0499 & 0 & 0 \\ 0.0499 & -0.0667 & 0 & 0 \\ 0 & 0 & -0.0251 & 0 \\ 0 & 0 & 0.0335 & -0.0335 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0.00510 & 0.00510 \\ 0 & 0 \\ 0.0377 & -0.0377 \\ 0 & 0 \end{bmatrix},$$

$$\mathbf{B}_v = \begin{bmatrix} 0 & 0 & 0 \\ -4.177 & 0 & 0 \\ 0 & 0.01255 & 0.01255 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}.$$

A state controller with integration is to be designed for the system outputs. A simplified block diagram of the entire system with the controller is shown on Fig. 4.25 (Compare this with Fig. 2.19). The block named *Tank system* contains the nonlinear set of differential equations governing the system. So the figure illustrates a model of the *nonlinear* system with a controller which will be designed on the basis of the *linear* model. As is the case with most controllers designed from a linear model, there is an inconsistency between the controller and the system it is meant to control. There will only be complete agreement between the controller and the nonlinear system model if the system is operating exactly at the linearization point given by the stationary state, input and disturbance given above.

It is therefore extremely important that the overall system's behaviour is investigated carefully under operating conditions different than the linearization point. For this reason it should be clear that access to reliable simulation software is just as important as analysis and design software.

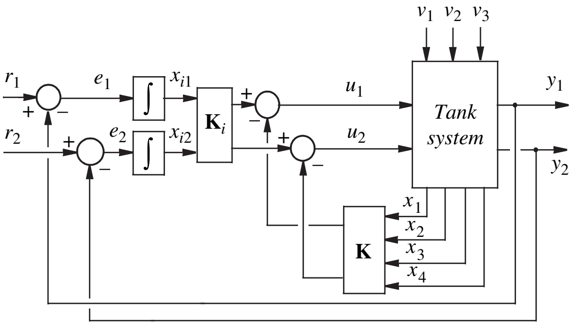


**Fig. 4.25** Tank system with integral controller

The system is sixth order and thus six eigenvalues must be selected. A natural frequency for the eigenvalues of 0.1 rad/s is desired with a damping ratio of at least 0.5. In accordance with these demands the closed loop eigenvalues will be selected to be

$$\lambda_{cl} = \begin{cases} -0.095 \pm j0.02 \\ -0.08 \pm j0.06 \\ -0.05 \pm j0.085 \end{cases} .$$

These eigenvalues are shown on Fig. 4.26: they lie approximately on a semicircle. Using the MATLAB **`place`** function gives the gain matrix

$$\mathbf{K}_1 = [\mathbf{K} \quad -\mathbf{K}_i] = \left[\begin{array}{cccc:cc} 2.642 & 5.308 & 2.466 & 7.159 & 0.1975 & -4.242 \\ 19.46 & 46.39 & -1.838 & -4.028 & -1.77 & 1.919 \end{array}\right].$$

Figure 4.27 shows 1000 s of a simulation of the closed loop system. The initial values are the stationary linearization values. At $t = 100$ sec a step occurs on the reference value $r_1$ requiring the system's level $H_2$ to change from the initial value 1.519 to 1.6 m. At $t = 500$ s a step on $r_2$ orders the temperature $T_2$ to rise it from 45 to 48°C.

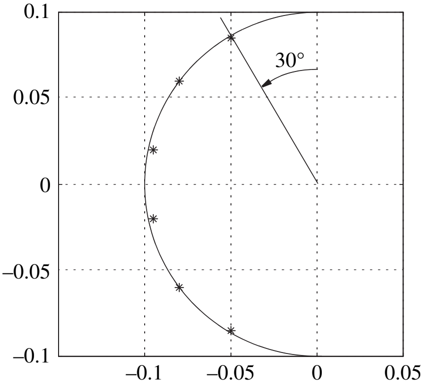


**Fig. 4.26** Eigenvalue placement for the overall system

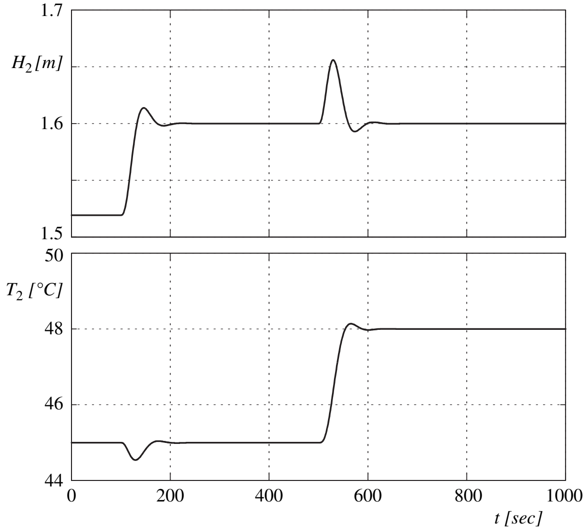


**Fig. 4.27** Step responses for the outputs

The stationary values are achieved exactly, before as well as after the changes, because of the integrators. The dynamic responses are also satisfactory but it is seen that there is dynamic cross-coupling between the two outputs. When the level changes a transient is seen on the temperature and vice versa.

A plot of the two control signals is seen on Fig. 4.28. If it is assumed that the maximum span of the control signals is 0–10 volts, it is noted that both signals keep within this limit.

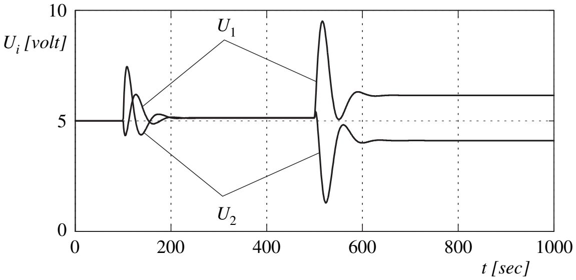


**Fig. 4.28** Control signals

To see the effect of changes in the disturbance variables, another simulation is carried out where the disturbances are changed as follows,

$$A_v : 0.0122\,\mathrm{m}^2 \rightarrow 0.0134\,\mathrm{m}^2 \,\mathrm{at}\ t = 100\,s,$$

$$T_w : 60\,^\circ\mathrm{C} \rightarrow 95\,^\circ\mathrm{C}\,\mathrm{at}\,t = 400\,s,$$

$$T_c : 30\,^\circ\mathrm{C} \rightarrow 10\,^\circ\mathrm{C}\,\mathrm{at}\,t = 700\,s.$$

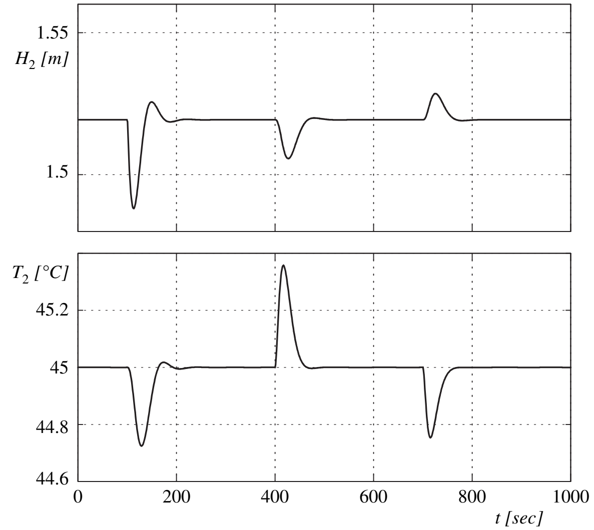


**Fig. 4.29** Outputs for disturbance changes

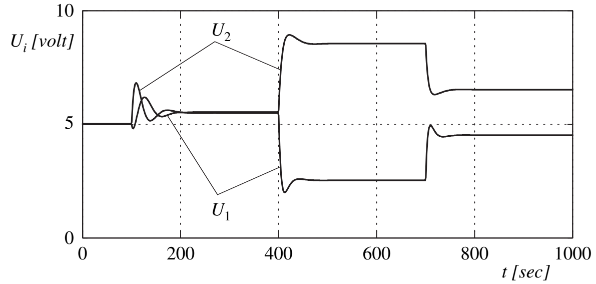


**Fig. 4.30** Control signals for disturbance changes

The responses of the outputs and the control signals are seen on Fig. 4.29 and 4.30. The desired stationary values are held perfectly in all cases but dynamic errors do occur after the changes in the disturbance variables. The errors are well damped. In the close vicinity of the linearization point this is of course in accordance with the positions of the eigenvalues but it is not known in advance how the system would behave if the system is moved away from this operating condition. As seen on the plots, there is no indication of any deterioration of the system properties even for large changes in the disturbances. ❐

### *Example 4.14.* Two Link Robot Arm Control with Multiple Integrators

Here a discrete time controller with integration will be designed for the two-link robot from Example 2.10. The design will be based on the model linearized around the state given by the two link angles: $\theta_1 = 45°$ and $\theta_2 = -30°$, see p. 48.

Prior to the controller design the model must be discretized with the sample period desired for the controller. The robot is a relatively fast mechanical system, so $T = 0.02$ s will be used. This means that it will be possible to handle frequencies up to about 5 Hz (one tenth of the sample frequency) with reasonable accuracy. If further investigations show that this is not adequate, the design must be revised with a different sample period selection.

The discrete time matrices become (using MATLAB's `c2d` function),

$$\mathbf{F} = \begin{bmatrix} 1003.6 & 20.024 & -0.60095 & -0.0040071 \\ 357.19 & 1003.6 & -60.152 & -0.60118 \\ -6.0182 & -0.040106 & 1002.1 & 20.014 \\ -602.38 & -6.0182 & 209.36 & 1002.1 \end{bmatrix} \cdot 10^{-3},$$

$$\mathbf{G} = \begin{bmatrix} 0.25048 & -0.48716 \\ 25.068 & -48.758 \\ -0.48716 & 1.3112 \\ -48.758 & 131.21 \end{bmatrix} \cdot 10^{-3}.$$

The eigenstructure design is characterized by the position of the eigenvalues. For good damping, $\zeta \geq 0.7$ is selected, and for fast response, a natural frequency in the neighborhood of 10 rad/s. Consequently the eigenvalues are

$$\lambda_{cl} = \begin{cases} -7 \pm j7 \\ -8.6 \pm j5 \\ -9.7 \pm j2.6 \end{cases},$$

and the discrete eigenvalues become,

$$\lambda_{cld} = e^{0.02\lambda_{cl}} = \begin{cases} 0.86085 \pm j0.12131 \\ 0.83777 \pm j0.084058 \\ 0.82254 \pm j0.042811 \end{cases}.$$

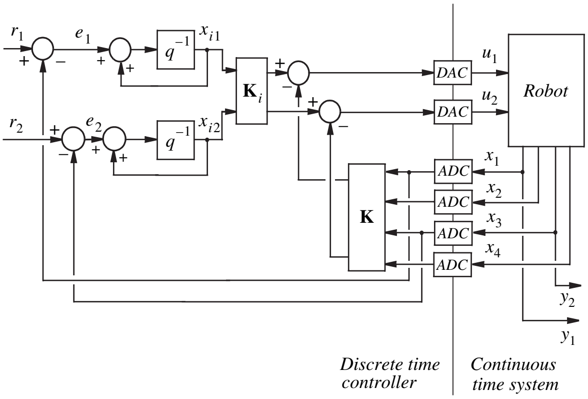
The MATLAB **place** function gives the gain matrix

$$\mathbf{K}_1 = [\mathbf{K} - \mathbf{K}_i] = \left[\begin{array}{cccc|cc} 841.3 & 76.385 & 153.99 & 20.175 & -63.628 & -8.1492 \\ 343.88 & 30.304 & 90.155 & 10.891 & -27.411 & -5.175 \end{array}\right].$$

The gains are to be realized as constants in the control computer. This is usually easier than realizing the gains using analog operational amplifiers (as will usually be the case for a continuous time controller). It should be kept in mind though that large gains may cause noise sensitivity, especially in the discrete case.

The overall system with controller is shown on Fig. 4.31. Note that the outputs $y_1$ and $y_2$ are equal to the states $x_1$ and $x_3$ respectively. Note also that the figure is divided into two parts. The continuous time nonlinear robot system to the right of the thin vertical line and the discrete time controller to the left of the line. When passing this interface, the continuous ('analog') signals are converted to discrete ('digital') signals by the *analog-to-digital converters*, ADC. The discrete signals are converted to continuous signals by the *digital-to-analog converters*, DAC.

**Fig. 4.31** Discrete time controller for a two-link robot



All the operations to the left of the vertical line take place in a *real time computer*.

The results of a simulation carried out on the system of Fig. 4.31 are shown on the two Figs. 4.32 and 4.33. The initial state is the same as the linearization point of Example 2.10:

$$\mathbf{x}_0 = \begin{bmatrix} 45^\circ \\ 0 \\ -30^\circ \\ 0 \end{bmatrix} = \begin{bmatrix} \pi/4 \\ 0 \\ -\pi/6 \\ 0 \end{bmatrix}.$$

Two steps are imposed upon the system. At $t = 0.1$ s the reference $r_1$ is changed from $\pi/4$ rad to $-\pi/4$ rad and at $t = 3$ s the reference $r_2$ is changed from $-\pi/6$ rad to $\pi/6$ rad. The initial and the final configurations are shown on Fig. 4.34.

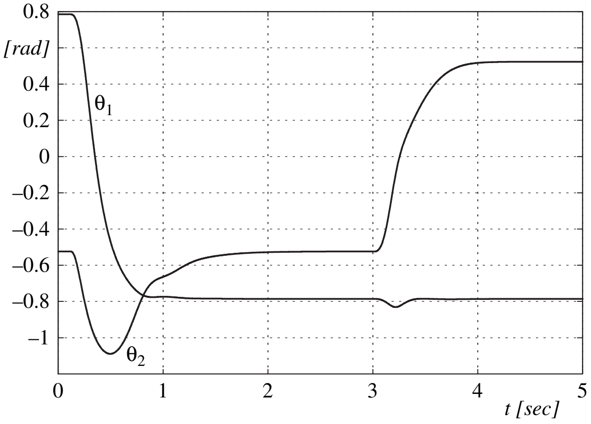


**Fig. 4.32** Robot link angle step responses

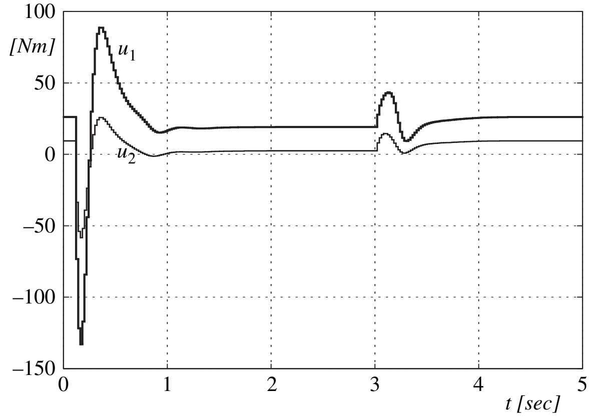


**Fig. 4.33** Control signals (motor torques)

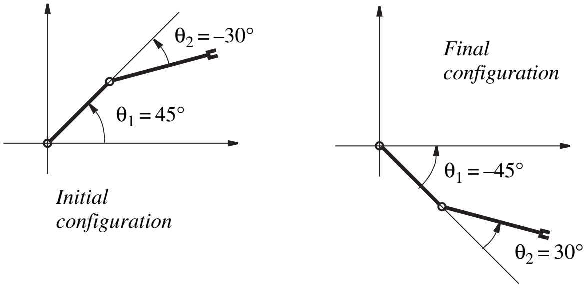


**Fig. 4.34** Link configurations before and after steps

After the transients the new angles are accurately achieved but with some dynamic interaction between the two outputs. The control signals (the motor torques) on Fig. 4.33 are stair case functions because they are outputs from the DAC-converters.

Although the system properties change considerably under this change of configuration (see Example 3.10), there are no signs of performance deterioration in the responses. The controller seems to be robust enough to allow large changes in the operating point. ❒

***Example 4.15.*** **Discrete Control of a Stock Inventory**

A discrete time example of a non-technical problem is the control of the stock inventory of a commodity (for example candy, beer, underpants, etc.) given the requirement that some stock must always be available for sale. Given is that the quantity of a commodity $c(k)$ must be equal to its value on the previous day plus that which is ordered, $o(k)$, on the previous day from a supplier. The quantity sold on any day, $s(k)$, has to be subtracted from the existing stock. The state equations which describes this system are:

$$
\begin{aligned}
c(k+1) &= c(k) + o(k) - s(k),\\
o(k+1) &= u(k),\\
y(k) &= c(k),
\end{aligned}
$$

or in vector-matrix form:

$$
\mathbf{x}(k+1) = \begin{bmatrix} c(k+1) \\ o(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c(k) \\ o(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(k) + \begin{bmatrix} -1 \\ 0 \end{bmatrix} s(k), \qquad (4.111)
$$

$$
y(k) = [\,1 \;\; 0\,]\mathbf{x}(k).
$$

The sampling period is $T = 1$ day.

What is desired in this example is to keep the stationary value of the stock constant independent of arbitrary constant sales. The sales must be considered a disturbance to the system as indicated in Eq. (4.111). This can be accomplished by adding a discrete time integral state as on Fig. 4.20,

$$
x_i(k+1) = r(k) - y(k) + x_i(k) = r(k) - \mathbf{C}\mathbf{x}(k) + x_i(k).
$$

The system is augmented in the usual way,

$$
\begin{bmatrix} c(k+1) \\ o(k+1) \\ x_i(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} c(k) \\ o(k) \\ x_i(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u(k) + \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} s(k),
$$

with the augmented matrices,

$$
\mathbf{F}_1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{G}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.
$$

It is easy to find the controller gain using Ackermann's formula.

Selecting the closed loop eigenvalues,

$$
\lambda_{cl} = \begin{cases} 0.2 \\ 0.2 \\ 0.2 \end{cases},
$$

it is found that

$$
\mathbf{K}_1 = [\,k_1 \;\; k_2 \;\; -k_i\,] = [\,1.92 \;\; 1.4 \;\; -0.512\,].
$$

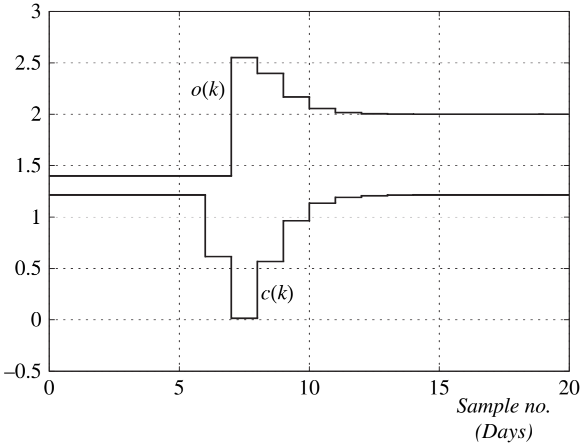


**Fig. 4.35** Disturbance response of stock control system

The daily order can then be calculated from the formula:

$$o(k+1) = u(k) = -\mathbf{Kx}(k) + k_i x_i(k) = -k_1 c(k) - k_2 o(k) + k_i x_i(k).$$

A simulation on the closed loop system (see Eq. (4.110)) gives the result on Fig. 4.35. At the start of the simulation the sales and the ordered inventory are 1.4 per day (measured in some appropriate unit) and the stock is 1.2 which is what is desired. At day no. 5 the sales go up from 1.4 to 2 and consequently the stock has decreased at the morning of day 6. This causes the 'controller' to take action and the ordered inventory increases on day 7. After the transients a new equilibrium is established at day 13, with a new constant daily order equal to the sales with the stock available the same size as before the disturbance. ❐

## 4.6 Deterministic Observers and State Estimation

From previous sections of this chapter it is clear that full state feedback is very useful in obtaining control systems which have good stability and robustness properties. In order to use full state feedback it is necessary that all states of a given system are measured. In practice this is not always possible or desirable. In some cases sensors are simply not available or cannot be made for the states which one would desire to measure. In other cases the required sensor is too expensive for the intended application. It has to be remembered that each measurement requires a sensor and its associated power supply, signal conditioning equipment and measurement connections. This means that each sensor channel represents a significant investment of time and money. Moreover in addition to the noise in the process to be controlled each measurement is also a source of noise which can reduce the accuracy with which a control object can be controlled.

If the process which is to be controlled is observable and a reasonably accurate model is available for it, then it is possible to use a modified model of the process to *estimate* the states of the process which are not measured. The main model modification is to add a term which corrects for modelling error and internal disturbances. Under the proper conditions, the state estimates can then be used instead of direct measurements in a full state feedback system. Such a modified model, used in a feedback system is called a *Luenberger observer* or a *Kalman filter*, depending on how it is designed. Often this is shortened to 'observer' with it being understood from the problem formulation or context which kind of observer or filter is being considered. As will become apparent in what follows, observers also have the very useful property that they can be used to effectively suppress or reduce both process and measurement noise in feedback systems and in some cases, this is their main advantage. The intention of the modifications which are introduced in the system models which form an observer is to cause the state estimates to follow more or less exactly and rapidly the states of the control object independent of internal or external disturbances. Here 'disturbances' can be actual signals as well as modelling errors.

Luenberger observers are modified models of control objects which are designed from deterministic considerations. Kalman filters are in structure identical to Luenberger observers but they are modified models of control objects which are designed from statistical considerations. In this chapter only deterministic (or Luenberger) observers will be considered.

### *4.6.1 Continuous Time Full Order Observers*

In what follows the control object is to be described by the usual system equations:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{4.112}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t). \tag{4.113}$$

If the Eqs. (4.112) and (4.113) were exact and undisturbed by noise, it would be reasonable to use the states estimated with the model of the system alone as the basis of a state feedback system. Such an observer might be called an *open loop observer*, see Kailath (1980). Unfortunately it is rare that a system is known with such accuracy and is noise free. Thus it is necessary to introduce feedback around the system model itself in order to force it to follow the control object accurately. Such a configuration might be called a *closed loop observer*. In control circles only such systems are called observers: open loop observers are most correctly called feed forward systems.

In order to practically estimate the state $\mathbf{x}(t)$ of the linear system described by Eqs. (4.112) and (4.113) a system of differential equations can be constructed which has the form,

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{M}\hat{\mathbf{x}}(t) + \mathbf{N}\mathbf{u}(t) + \mathbf{L}\mathbf{y}(t), \tag{4.114}$$

where $\hat{\mathbf{x}}(t)$ is estimate of the state variable $\mathbf{x}(t)$. Equation (4.114) is the state equation of an observer for (4.112) if

$$\hat{\mathbf{x}}(t_0) = \mathbf{x}(t_0) \text{ implies } \hat{\mathbf{x}}(t) = \mathbf{x}(t) \text{ for all } t \geq t_0 \text{ and for all } \mathbf{u}(t). \tag{4.115}$$

As all of the states of the system are estimated, the observer is a *full order observer*.

Equation (4.114) is now subtracted from (4.112), (4.113) is inserted and the following equation is obtained

$$\dot{\mathbf{x}}(t) - \dot{\hat{\mathbf{x}}}(t) = \dot{\mathbf{e}}_e(t) = (\mathbf{A} - \mathbf{L}\mathbf{C})\mathbf{x}(t) - \mathbf{M}\hat{\mathbf{x}}(t) + (\mathbf{B} - \mathbf{N})\mathbf{u}(t). \tag{4.116}$$

The vector,

$$\mathbf{e}_e(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t), \tag{4.117}$$

is called the *estimation error*. Letting

$$\mathbf{M} = \mathbf{A} - \mathbf{L}\mathbf{C} \text{ and } \mathbf{N} = \mathbf{B} \tag{4.118}$$

Equation (4.116) becomes

$$\dot{\mathbf{e}}_e(t) = (\mathbf{A} - \mathbf{L}\mathbf{C})\mathbf{e}_e(t). \tag{4.119}$$

If the system in Eq. (4.119) is asymptotically stable, i.e., if the eigenvalues of the system matrix $\mathbf{A} - \mathbf{L}\mathbf{C}$ are all in the left half plane, then the condition (4.115) is fulfilled and (4.114) is an observer or a state estimator for (4.112).

Even if the estimator is 'started' in an initial state such that $\hat{\mathbf{x}}(t_0) \neq \mathbf{x}(t_0)$, the asymptotic stability will, according to Eq. (4.119) ensure that

$$\hat{\mathbf{x}}(t) \rightarrow \mathbf{x}(t) \text{ for } \rightarrow \infty.$$

For this reason (4.114) is called an *asymptotic state estimator* for (4.112).

If the estimated output is defined as

$$\hat{\mathbf{y}}(t) = \mathbf{C}\hat{\mathbf{x}}(t) \tag{4.120}$$

then (4.114) can be written with insertion of (4.118) as

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{A}\hat{\mathbf{x}}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{L}(\mathbf{y}(t) - \hat{\mathbf{y}}(t)). \tag{4.121}$$

The gain matrix **L** is called the *observer gain matrix* and is to be selected by the designer.

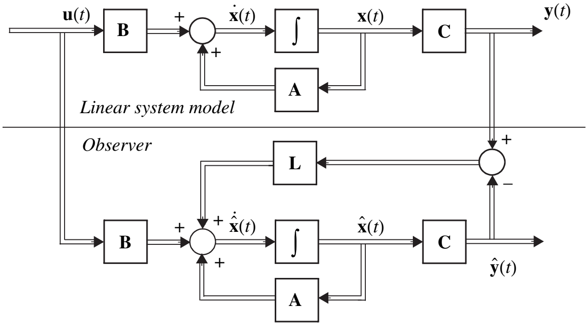


**Fig. 4.36** Full order continuous time observer

Figure 4.36 shows a block diagram of a full order observer. It is implicit in the block diagram that the system to be controlled and the observer have exactly the same **A**, **B** and **C** matrices. Both the control object and the observer are supplied with the same input. The outputs of the control object and observer model are subtracted from each other and this output difference is multiplied by the observer gain **L** and inserted into the summing point in front of the observer model. This difference is called a *residual* or an *innovation*. Figure 4.37 shows an alternative block diagram of the observer.

Clearly the properties of the observer in Eq. (4.121) are strongly dependent on the matrix **L**. The most important point here is that the Eq. (4.119) describing the estimation error is stable and sufficiently fast to ensure that initial estimation errors will vanish rapidly. These qualities are reflected in the eigenvalues of the observer system matrix,

$$\mathbf{A_L} = \mathbf{A} - \mathbf{LC}, \tag{4.122}$$

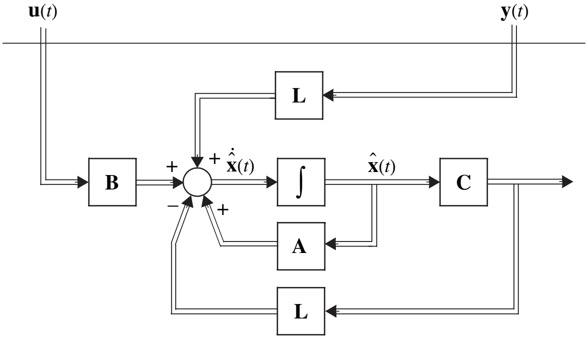


**Fig. 4.37** Alternative block diagram of the observer of Eq. (4.121)

which is sometimes called the *stability matrix* of the observer (4.121). The eigenvalues of the stability matrix are the solutions of the equation:

$$det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{LC}) = 0. \tag{4.123}$$

## *4.6.2 Discrete Time Full Order Observers*

In the discrete time case the methods used to analyze and describe the properties of observers are nearly the same as above for continuous systems. The observers which can be constructed may however have somewhat different properties. This is in large measure due to the fact that the pure time delay is an inherent dynamic element in discrete time systems. This means that no input can influence a control object or observer between sampling times and no discrete measurement can resolve what a control object is doing between sampling times.

For the discrete time system,

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{Fx}(k) + \mathbf{Gu}(k), \\ \mathbf{y}(k) &= \mathbf{Cx}(k), \end{aligned} \tag{4.124}$$

the discrete observer equivalent to (4.121) will be

$$\hat{\mathbf{x}}(k+1) = \mathbf{F}\hat{\mathbf{x}}(k) + \mathbf{Gu}(k) + \mathbf{L}(\mathbf{y}(k) - \hat{\mathbf{y}}(k)). \tag{4.125}$$

Defining the estimation error,

$$\mathbf{e}_e(k) = \mathbf{x}(k) - \hat{\mathbf{x}}(k), \tag{4.126}$$

and subtracting (4.125) from (4.124) gives the equation governing the error,

$$\mathbf{e}_e(k+1) = (\mathbf{F} - \mathbf{LC})\mathbf{e}_e(k), \tag{4.127}$$

and it is not surprising to find that the stability matrix is

$$\mathbf{F_L} = \mathbf{F} - \mathbf{LC}. \tag{4.128}$$

The observer (4.125) calculates the state estimate at time $k+1$ as soon the measurement $\mathbf{y}$ is taken at time $k$. In other words, the observer *predicts* what the state will be one sample period forward. Therefore (4.125) is called a *predictive observer*. Sometimes it is argued, that the observer (4.125) does not lead to the best possible result since is does not benefit from the *last* measurement, i.e., the measurement $\mathbf{y}(k+1)$.

It is of course not possible for any real time computer to make a measurement, to calculate the state estimate and to calculate the control signal instantaneously. However, if the plant to be controlled is much slower than the computer, a computation delay may be immaterial and in such cases the observer above can be improved.

To accomplish this an intermediate state, $\bar{\mathbf{x}}(k)$, is introduced and the state estimation calculation is split into two steps. During the sample period from time $k$ to time $k+1$ the *time update* is calculated,

$$\bar{\mathbf{x}}(k+1) = \mathbf{F}\hat{\mathbf{x}}(k) + \mathbf{G}\mathbf{u}(k), \tag{4.129}$$

and at time $k+1$ the system output $\mathbf{y}(k+1)$ is measured and the *measurement update* is calculated ('instantaneously'), i.e., the actual state estimate,

$$\hat{\mathbf{x}}(k+1) = \mathbf{F}\hat{\mathbf{x}}(k) + \mathbf{G}\mathbf{u}(k) + \mathbf{L}(\mathbf{y}(k+1) - \mathbf{C}\bar{\mathbf{x}}(k+1)). \tag{4.130}$$

Since the *current* value of the output is used, the observer (4.129) and (4.130) is called a *current observer*.

If Eq. (4.129) is inserted into (4.130), it can be rewritten as

$$\hat{\mathbf{x}}(k+1) = (\mathbf{F} - \mathbf{LCF})\hat{\mathbf{x}}(k) + \mathbf{G}\mathbf{u}(k) - \mathbf{LCG}\mathbf{u}(k) + \mathbf{LC}\mathbf{x}(k+1). \tag{4.131}$$

Using the state Eq. (4.124) leads to

$$\hat{\mathbf{x}}(k+1) = (\mathbf{F} - \mathbf{LCF})\hat{\mathbf{x}}(k) + \mathbf{G}\mathbf{u}(k) + \mathbf{LCF}\mathbf{x}(k). \tag{4.132}$$

Subtracting this from (4.124) gives

$$\mathbf{x}(k+1) - \hat{\mathbf{x}}(k+1) = \mathbf{x}(k) - (\mathbf{F} - \mathbf{LCF})\hat{\mathbf{x}}(k) - \mathbf{LCF}\mathbf{x}(k) \tag{4.133}$$

or, introducing again the estimation error (4.126),

$$\mathbf{x}(k+1) - \hat{\mathbf{x}}(k+1) = \mathbf{e}_e(k+1) = (\mathbf{F} - \mathbf{LCF})\mathbf{e}_e(k). \tag{4.134}$$

The stability matrix in (4.134) resembles that of Eq. (4.127). The only difference is that the output matrix $\mathbf{C}$ is replaced by the product $\mathbf{CF}$, which has the same dimension as $\mathbf{C}$.

## 4.7 Observer Design for SISO Systems

### *4.7.1 Observer Design Based on the Observer Canonical Form*

Any observable SISO system can be brought into observer canonical form by a similarity transformation as shown in Sect. 3.9.2. The state transformation is carried out by the expression $\mathbf{z} = \mathbf{Q}\mathbf{x}$ where $\mathbf{Q}$ is found from Eq. (3.368).

The observer canonical form is:

$$\begin{aligned}\dot{\mathbf{z}} &= \mathbf{A}_{oc}\mathbf{z} + \mathbf{B}_{oc}u,\\ y &= \mathbf{C}_{oc}\mathbf{z},\end{aligned} \tag{4.135}$$

where

$$\mathbf{A}_{oc} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & 0 & \dots & 0 & -a_1 \\ 0 & 1 & 0 & \dots & 0 & -a_2 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & -a_{n-2} \\ 0 & 0 & 0 & \dots & 1 & -a_{n-1} \end{bmatrix}, \ \mathbf{B}_{oc} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{bmatrix}, \tag{4.136}$$

$$\mathbf{C}_{oc} = [\,0\ 0\ \dots\ 0\ 1\,]$$

The observer for this system will be,

$$\dot{\hat{\mathbf{z}}}(t) = \mathbf{A}_{oc}\hat{\mathbf{z}}(t) + \mathbf{B}_{oc}u(t) + \mathbf{L}_{oc}(y(t) - \hat{y}(t)), \tag{4.137}$$

with the gain matrix,

$$\mathbf{L}_{oc} = \begin{bmatrix} l'_1 \\ l'_2 \\ \vdots \\ l'_n \end{bmatrix}, \tag{4.138}$$

and the stability matrix,

$$\mathbf{A}_{\mathbf{L}_{oc}} = \mathbf{A}_{oc} - \mathbf{L}_{oc}\mathbf{C}_{oc}. \tag{4.139}$$

Inserting these matrices into (4.139) gives

$$\mathbf{A}_{\mathbf{L}_{oc}} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & -a_0 - l'_1 \\ 1 & 0 & 0 & \dots & 0 & -a_1 - l'_2 \\ 0 & 1 & 0 & \dots & 0 & -a_2 - l'_3 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & -a_{n-2} - l'_{n-1} \\ 0 & 0 & 0 & \dots & 1 & -a_{n-1} - l'_n \end{bmatrix}. \tag{4.140}$$

The stability matrix is also in observer canonical form, so the last column contains the coefficients of the characteristic polynomial of the stability matrix.

If it is desired that the eigenvalues of the observer be placed in the specified positions in the complex plane,

$$\lambda_0 = \lambda_{01}, \lambda_{02}, \ldots, \lambda_{0n}, \tag{4.141}$$

then the stability matrix characteristic polynomial can be written,

$$P_{ch,\,\mathbf{A}_{\mathbf{L}_{oc}}} = \prod_{i=1}^{n} (\lambda - \lambda_{oi}) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \ldots + \alpha_1\lambda + \alpha_0. \tag{4.142}$$

Comparing (4.140) and (4.142) leads to a set of equations for determination of the gains:

$$\begin{aligned} \alpha_0 &= a_0 + l'_1, \\ \alpha_1 &= a_1 + l'_2, \\ &\vdots \\ \alpha_{n-1} &= a_{n-1} + l'_n, \end{aligned} \tag{4.143}$$

or

$$\begin{aligned} l'_1 &= \alpha_0 - a_0, \\ l'_2 &= \alpha_1 - a_1, \\ &\vdots \\ l'_n &= \alpha_{n-1} - a_{n-1}. \end{aligned} \tag{4.144}$$

For the original system the stability matrix is $\mathbf{A_L} = \mathbf{A} - \mathbf{LC}$. The transformation means that,

$$\begin{aligned} \mathbf{A_L} = \mathbf{A} - \mathbf{LC} &= \mathbf{Q}^1\mathbf{A}_{oc}\mathbf{Q} - \mathbf{LC}_{oc}\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{A}_{oc}\mathbf{Q} - \mathbf{Q}^{-1}\mathbf{QLC}_{oc}\mathbf{Q} \\ &= \mathbf{Q}^{-1}(\mathbf{A}_{oc} - \mathbf{QLC}_{oc})\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{AL}_{oc}\mathbf{Q} = \mathbf{Q}^{-1}(\mathbf{A}_{oc} - \mathbf{L}_{oc}\mathbf{C}_{oc})\mathbf{Q}, \end{aligned} \tag{4.145}$$

which shows that

$$\mathbf{L} = \mathbf{Q}^{-1}\mathbf{L}_{oc}. \tag{4.146}$$

The stability matrix of the predictive discrete time observer has the same form as in the continuous case so this design method will also work for the predictive observer.

## 4.7.2 Ackermann's Formula for the Observer

The problems of designing a state feedback controller and a full order observer are very much alike. The controller closed loop system matrix and the observer stability matrix are

$$\begin{aligned}\mathbf{A_K} &= \mathbf{A} - \mathbf{BK},\\ \mathbf{A_L} &= \mathbf{A} - \mathbf{LC}.\end{aligned} \tag{4.147}$$

Looking at the dual systems introduced in Sect. 3.8.9 and specializing to the SISO case,

$$S_x : \begin{cases} \dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{B}u \\ y = \mathbf{Cx} \end{cases} \tag{4.148}$$

and

$$S_z : \begin{cases} \dot{\mathbf{z}} = \mathbf{A}^T\mathbf{z} + \mathbf{C}^T u. \\ y = \mathbf{B}^T\mathbf{z} \end{cases} \tag{4.149}$$

Designing an observer for the system $S_x$ leads to the stability matrix above. It's transpose is

$$\mathbf{A}_{\mathbf{L}_x}^T = \mathbf{A}^T - \mathbf{C}^T\mathbf{L}_x^T. \tag{4.150}$$

Design of a controller for system $S_z$ leads to the matrix,

$$\mathbf{A}_{\mathbf{K}_z} = \mathbf{A}^T - \mathbf{C}^T\mathbf{K}_z. \tag{4.151}$$

If $\mathbf{L}_x^T = \mathbf{K}_z$ the two matrices in (4.150) and (4.151) are equal. This shows that designing an observer for the system $S_x$ is precisely the same as designing a controller for the dual system $S_z$.

Formulation of Ackermann's formula for the design of a controller for $S_z$ yields

$$\mathbf{K}_z = [0\ 0\ \dots\ 0\ 1]\mathbf{M}_{c,z}^{-1} P_{ch,\mathbf{A}_{\mathbf{K}_z}}(\mathbf{A}^T), \tag{4.152}$$

where $\mathbf{M}_{c,z}$ is the controllability matrix for $S_z$ and $P_{ch,\mathbf{A}\mathbf{K}_z}(\mathbf{A}^T)$ is the closed loop characteristic polynomial with the system matrix $\mathbf{A}^T$ replacing $\lambda$. Now the observer gain matrix for $S_x$ can be found as

$$\mathbf{L}_x = \mathbf{K}_z^T = P_{ch,\mathbf{A}_{\mathbf{K}_z}}^T(\mathbf{A}^T)(\mathbf{M}_{c,z}^{-1})^T \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{4.153}$$

From Eq. (3.326) it is known that $\mathbf{M}_{c,z}^T = \mathbf{M}_{o,x}$ and therefore also that

$$(\mathbf{M}_{c,z}^{-1})^T = \mathbf{M}_{o,x}^{-1} \tag{4.154}$$

and for the characteristic polynomial one finds that

$$\begin{aligned} P^T_{ch,\mathbf{A}_{\mathbf{K}_z}}(\mathbf{A}^T) &= P^T_{ch,\mathbf{A}_{\mathbf{L}_x^T}}(\mathbf{A}^T) \\ &= ((\mathbf{A}^T)^n + \alpha_{n-1}(\mathbf{A}^T)^{n-1} + \ldots + \alpha_1\mathbf{A}^T + \alpha_0\mathbf{I})^T \\ &= \mathbf{A}^n + \alpha_{n-1}\mathbf{A}^{n-1} + \ldots + \alpha_1\mathbf{A} + \alpha_0\mathbf{I} = P_{ch,\mathbf{A}_{\mathbf{L}_x}}(\mathbf{A}). \end{aligned} \tag{4.155}$$

By these simple manipulations Ackermann's formula for the full order observer for the system of Eq. (4.148) has been derived:

$$\mathbf{L} = P_{ch,\mathbf{A}_\mathbf{L}}(\mathbf{A})\mathbf{M}_o^{-1}\begin{bmatrix} 0 \\ 0 \\ : \\ 0 \\ 1 \end{bmatrix}. \tag{4.156}$$

Ackermann's formula can be applied to continuous as well as discrete time systems. For the predictive observer the formula is used unchanged, with the system matrix **F** replacing **A** and with the usual observability matrix,

$$\mathbf{M}_o = \begin{bmatrix} \mathbf{C} \\ \mathbf{CF} \\ \mathbf{CF}^2 \\ : \\ \mathbf{CF}^{n-1} \end{bmatrix}. \tag{4.157}$$

For the current estimator the output matrix **C** is replaced by the product **CF** (see Eq. (4.134)) and consequently $\mathbf{M}_o$ in Ackermann's formula is replaced by

$$\mathbf{M}_o' = \begin{bmatrix} \mathbf{CF} \\ \mathbf{CF}^2 \\ \mathbf{CF}^3 \\ : \\ \mathbf{CF}^n \end{bmatrix}. \tag{4.158}$$

***Example 4.16*. Observer Control of a Marginally stable System**

Many mechanical and electrical systems (or combinations of them) can be described as un- or underdamped harmonic oscillators. For example hydraulic/mechanical motors, tuned circuits (electrical or mechanical/electrical), etc. The state and output equations of such a system are easy to write down. Choosing the position and the velocity as the states, the equations will be

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u,$$

$$y = [1\ 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

$\omega_n$ is the natural frequency of the system. The eigenvalues are purely imaginary,

$$\lambda = \pm j\omega_n,$$

so the system is Lyapunov stable.

It is assumed that only the position of the body is measured. An observer for this system can be found from Eq. (4.121),

$$\dot{\hat{\mathbf{x}}} = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & 0 \end{bmatrix} \hat{\mathbf{x}} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \mathbf{L}(y - \mathbf{C}\hat{\mathbf{x}}),$$

where **L** is two-dimensional,

$$\mathbf{L} = \begin{bmatrix} l_1 \\ l_2 \end{bmatrix}.$$

In this simple case the gains can be found directly by comparing coefficients. The characteristic polynomial of the stability matrix is

$$\begin{aligned} det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{LC}) &= det\left( \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & 0 \end{bmatrix} + \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} [1\ 0] \right) \\ &= \lambda^2 + l_1\lambda + \omega_n^2 + l_2 = \lambda^2 + 2\zeta_o\omega_{no}\lambda + \omega_{no}^2, \end{aligned}$$

where $\omega_{no}$ is the natural frequency and $\zeta_o$ is the damping ratio of the observer.

In order for the observer to be reasonably fast with respect to the control object, the observer eigenvalues should be on the order of 5–10 times as ‘fast’ as the system itself. This means that one could select $\omega_{no} = 5\omega_n$. The damping of these eigenvalues should be good: $\zeta_o = 0.707$. In other words,

$$l_1 = 2\zeta_o\omega_{no} = 7.07\omega_n$$

$$\omega_n^2 + l_2 = 25\omega_n^2 \Rightarrow l_2 = 24\omega_n^2.$$

It should be noticed here that the control object is a system which is not stable itself. The observer for this system is however stable and fast so that it can follow the control object independent of its nonstable behavior or its initial conditions. There is no contradiction in this: the situation is parallel to the design of a stable controller for a nonstable or unstable control object. ❒

***Example 4.17.* Discrete Observer for a SISO System**

In this example a discrete time predictive observer for the system in Examples 4.5 and 4.6 is to be designed. The system is SISO and therefore Ackermann's Formula (4.156) can be applied.

The continuous system's eigenvalues are

$$\lambda_c = \begin{cases} -0.8986 \\ -0.1457 \pm j0.2157 \end{cases}.$$

Again, it is desired that the observer to be much faster than the system itself, so the following observer eigenvalues are selected:

$$\lambda_{co} = \begin{cases} -5 \\ -3.5 \pm j3.5 \end{cases},$$

which means that the observer will not only be fast, it will also be well damped.The discrete time eigenvalues with the sample period $T = 0.2$ s will be

$$\lambda_{do} = e^{0.2\lambda_{co}} = \begin{cases} 0.3679 \\ 0.3798 \pm j0.3199 \end{cases}.$$

With these eigenvalues the characteristic polynomial for the stability matrix becomes,

$$P_{ch,\,\mathbf{F_L}}(\mathbf{F}) = \mathbf{F}^3 - 1.1275\mathbf{F}^2 + 0.5260\mathbf{F} - 0.09072\mathbf{I},$$

where

$$\mathbf{F} = \begin{bmatrix} 0.9730 & 0.05293 & -0.06031 \\ 0.01918 & 0.9461 & -0.00061 \\ 0.003142 & 0.3063 & 0.8572 \end{bmatrix}.$$

The observability matrix is,

$$\mathbf{M}_o = \begin{bmatrix} \mathbf{C} \\ \mathbf{CF} \\ \mathbf{CF}^2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 1.9459 & 0.10586 & -0.12062 \\ 1.8949 & 0.16621 & -0.22082 \end{bmatrix},$$

and the observer gain matrix is readily computed,

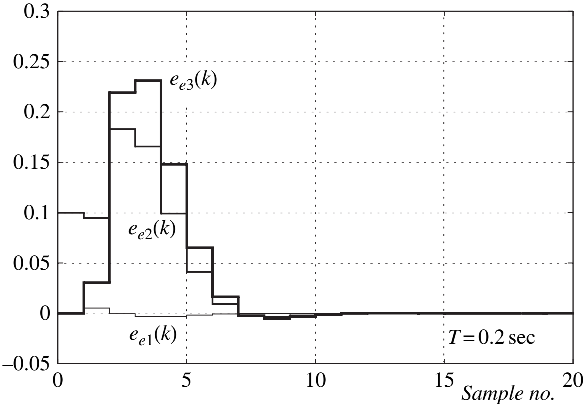
$$\mathbf{L} = \begin{bmatrix} 0.8244 \\ -8.8262 \\ -15.592 \end{bmatrix}.$$

Carrying out a simulation of the estimation error determined by the state Equation (4.127), the results as shown on Fig. 4.38 will appear. In this case the initial error,

$$\mathbf{e}_{e0} = \begin{bmatrix} 0 \\ 0.1 \\ 0 \end{bmatrix},$$

was assumed.

**Fig. 4.38** Estimation error for $\mathbf{e}_{e0} = [0\ 0.1\ 0]^T$



As should be expected, the error tends to zero and it has disappeared in about 11 samples or approx. 2.2 s. ❐

***Example 4.18.*** **Current Observer for the SISO System**

The current observer (4.129) and (4.130) can be designed the same way as in the predictive case. The only difference is that the output matrix **C** must be replaced by the product **CF**.

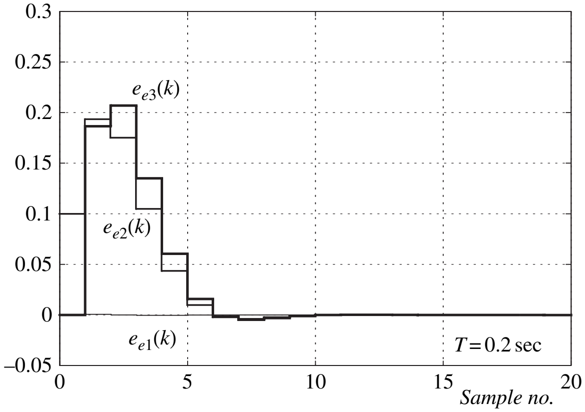
The modified observability matrix becomes

$$\mathbf{M'}_o = \begin{bmatrix} \mathbf{CF} \\ \mathbf{CF}^2 \\ \mathbf{CF}^3 \end{bmatrix} = \begin{bmatrix} 1.9459 & 0.10586 & -0.12062 \\ 1.8949 & 0.16621 & -0.22082 \\ 1.8462 & 0.18991 & -0.30368 \end{bmatrix}.$$

The new observer gain matrix is found to be

$$\mathbf{L} = \begin{bmatrix} 0.44245 \\ -9.3475 \\ -14.735 \end{bmatrix},$$

**Fig. 4.39** Estimation error for $\mathbf{e}_{e0} = [0\ 0.1\ 0]^T$ (current observer)



which is not far from the gain matrix in the previous example. The estimation error is now governed by the state equation:

$$\mathbf{e}_e(k+1) = (\mathbf{F} - \mathbf{LCF})\mathbf{e}_e(k).$$

The solution for the same initial error as in Example 4.17 is found by simulation. See Fig. 4.39. Comparing with Fig. 4.38, a slight decrease in the errors can be noted on Fig. 4.39, but in this case the difference is not very significant. ❐

### 4.7.3 Conditions for Eigenvalue Assignment

The Formula (4.156) show that $\mathbf{L}$ can be found if $\mathbf{M}_o$ is nonsingular, i.e., if the system is observable. It can be concluded that observability is a *sufficient* condition for arbitrary eigenvalue placement for the observer.

If a unobservable system is considered the observable subspace decomposition in Sect. 3.8.12 can be applied. A similarity transformation $\mathbf{z} = \mathbf{Px}$ is applied where $\mathbf{P}$ is found from (3.345). If it is assumed that an observer gain matrix $\mathbf{L}$ has been found then it is known from Eq. (4.146) that

$$\mathbf{L}_t = \mathbf{PL}, \tag{4.159}$$

where $\mathbf{L}_t$ is the gain matrix for the transformed system. The eigenvalues for the stability matrix are the solutions of the equation:

$$det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{LC}) = 0. \tag{4.160}$$

Multiply now by the determinants of $\mathbf{P}$ and $\mathbf{P}^{-1}$ and make the transformation:

$$\begin{aligned} det(\mathbf{P}) \cdot det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{LC}) \cdot det(\mathbf{P}^{-1}) &= det(\mathbf{P}(\lambda\mathbf{I} - \mathbf{A} + \mathbf{LC})\mathbf{P}^{-1}) \\ = det(\lambda\mathbf{I} - \mathbf{PAP}^{-1} + \mathbf{PLCP}^{-1}) &= det(\lambda\mathbf{I} - \mathbf{A}_t + \mathbf{L}_t\mathbf{C}_t) = 0. \end{aligned} \tag{4.161}$$

If the system's observability matrix has the rank *p,* the gain matrix is partitioned as

$$\mathbf{L}_t = \begin{bmatrix} \mathbf{L}_{t1} \\ \mathbf{L}_{t2} \end{bmatrix}, \tag{4.162}$$

such that $\mathbf{L}_{t1}$ has the length $p$ and $\mathbf{L}_{t2}$ has length $n - p$. Insertion of the partitioned matrices into the last expression in Eq. (4.161) yields

$$\begin{aligned} det(\lambda\mathbf{I} - \mathbf{A}_t + \mathbf{L}_t\mathbf{C}_t) &= det\left(\lambda\mathbf{I} - \begin{bmatrix} \mathbf{A}_0 & \mathbf{0} \\ \mathbf{A}_{12} & \mathbf{A}_{no} \end{bmatrix} + \begin{bmatrix} \mathbf{L}_{t1} \\ \mathbf{L}_{t2} \end{bmatrix} [\mathbf{C}_0 \quad \mathbf{0}]\right) \\ &= det\left(\left[\begin{array}{c:c} \lambda\mathbf{I}_p - \mathbf{A}_0 + \mathbf{L}_{t1}\mathbf{C}_0 & \mathbf{0} \\ \hdashline -\mathbf{A}_{12} + \mathbf{L}_{t2}\mathbf{C}_0 & \lambda\mathbf{I}_{n-p} - \mathbf{A}_{no} \end{array}\right]\right) \\ &= det(\lambda\mathbf{I}_p - \mathbf{A}_0 + \mathbf{L}_{t1}\mathbf{C}_0) \cdot det(\lambda\mathbf{I}_{n-p} - \mathbf{A}_{no}) = 0. \end{aligned} \tag{4.163}$$

The last expression shows that the eigenvalues consist of the $p$ eigenvalues which can be influenced by the gain matrix $\mathbf{L}_{t1}$ and the remaining $n - p$ eigenvalues of the non observable system matrix $\mathbf{A}_n o$. These eigenvalues cannot be influenced by the observer gains and will therefore remain in their original positions. Thus it can be concluded that observability is also a *necessary* condition for arbitrary eigenvalue placement.

Another important conclusion can be drawn from the result above. It is clear that all the eigenvalues in the observable subspace can be assigned specific values, even if the observable subsystem is not stable. If the eigenvalues of the unobservable system are also in the left half plane then the observer can be made stable by proper choice of $\mathbf{L}_{t1}$. A system where the unobservable eigenvalues are stable is called *detectable*. See p. 124. Precisely similar conditions apply for discrete time systems.

## 4.8 Observer Design for MIMO Systems

It can be shown that the conditions for eigenvalue assignment for SISO observers discussed in the previous section are also valid in the MIMO case:

1. The observer eigenvalues can be placed arbitrarily if and only if the system is observable.

2. The stability matrix can be stabilized if and only if the unobservable states are stable (the system is detectable).

To find a design method for MIMO observers the duality properties can once again be utilized. The essential matrices for the controller and the observer cases are:

$$\begin{aligned}\mathbf{A_K} &= \mathbf{A} - \mathbf{BK},\\ \mathbf{A_L} &= \mathbf{A} - \mathbf{LC}.\end{aligned} \tag{4.164}$$

Following the same line of development as in Sect. 4.7.2, it is easy to see that designing a controller for the system,

$$S_x : \begin{cases}\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \\ \mathbf{y} = \mathbf{Cx}\end{cases}, \tag{4.165}$$

is the same as designing a state feedback controller for the dual system,

$$S_z : \begin{cases}\dot{\mathbf{z}} = \mathbf{A}^T\mathbf{z} + \mathbf{C}^T\mathbf{u} \\ \mathbf{y} = \mathbf{B}^T\mathbf{z}\end{cases}. \tag{4.166}$$

## 4.8.1 Eigenstructure Assignment for MIMO Observers

The robust eigenstructure assignment implemented in the MATLAB function `place`, can thus be used for observer design with a simple modification. The observer gain matrix **L** for the MIMO system (4.165) is determined by the command,

$$\texttt{L=place(A',C',[}\lambda_1,\lambda_2,\ldots,\lambda_n\texttt{])'},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the observer eigenvalues. Note that prime (') in MATLAB denotes matrix transposition.

## 4.8.2 Dead Beat Observers

By a derivation completely parallel to the treatment in Sect. 4.4.2 it can be shown that a discrete time observer with dead beat behaviour is obtained if the eigenvalues of the stability matrix are all placed in zero.

In this case the characteristic polynomial becomes

$$det(\lambda\mathbf{I} - \mathbf{F} + \mathbf{LC}) = P_{ch,\mathbf{F_L}}(\lambda) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \ldots + \alpha_1\lambda + \alpha_0 = \lambda^n. \tag{4.167}$$

A deadbeat observer will now be desiged for the system of Example 4.19.

***Example 4.19.*** **Predictive Deadbeat Observer**

A predictive dead beat observer is designed by using the $n$ eigenvalues,

$$\lambda_o = 0, 0, \ldots, 0.$$

Repeating the procedure from Example 4.17 leads to the gain matrix

$$\mathbf{L} = \begin{bmatrix} 1.3881 \\ -30.616 \\ -47.111 \end{bmatrix}$$

and a simulation of the estimation error under the same circumstances as in the two previous examples gives the responses on Fig. 4.40. All error components are exactly zero from the third sample instant, clearly indicating the dead beat behaviour. The price is, as it is usually the case with dead beat systems, that the gains as well as the signal amplitudes are large compared to the cases with more 'normal' eigenvalue placements. ❐

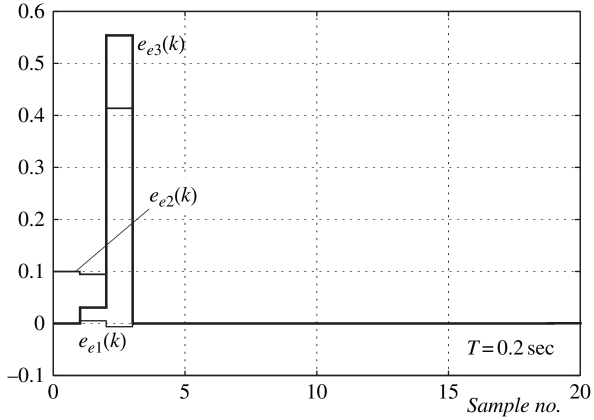


**Fig. 4.40** *Estimation error for* $\mathbf{e}_{e0} = [0 \quad 0.1 \quad 0]^T$ *(dead beat observer)*

## 4.9 Reduced Order Observers

The observers which have been described above are all full order. This means that all the states of the control object are estimated whether or not this is necessary. In some cases control objects have states which can be effectively measured without internal state or measurement noise while other states have large noise components. It may also be true that the modelling of some states is very accurate while the measurements of others are inaccurate.

Under these circumstances it might be advantageous to estimate a *subset* of the state variables. Such *model reductions* may be advantageous because they make it possible to simplify the control system and thus reduce system costs. This is nearly always an interesting proposition because not only does this reduce hardware costs but also implementation time and tuning difficulties. For these reasons reduced order observers are of great interest for practical applications.

Consider now a system divided into blocks containing the states which can be accurately measured directly, $\mathbf{x}_1$, and another set for which an observer is required, $\mathbf{x}_2$. In contrast to what is usually the case, it is clearly important here to order the states properly.

The state vector,

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \tag{4.168}$$

defines the blocks of the overall system according to the state equation,

$$\begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} \mathbf{u}, \tag{4.169}$$

where the state vector $\mathbf{x}$ is measured according to

$$\mathbf{y} = \mathbf{C}_1 \mathbf{x}_1. \tag{4.170}$$

It will be assumed for simplicity here that the matrix $\mathbf{C}_1$ is quadratic and nonsingular. This will very often be the case and the method presented will therefore cover most practical problems, although it is not applicable in all cases. The assumption is not absolutely necessary but it introduces a useful simplification and a straightforward design technique. For a more general approach the reader should see Friedland (1987) or Kwakernaak and Sivan (1972).

If a full order observer is to be constructed for this system, it should estimate the entire state vector,

$$\hat{\mathbf{x}} = \begin{bmatrix} \hat{\mathbf{x}}_1 \\ \hat{\mathbf{x}}_2 \end{bmatrix}. \tag{4.171}$$

However there is no reason to estimate the upper component of the state vector since it is available through the measurement (4.170). Therefore let

$$\hat{\mathbf{x}}_1 = \mathbf{x}_1 = \mathbf{C}_1^{-1} \mathbf{y}. \tag{4.172}$$

On the other hand it is necessary to introduce a general observer for the lower component $\mathbf{x}_2$ of the state vector,

$$\hat{\mathbf{x}}_2 = \mathbf{z} + \mathbf{L}\mathbf{y}, \tag{4.173}$$

where the auxiliary state vector $\mathbf{z}$ is determined from a state equation of the form,

$$\dot{\mathbf{z}} = \mathbf{Mz} + \mathbf{Nu} + \mathbf{Py}. \tag{4.174}$$

This vector has the same order as $\mathbf{x}_2$, i.e., $n - r$ since $\mathbf{x}_1$ and $\mathbf{y}$ have the dimension $r$.

If the estimation error is defined as before, the expression for the error will be

$$\mathbf{e}_e = \begin{bmatrix} \mathbf{x}_1 - \hat{\mathbf{x}}_1 \\ \mathbf{x}_2 - \hat{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{e}_{e1} \\ \mathbf{e}_{e2} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{e}_{e2} \end{bmatrix}. \tag{4.175}$$

Using Eqs. (4.169) and (4.173) an expression for $\dot{\mathbf{e}}_{e2}$ can be derived:

$$\dot{\mathbf{e}}_{e2} = \dot{\mathbf{x}}_2 - \dot{\hat{\mathbf{x}}}_2 = \mathbf{A}_{21}\mathbf{x}_1 + \mathbf{A}_{22}\mathbf{x}2 + \mathbf{B}_2\mathbf{u} - \mathbf{L}\dot{\mathbf{y}} - \dot{\mathbf{z}}. \tag{4.176}$$

Further using Eqs. (4.170) and (4.174) yields

$$\begin{aligned} \dot{\mathbf{e}}_{e2} &= \dot{\mathbf{x}}_2 - \dot{\hat{\mathbf{x}}}_2 = \mathbf{A}_{21}\mathbf{x}_1 + \mathbf{A}_{22}\mathbf{x}_2 + \mathbf{B}_2\mathbf{u} - \mathbf{L}\mathbf{C}_1\dot{\mathbf{x}}_1 - \mathbf{Mz} - \mathbf{Nu} - \mathbf{Py} \\ &= \mathbf{A}_{21}\mathbf{x}_1 + \mathbf{A}_{22}\mathbf{x}_2 + \mathbf{B}_2\mathbf{u} - \mathbf{L}\mathbf{C}_1(\mathbf{A}_{11}\mathbf{x}_1 + \mathbf{A}_{12}\mathbf{x}_2 + \mathbf{B}_1\mathbf{u}) - \mathbf{Mz} - \mathbf{Nu} - \mathbf{Py}. \end{aligned} \tag{4.177}$$

From (4.173), (4.170) and (4.175) is seen that

$$\mathbf{z} = \mathbf{x}_2 - \mathbf{e}_{e2} - \mathbf{L}\mathbf{C}_1\mathbf{x}_1. \tag{4.178}$$

Inserting this into (4.177) and after some manipulation this gives

$$\begin{aligned} \dot{\mathbf{e}}_{e2} = (\mathbf{A}_{21} - \mathbf{L}\mathbf{C}_1\mathbf{A}_{11} + \mathbf{M}\mathbf{L}\mathbf{C}_1 - \mathbf{P}\mathbf{C}_1)\mathbf{x}_1 + (\mathbf{A}_{22} - \mathbf{L}\mathbf{C}_1\mathbf{A}_{12} - \mathbf{M})\mathbf{x}_2 \\ + (\mathbf{B}_2 - \mathbf{L}\mathbf{C}_1\mathbf{B}_1 - \mathbf{N})\mathbf{u} + \mathbf{M}\mathbf{e}_{e2}. \end{aligned} \tag{4.179}$$

The first three right hand terms can be made zero by setting

$$\begin{aligned} \mathbf{M} &= \mathbf{A}_{22} - \mathbf{L}\mathbf{C}_1\mathbf{A}_{12}, \\ \mathbf{N} &= \mathbf{B}_2 - \mathbf{L}\mathbf{C}_1\mathbf{B}_1, \\ \mathbf{P} &= (\mathbf{A}_{21} - \mathbf{L}\mathbf{C}_1\mathbf{A}_{11})\mathbf{C}_1^{-1} + \mathbf{ML}, \end{aligned} \tag{4.180}$$

and (4.179) becomes

$$\dot{\mathbf{e}}_{e2} = \mathbf{M}\mathbf{e}_{e2} = (\mathbf{A}_{22} - \mathbf{L}\mathbf{C}_1\mathbf{A}_{12})\mathbf{e}_{e2}. \tag{4.181}$$

This last differential equation shows that $\mathbf{M}$ is the stability matrix of the reduced order observer and the design task is once again, reduced to determining $\mathbf{L}$ such that $\mathbf{M}$ will have a desired set of eigenvalues. The problem is quite parallel to finding $\mathbf{L}$ in the full order observer such that the matrix, $\mathbf{A_L} = \mathbf{A} - \mathbf{LC}$, gives the desired eigenvalues. The usual eigenvalue or eigenstructure design can therefore be used with the only modification that $\mathbf{C}$ must be replaced by $\mathbf{C}_1\mathbf{A}_{12}$.

The eigenvalues of the reduced order observer stability matrix are the solutions to

$$det(\lambda I - \mathbf{A}_{22} + \mathbf{LC}_1\mathbf{A}_{12}) = 0. \tag{4.182}$$

The condition for eigenvalue or eigenstructure assignment is also as before. The necessary and sufficient condition is that the observability matrix,

$$\mathbf{M}_o' = \begin{bmatrix} \mathbf{C}_1\mathbf{A}_{12} \\ \mathbf{C}_1\mathbf{A}_{12}\mathbf{A}_{22} \\ \mathbf{C}_1\mathbf{A}_{12}\mathbf{A}_{22}^2 \\ : \\ \mathbf{C}_1\mathbf{A}_{12}\mathbf{A}_{22}^{n-r-1} \end{bmatrix}, \tag{4.183}$$

has the full rank $n - r$.

A block diagram of the reduced order observer is shown on Fig. 4.41.

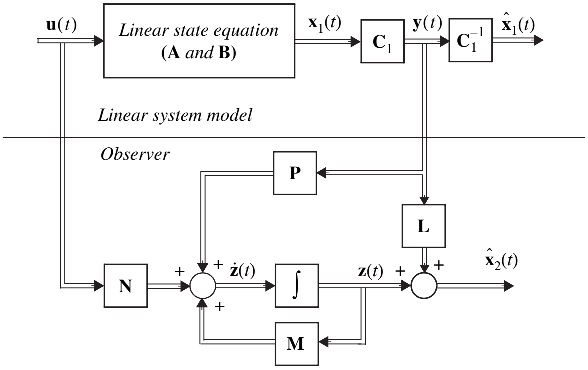


**Fig. 4.41** Reduced order observer

### Discrete Time Case

It should be noticed that the derivation above can be carried out in exactly the same way with difference equations and discrete time versions of the observer design shown above can easily be constructed. If the discrete time system matrix is partitioned in the same way as in Eq. (4.169),

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix}, \tag{4.184}$$

then the estimation error is determined by the difference equation,

$$\mathbf{e}_{e2}(k+1) = \mathbf{M}\mathbf{e}_{e2}(k) = (\mathbf{F}_{22} - \mathbf{L}\mathbf{C}_1\mathbf{F}_{12})\mathbf{e}_{e2}(k), \tag{4.185}$$

and the stability matrix is seen to be

$$\mathbf{M} = \mathbf{F}_{22} - \mathbf{L}\mathbf{C}_1\mathbf{F}_{12}. \tag{4.186}$$

***Example 4.20.*** **Reduced Order Observer for a Third Order System**

Returning to the third order system from Example 4.5. The model for this system is

$$\dot{\mathbf{x}} = \begin{bmatrix} -0.14 & 0.33 & -0.33 \\ 0.1 & -0.28 & 0 \\ 0 & 1.7 & -0.77 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ -0.025 \end{bmatrix}u, \; y = [2 \quad 0 \quad 0]\mathbf{x}.$$

Supposing that only the first state variable can be measured according to the output equation, an observer for estimation of the two remaining states is needed.

The system matrix partition (4.169) leads to

$$\mathbf{A}_{11} = -0.14, \; \mathbf{A}_{12} = [0.33 \; -0.33],$$
$$\mathbf{A}_{12} = \begin{bmatrix} 0.1 \\ 0 \end{bmatrix}, \; \mathbf{A}_{22} = \begin{bmatrix} -0.28 & 0 \\ 1.7 & -0.77 \end{bmatrix},$$

and the output equation can be written

$$y = \mathbf{C}_1 x_1 = 2x_1.$$

**L** will have dimension 2 × 1 and can be found if the matrix pair $\{\mathbf{A}_{22}, \mathbf{C}_1\mathbf{A}_{12}\}$ is observable. It is found that

$$\mathbf{M}'_o = \begin{bmatrix} \mathbf{C}_1\mathbf{A}_{12} \\ \mathbf{C}_1\mathbf{A}_{12}\mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} 0.66 & -0.66 \\ -1.3068 & 0.5082 \end{bmatrix} \text{and} \;\; det(\mathbf{M}'_o) = -0.5281.$$

so the observability requirement is fulfilled.

The observer should be much faster than the system and the following eigenvalues are selected,

$$\lambda_0 = -3.5 \pm j3.5.$$

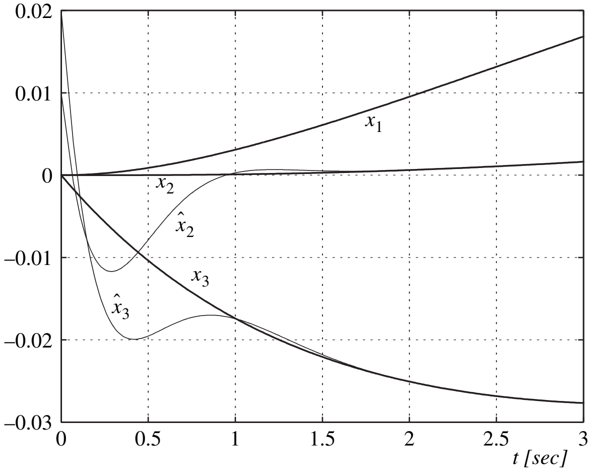


**Fig. 4.42** Unit step response of system and reduced order observer

Ackermann's Formula (4.156) for the observer case can be applied and the following result is found:

$$\mathbf{L} = \begin{bmatrix} -28.32 \\ -37.34 \end{bmatrix}.$$

The result of a simulation of the overall system is shown on Fig. 4.42. The block diagram of the system is seen on Fig. 4.41 and the system's initial state vector is a zero vector. The observer is however started with incorrect initial conditions, namely:

$$\hat{\mathbf{x}}_{20} = \begin{bmatrix} \hat{\mathbf{x}}_{20} \\ \hat{\mathbf{x}}_{30} \end{bmatrix} = \begin{bmatrix} 0.01 \\ 0.02 \end{bmatrix}.$$

A deviation between the states $x_2$ and $x_3$ on one side and the corresponding estimated states $\hat{\mathbf{x}}_2$ and $\hat{\mathbf{x}}_3$ on the other side is clearly visible over the first 1.5 s of the simulation. After this the curves coincide, showing that the estimation errors have vanished as was intended. ❐

## 4.10 State Feedback with Observers

It is clear from Sects. 4.6, 4.7, 4.8, 4.9 that it is possible to estimate the states of a linear system using a linear system model called an observer. Even if only a subset of the states of a control object is measured, it is still possible to make all the states available for feedback purposes, albeit some only as estimates. It is clear however that proper design techniques must be used to ensure accurate

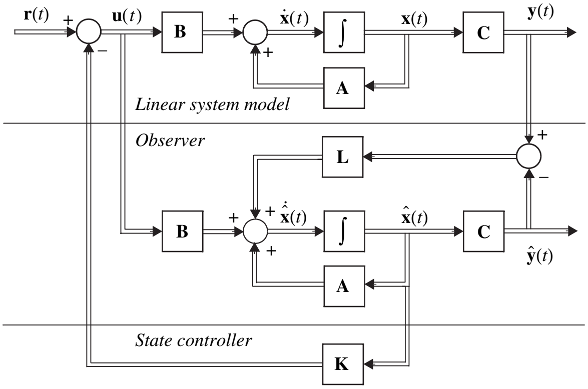


**Fig. 4.43** State feedback with full order observer, the design situation

estimates. It now has to be investigated how the combination of observer state estimates and full state feedback will work in conjunction with each other.

A block diagram of the observer structure to be studied in this chapter is in Fig. 4.43. It may be seen on the figure that the observer structure of Sect. 4.6 has been retained without any changes. What is new is that the state estimates are used in the control loop instead of direct state measurements. Most importantly, it is necessary to specify design guidelines for such apparently complex systems given the requirement that overall asymptotic stability must be attained.

It should be kept in mind that the observer (full order as well as reduced order) is a *linear approximation* to the real system and therefore it cannot exactly follow the states of the real system which is most likely nonlinear. It is very important to study the possible consequences of this fact. Usually this is most conveniently done by simulation as shown in Example 4.22.

## *4.10.1 Combining Observers and State Feedback*

As before the control object is to be described by the state equations:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{4.187}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t), \tag{4.188}$$

and the observer by the equation,

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{A}\hat{\mathbf{x}}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{L}(\mathbf{y}(t) - \mathbf{C}\hat{\mathbf{x}}(t)). \tag{4.189}$$

To establish full state feedback using the state estimates, the feedback law is

$$\mathbf{u}(t) = -\mathbf{K}\hat{\mathbf{x}}(t) + \mathbf{r}(t). \tag{4.190}$$

This immediately leads to the observer state equation (the time argument is omitted),

$$\dot{\hat{\mathbf{x}}} = (\mathbf{A} - \mathbf{BK} - \mathbf{LC})\hat{\mathbf{x}} + \mathbf{Ly} + \mathbf{Br}. \tag{4.191}$$

The estimation error is

$$\mathbf{e}_e = \mathbf{x} - \hat{\mathbf{x}} \Rightarrow \hat{\mathbf{x}} = \mathbf{x} - \mathbf{e}_e. \tag{4.192}$$

If $\hat{\mathbf{x}}$ is inserted into (4.190) and the resulting $\mathbf{u}$ into (4.187), the following equation is obtained

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{BK})\mathbf{x} + \mathbf{BK}\mathbf{e}_e + \mathbf{Br}. \tag{4.193}$$

The estimation error is governed by the equation

$$\dot{\mathbf{e}}_e = (\mathbf{A} - \mathbf{LC})\mathbf{e}_e. \tag{4.194}$$

Equations (4.193) and (4.194) can be combined to the $2n$-dimensional state equation,

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{e}}_e \end{bmatrix} = \begin{bmatrix} \mathbf{A} - \mathbf{BK} & \mathbf{BK} \\ \mathbf{0} & \mathbf{A} - \mathbf{LC} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{e}_e \end{bmatrix} + \begin{bmatrix} \mathbf{B} \\ \mathbf{0} \end{bmatrix} \mathbf{r}. \tag{4.195}$$

The characteristic equation of the overall system is

$$\begin{aligned} &det \begin{bmatrix} \lambda\mathbf{I} - \mathbf{A} + \mathbf{BK} & -\mathbf{BK} \\ \mathbf{0} & \lambda\mathbf{I} - \mathbf{I} - \mathbf{A} + \mathbf{LC} \end{bmatrix} \\ &= det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{BK}) \cdot det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{LC}) = 0. \end{aligned} \tag{4.196}$$

This shows that the eigenvalues of the overall system (4.195) consist of the union of the eigenvalues of the controller and the eigenvalues of the observer.

In other words, the eigenvalues of the controller and those of the observer are *independent of each other*. If the controller as well as the observer separately have been given eigenvalues in the left half plane, then one can be sure that the overall system will be asymptotically stable. This important result is often called the *separation principle* though this label was originally applied to stochastic systems.

In the presentation of the observer/full state feedback control system above, no account has been taken of the response time of the observer. In fact this has not been necessary at all. The reason for this is that if overall stability can be achieved then the response time of the observer is not important in itself, as long as it is at least as fast as the system itself, given the input which one wishes for the control system to follow. If the observer is not fast enough to follow the control object then this will degrade the quality of the control which can be attained, though not its stability, given a reasonably accurate system model for the controller and observer design.

To ensure that the observer is fast enough, the observer eigenvalues should normally be placed to the left of the controller eigenvalues in the complex plane. For continuous time systems it is often suggested that the observer eigenvalues are selected according to

$$\lambda_o = a \cdot \lambda_c, \tag{4.197}$$

where $\lambda_c$ is the controller eigenvalue and the multiplication factor $a$ is 3–10.

If this rule is applied a good result is achieved in most cases, but it should remembered that it is merely a coarse rule of thumb. Sometimes adequate performance is obtained even if the observer eigenvalues are placed quite close to the controller eigenvalues. It should also be noted that if the eigenvalues are placed too far into the left half plane, large observer gains and severe noise and disturbance sensitivity may be the result.

### Discrete Time Case

As before for controllers and observers, it is obvious that the discrete time systems can be treated precisely the same way as the continuous ones above. The discrete version of the $2n$-dimensional overall system state Equation (4.195) is

$$\begin{bmatrix} \mathbf{x}(k+1) \\ \mathbf{e}_e(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{F} - \mathbf{GK} & \mathbf{GK} \\ \mathbf{0} & \mathbf{F} - \mathbf{LC} \end{bmatrix} \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{e}_e(k) \end{bmatrix} + \begin{bmatrix} \mathbf{G} \\ \mathbf{0} \end{bmatrix} \mathbf{r}(k) \tag{4.198}$$

and consequently the characteristic equation reflecting the separation principle becomes

$$\begin{aligned} &det \begin{bmatrix} \lambda\mathbf{I} - \mathbf{F} + \mathbf{GK} & -\mathbf{GK} \\ \mathbf{0} & \lambda\mathbf{I} - \mathbf{F} + \mathbf{LC} \end{bmatrix} \\ &= det(\lambda\mathbf{I} - \mathbf{F} + \mathbf{GK}) \cdot det(\lambda\mathbf{I} - \mathbf{F} + \mathbf{LC}) = 0. \end{aligned} \tag{4.199}$$

### *Example 4.21.* Combined Observer/Pole Placement Control

For the third order system in Example 4.5 a state controller was designed with the gain matrix,

$$\mathbf{K} = [81.54 \quad 56.26 \quad -32.8],$$

resulting in the closed loop eigenvalues,

$$\lambda_{cl} = \begin{cases} -0.67 \\ -0.67 \\ -0.67 \end{cases} .$$

If the full state feedback is to be replaced with an observer and feedback of the estimated states instead of the real states, an observer gain matrix must be found such that the observer stability matrix has suitable eigenvalues. If the observer eigenvalues,

$$\lambda_o = \begin{cases} -3 \\ -2 \pm j2 \end{cases},$$

are selected, it is obvious that the observer is considerably faster than the controller.

The block diagram of the overall system is the one depicted on Fig. 4.43 with the exception that the input $u$ and the output $y$ are both scalars and the **L** matrix can be determined by application of Ackermann's Formula (4.156). The following matrix is obtained:

$$\mathbf{L} = \begin{bmatrix} 2.905 \\ -23.65 \\ -44.21 \end{bmatrix}. \tag{4.200}$$

If the system and the observer are both started with a zero initial vector (i.e $\mathbf{x}_0 = \hat{\mathbf{x}}_0 = \mathbf{0}$), it will hardly be possible to distinguish the step responses $\mathbf{x}(t)$ and $\hat{\mathbf{x}}(t)$ from each other and the output and the control signal will be as on Fig. 4.12. However if the two initial vectors are not the same, an estimation error will occur, at least for a period of time depending on how fast the observer is.

Now, two simulations are carried out using two different observer gain matrices, namely (4.200) and the matrix,

$$\mathbf{L} = \begin{bmatrix} 0.1050 \\ -0.01988 \\ -03985 \end{bmatrix},$$

giving the observer eigenvalues,

$$\lambda_o = \begin{cases} -0.6 \\ -0.4 \pm j0.4 \end{cases},$$

which means an observer which is five times 'slower' that the first one.

The estimation errors $\mathbf{e}_e(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t)$ are shown on Fig. 4.44 for the two cases. The initial values are

$$\mathbf{e}_{e0} = [\,0.0005 \quad -0.0005 \quad 0.001\,]^T.$$

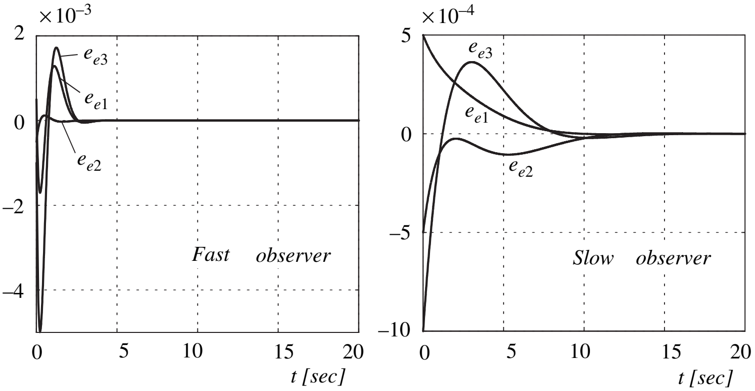


**Fig. 4.44** Estimation error for the two different observer cases

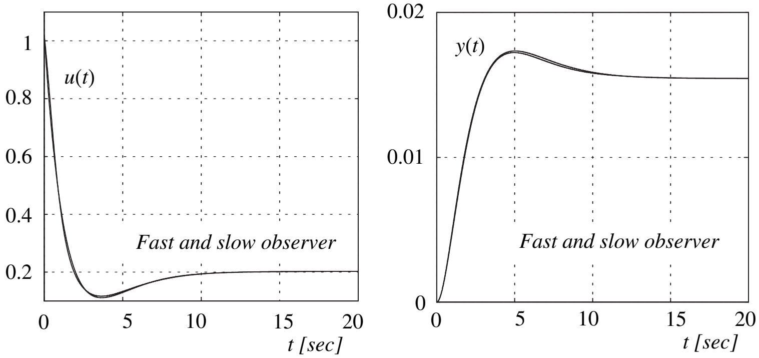


**Fig. 4.45** Input end output for the two different observer cases

If the fast observer is used, the initial estimation errors have disappeared in less than 4 s, whereas the slow observer needs more than 15 s to remove the error. The apparently large differences in the estimation errors do not reflect similar differences in the input and output signals, which are shown on Fig. 4.45. Both signals are almost identical for the fast and the slow observer.

This example indicates that it is not *necessarily* advantageous to make the observer very fast compared with the controller. In fact, the slow observer has much smaller gains than the fast one and it will therefore be less prone to exhibiting undesired noise sensitivity.

## *4.10.2 State Feedback with Integral Controller and Observer*

The observer based state feedback controller can be combined with the error integration introduced in Sect. 4.5. A simplified block diagram of an overall

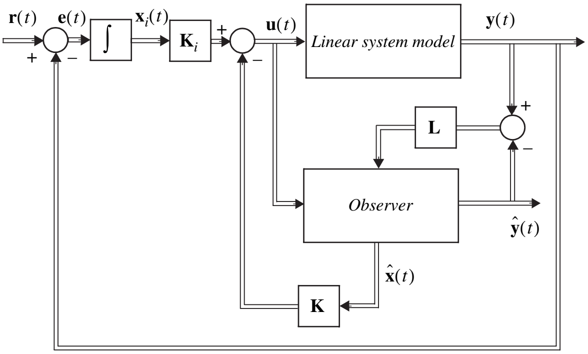


**Fig. 4.46** State feedback with integral controller and observer. Note, that in the real situation the linear system model above is replaced by the physical system which in most cases is nonlinear

system with such a controller structure is shown on Fig. 4.46. The entire set of equations for the system consists of the already known relationships:

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu},\\ \dot{\mathbf{x}}_i &= \mathbf{r} - \mathbf{Cx},\\ \dot{\hat{\mathbf{x}}} &= \mathbf{A}\hat{\mathbf{x}} + \mathbf{Bu} + \mathbf{LC}(\mathbf{x} - \hat{\mathbf{x}}),\\ \mathbf{u} &= -\mathbf{K}\hat{\mathbf{x}} + \mathbf{K}_i\mathbf{x}_i.\end{aligned} \tag{4.201}$$

If an overall system state vector is defined, the following model can be obtained:

$$\begin{bmatrix}\dot{\mathbf{x}}\\ \dot{\mathbf{x}}_i\\ \dot{\hat{\mathbf{x}}}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{BK}_i & -\mathbf{BK}\\ -\mathbf{C} & \mathbf{0} & \mathbf{0}\\ \mathbf{LC} & \mathbf{BK}_i & \mathbf{A} - \mathbf{LC} - \mathbf{BK}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{x}_i\\ \hat{\mathbf{x}}\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\ \mathbf{I}_r\\ \mathbf{0}\end{bmatrix}\mathbf{r}. \tag{4.202}$$

Similarity transformation of this system using the matrix,

$$\mathbf{P} = \begin{bmatrix}\mathbf{I}_n & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_r & \mathbf{0}\\ \mathbf{I}_n & \mathbf{0} & -\mathbf{I}_n\end{bmatrix} = \mathbf{P}^{-1}, \tag{4.203}$$

leads to the state vector,

$$\mathbf{z} = \mathbf{P}\begin{bmatrix}\mathbf{x}\\ \mathbf{x}_i\\ \hat{\mathbf{x}}\end{bmatrix} = \begin{bmatrix}\mathbf{x}\\ \mathbf{x}_i\\ \mathbf{x} - \hat{\mathbf{x}}\end{bmatrix} = \begin{bmatrix}\mathbf{x}\\ \mathbf{x}_i\\ \mathbf{e}_e\end{bmatrix}, \tag{4.204}$$

and the overall system matrix,

$$\mathbf{A}_{at} = \mathbf{PA}_a\mathbf{P}^{-1} = \begin{bmatrix}\mathbf{A} - \mathbf{BK} & \mathbf{BK}_i & \mathbf{BK}\\ -\mathbf{C} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{A} - \mathbf{LC}\end{bmatrix}, \tag{4.205}$$

where $\mathbf{A}_a$ is the system matrix in Eq. (4.202).
The matrix is block triangular and the eigenvalues can be found from

$$\begin{aligned} det(\lambda\mathbf{I} - \mathbf{A}_{at}) &= det\left(\lambda\mathbf{I} - \left[\begin{array}{cc:c} \mathbf{A}-\mathbf{BK} & \mathbf{BK}_i & \mathbf{BK} \\ -\mathbf{C} & \mathbf{0} & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} & \mathbf{A}-\mathbf{LC} \end{array}\right]\right) \\ &= det\left(\lambda\mathbf{I} - \begin{bmatrix} \mathbf{A}-\mathbf{BK} & \mathbf{BK}_i \\ -\mathbf{C} & \mathbf{0} \end{bmatrix}\right) \cdot det(\lambda\mathbf{I} - (\mathbf{A}-\mathbf{LC})) = 0. \end{aligned} \tag{4.206}$$

This means that the overall set of eigenvalues consists of the eigenvalues for the feedback integral controller (see Eq. (4.100)) plus the observer eigenvalues.

### Discrete Time Case

For the discrete time systems, quite similar manipulations can be carried out and the overall system matrix corresponding to the state vector,

$$\mathbf{z}(k) = \begin{bmatrix} \mathbf{x}(k) \\ \mathbf{x}_i(k) \\ \mathbf{e}_e(k) \end{bmatrix}, \tag{4.207}$$

becomes

$$\mathbf{F}_{at} = \begin{bmatrix} \mathbf{F}-\mathbf{GK} & \mathbf{GK}_i & \mathbf{GK} \\ -\mathbf{C} & \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{F}-\mathbf{LC} \end{bmatrix}. \tag{4.208}$$

One can immediately see that the conclusion concerning the separation eigenvalues is precisely the same as for (4.205).

***Example 4.22.* MIMO Observer/Integrator Control of the Robot Arm**

Example 4.14 investigated the integration state controller for the two link robot from Example 2.10.

If the full state feedback controller is replaced with a full order observer and the additional output feedback and the integrators maintained, the overall system on Fig. 4.47 results. The system is the continuous/discrete 'hybrid' version of the system on figure 4.46. To obtain a design for the entire control system it is only necessary to combine the integrating controller from example 4.14 with the controller gain matrix,

$$\begin{aligned} \mathbf{K}_1 &= [\mathbf{K} \quad -\mathbf{K}_i] \\ &= \begin{bmatrix} 841.3 & 76.385 & 153.99 & 20.175 & -63.628 & -8.1492 \\ 343.88 & 30.304 & 90.155 & 10.891 & -27.411 & -5.175 \end{bmatrix}, \end{aligned}$$

and with an observer gain matrix ensuring an appropriate eigenvalue placement for the stability matrix $\mathbf{F_L} = \mathbf{F} - \mathbf{LC}$.

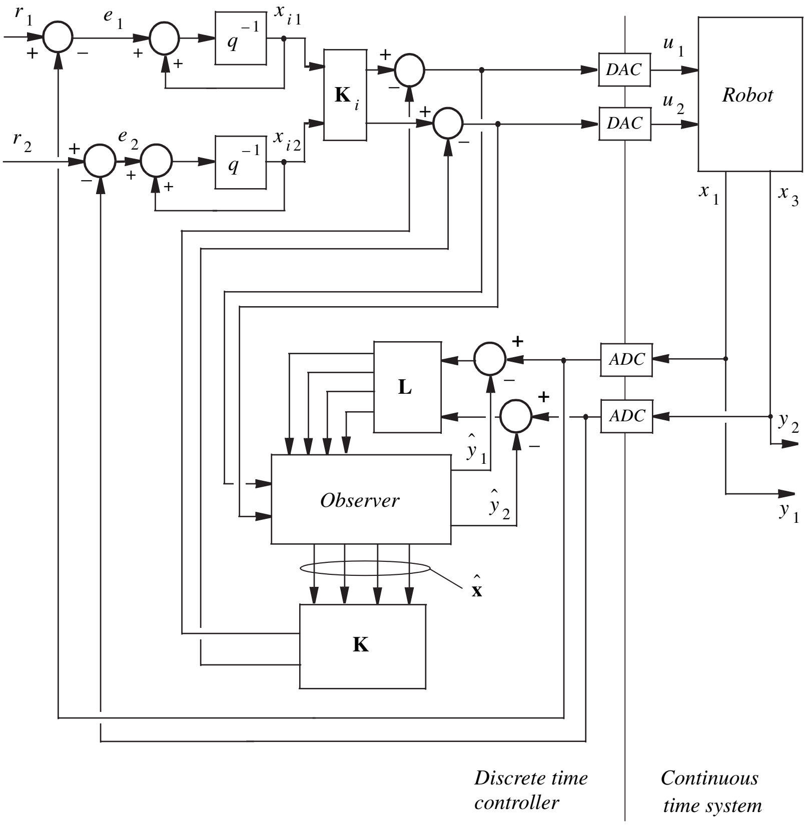


**Fig. 4.47** Controller and observer for two link robot

The six controller eigenvalues are

$$\lambda_{cl} = \begin{cases} -7 \pm j7 \\ -8.6 \pm j5 \\ -9.7 \pm j2.6 \end{cases}$$

in continuous time and

$$\lambda_{cld} = e^{0.02\lambda_{cl}} = \begin{cases} 0.86085 \pm j0.12131 \\ 0.83777 \pm j0.084058 \\ 0.82254 \pm j0.042811 \end{cases} \tag{4.209}$$

in discrete time. The sample period is $T = 0.02$ s.

According to the usual rule of thumb, the observer should be 3–10 times faster than the controller. If the 4 best damped controller eigenvalues above are picked as a starting point and if a factor 5 is selected, the following discrete time observer eigenvalues will be obtained:

$$\lambda_o = \begin{cases} 0.3714 \pm j0.2029 \\ 0.3663 \pm j0.0975 \end{cases} . \tag{4.210}$$

Applying the MATLAB `place` function leads to the gain matrix

$$\mathbf{L} = \begin{bmatrix} 1.2644 & 2.341 \cdot 10^{-4} \\ 22.373 & -0.083636 \\ -0.013327 & 1.2715 \\ -1.0311 & 20.879 \end{bmatrix} . \tag{4.211}$$

If all the known matrices are inserted into the 10 × 10-dimensional overall system matrix (4.208), it will be found that the 10 eigenvalues will be the union of (4.209) and (4.210), precisely as expected.

A simulation corresponding to the one carried out in Example 4.14 gives the results on Figs. 4.48 and 4.49. Comparison with Figs. 4.32 and 4.33 reveals that the system with observer is not quite as robust as the system in Example 4.14. The performance is slightly inferior (more oscillatory) after $t = 3$ s. where the links have moved far away from the linearization point. This is not uncommon for control systems containing observers.

It may be possible to improve the performance by making the observer faster. If a factor 10 is used for the observer eigenvalues instead of 5, the result on Fig. 4.50 is obtained. Only the angle $\theta_2$ and the input $u_1$ are shown for simplicity. The angle response is clearly better than the previous one for a slower observer. However some of the gains are also larger,

$$\mathbf{L} = \begin{bmatrix} 1.7578 & -1.594 \cdot 10^{-4} \\ 39.190 & -0.031706 \\ -9.1553 \cdot 10^{-3} & 1.8106 \\ -0.89259 & 42.294 \end{bmatrix} , \tag{4.212}$$

and this will increase the noise sensitivity.

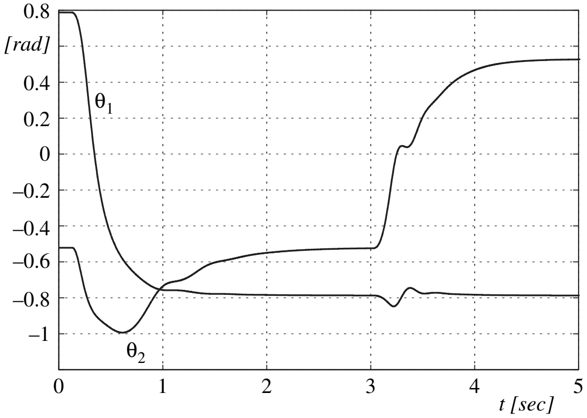


**Fig. 4.48** Robot link angle step responses

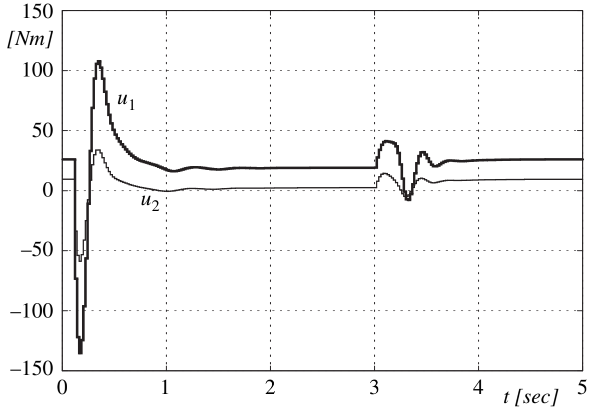


**Fig. 4.49** Control signals (motor torques)

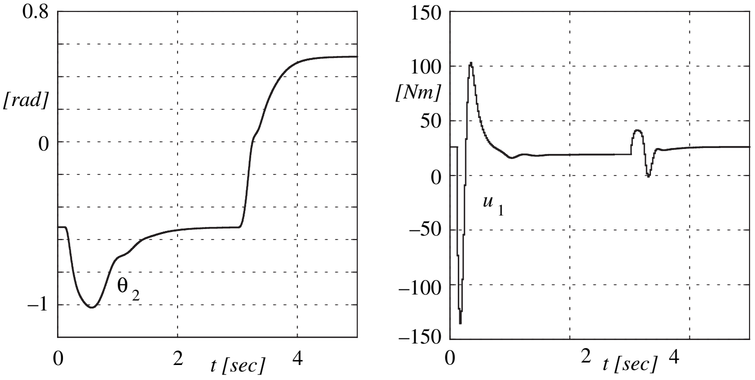
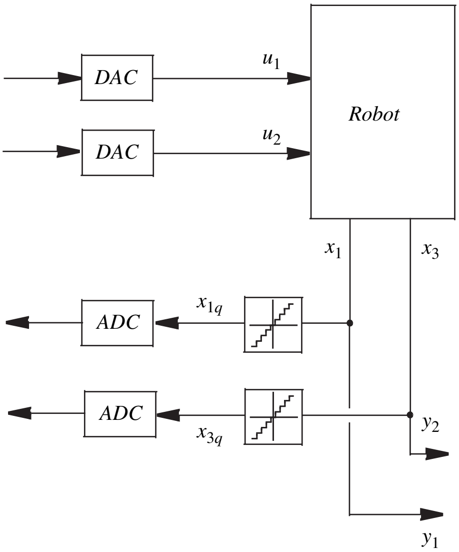


**Fig. 4.50** Responses with a faster observer

Noise effects are not pursued at this point, but it must be pointed out that at least one source of noise is inevitable in hybrid systems with analog-to-digital converters. The converter will not only sample the continuous signal at discrete instants of time, it will also *quantize* it. The deviation between the continuous signal and the quantized signal depends on the resolution of the ADC and this is given by the number of bits the converter can handle linearly. There is no quantization involved in the simulations above but this can easily be included in the simulation. It is just a matter of adding the quantizers in the output lines as seen on Fig. 4.51. A new simulation using the fast observer (4.212) and with a 12 bit converter gives the responses on Fig. 4.52. The angle seems unaffected but the control signal is clearly noisier than before because the *quantization noise* propagates through the system from the quantizer to the input signal. A part of a record of the quantization noise is seen on Fig. 4.53. If a 10 bit converter is used, the measurement noise problem becomes even more evident as seen on

**Fig. 4.51** Quantizers inserted into output lines



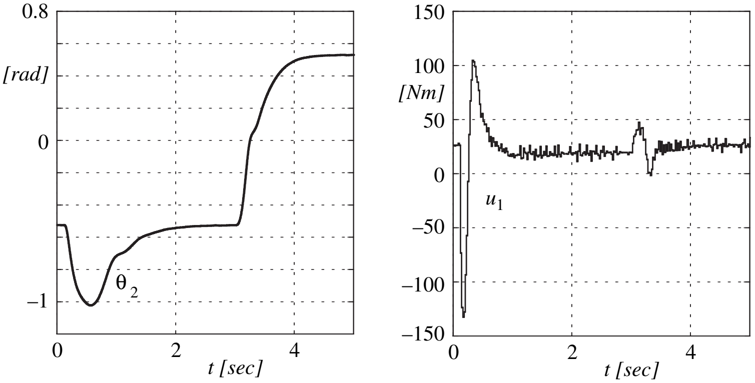


**Fig. 4.52** Responses with a faster observer (12 bit quantization)

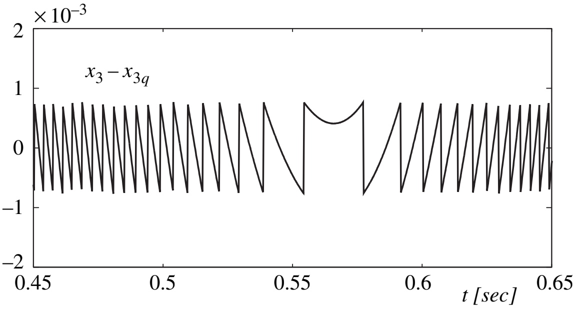


**Fig. 4.53** Quantization noise (12 bit quantization)

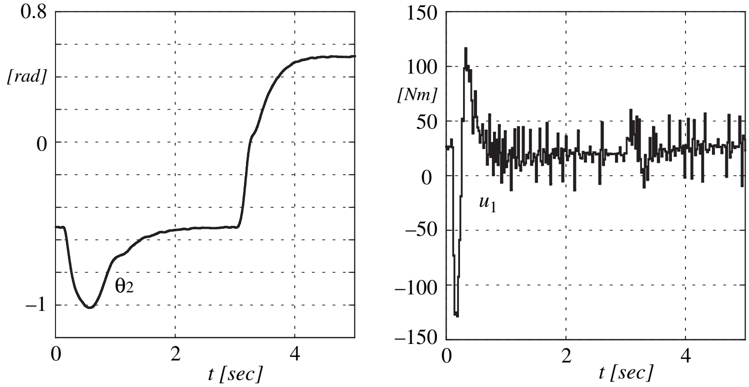


**Fig. 4.54** Responses with a faster observer (10 bit quantization)

Fig. 4.54. In this case the input noise has become worse and now one can also see faintly that the noise influences the link angle $\theta_2$. With the slower observer (4.211) this noise influence would be much more modest.

It can be concluded that there is a price to be paid if the observer is made too fast. This is in fact a general problem. It is usually the case that there is a *trade off* between the requirement that the observer should follow the states closely and its sensitivity to noise and disturbances. This problem is treated in more detail in later chapters. ❒

## *4.10.3 State Feedback with Reduced Order Observer*

Given the simplicity that the separation principle makes possible in the design of full order observer/feedback systems, it is of great interest to determine how full order feedback will function together with reduced order observers. The overall system is shown on Fig. 4.55.

A reduced order observer has been derived in Sect. 4.9 for the control object given by Equation (4.187). In this treatment the states $\mathbf{x}_1$ are measured while the remaining states are estimated on the basis of these measurements. The estimation errors are given by

$$\mathbf{e}_{e1} = \mathbf{x}_1 - \hat{\mathbf{x}}_1 = \mathbf{0} \tag{4.213}$$

$$\mathbf{e}_{e2} = \mathbf{x}_2 - \hat{\mathbf{x}}_2 \Rightarrow \hat{\mathbf{x}}_2 = \mathbf{x}_2 - \mathbf{e}_{e2} \tag{4.214}$$

and $\mathbf{e}_{e2}$ is governed by the state equation

$$\dot{\mathbf{e}}_{e2} = \mathbf{M}\mathbf{e}_{e2} = (\mathbf{A}_{22} - \mathbf{L}\mathbf{C}_1\mathbf{A}_{12})\mathbf{e}_{e2}. \tag{4.215}$$

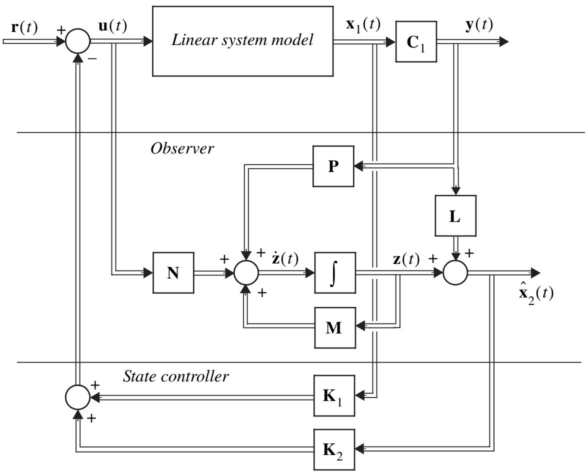


**Fig. 4.55** State feedback with reduced order observer (see the remarks to Fig. 4.46)

The control signal is given by

$$\mathbf{u} = -\mathbf{K}\hat{\mathbf{x}} + \mathbf{r} = -\begin{bmatrix} \mathbf{K}_1 & \mathbf{K}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \hat{\mathbf{x}}_2 \end{bmatrix} + \mathbf{r} \tag{4.216}$$

since $\hat{\mathbf{x}}_1 = \mathbf{x}_1$. Inserting (4.216) and (4.214) into (4.187) yields,

$$\begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} - \begin{bmatrix} \mathbf{B}_1\mathbf{K}_1 & \mathbf{B}_1\mathbf{K}_2 \\ \mathbf{B}_2\mathbf{K}_1 & \mathbf{B}_2\mathbf{K}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 - \mathbf{e}_{e2} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} \mathbf{r}, \tag{4.217}$$

where the partitioning of **A** and **B** is the same is in Eq. (4.169). Combining (4.217) with (4.215) leads to

$$\begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \\ \dot{\mathbf{e}}_{e2} \end{bmatrix} = \left[\begin{array}{cc:c} \mathbf{A}_{11} - \mathbf{B}_1\mathbf{K}_1 & \mathbf{A}_{12} - \mathbf{B}_1\mathbf{K}_2 & \mathbf{B}_1\mathbf{K}_2 \\ \mathbf{A}_{21} - \mathbf{B}_2\mathbf{K}_1 & \mathbf{A}_{22} - \mathbf{B}_2\mathbf{K}_2 & \mathbf{B}_2\mathbf{K}_2 \\ \hdashline \mathbf{0} & \mathbf{0} & \mathbf{M} \end{array}\right] \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{e}_{e2} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \mathbf{0} \end{bmatrix} \mathbf{r} \tag{4.218}$$

or

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{e}}_{e2} \end{bmatrix} = \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{K} & \mathbf{B}\mathbf{K}_2 \\ \mathbf{0} & \mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{e}_{e2} \end{bmatrix} + \begin{bmatrix} \mathbf{B} \\ \mathbf{0} \end{bmatrix} \mathbf{r}. \tag{4.219}$$

The overall system matrix is block triangular and the eigenvalues are therefore determined by the characteristic equation,

$$det \begin{bmatrix} \lambda\mathbf{I} - \mathbf{A} + \mathbf{B}\mathbf{K} & -\mathbf{B}\mathbf{K}_2 \\ \mathbf{0} & \lambda\mathbf{I}_{n-r} - \mathbf{M} \end{bmatrix} = det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{B}\mathbf{K}) \cdot det(\lambda\mathbf{I} - \mathbf{M}) = 0, \tag{4.220}$$

where $\mathbf{I}_{n-r}$ is the $n - r$-dimensional identity matrix.

Equation (4.220) shows that the separation principle is also valid in this case.

## 4.11 Summary

This chapter has presented three main subjects:

1. The use of full state feedback for control system design,
2. The use of deterministic observers for state estimation and
3. The combined use of observers and full state feedback in control system design.

The basic method of designing feedback control loops for multivariable systems is eigenvalue placement and this is the subject of the Sects. 4.1, 4.2, 4.3, 4.4, 4.5 of this chapter. It has been shown that it is possible to place the closed loop eigenvalues of any controllable system arbitrarily in the complex plane. A vital practical issue is that feedback loops must nearly always be offset so that they regulate around a point in input/state space which is not the origin. Once offset there is very little difference in how the feedback loop itself is designed: a simple application of superposition is sufficient given an accurate system model. When there are well defined state disturbances in a system these may be suppressed by the use of integrators as is well known in classical control system design.

Sections 4.6, 4.7, 4.8, 4.9 of this chapter have dealt with the design of linear deterministic observers, sometimes called Luenberger observers, for state feedback systems. These observers can be based on incomplete or possibly noise corrupted measurements of the states of the control object. Both full and reduced order observers have been treated. Only one type of observer structure has been considered but this is also the only type which is widely known in the literature. Later it will be apparent that exactly the same type of observer, having the same form, can also be derived from statistical considerations. In this case the observer structure is called a Kalman filter (see Chap. 7).

The main method of observer design which has been treated is that of eigenvalue placement and this method is entirely dependent on the characteristic equation of the observer. The characteristic equation defines the dynamic behavior of the observer with respect to the control object itself. In general an observer is designed to be somewhat faster than the control object itself in order to be able to follow the states of the system as rapidly and as accurately as possible. This is necessary for accurate control. In this connection it must be remembered that there is no requirement that the control object itself have non-minimum phase or even stable dynamics. The only requirement is that the observer is stable. This is in general possible within broad limits if the underlying control object is observable and the observer gains properly designed. If reduced order observers are considered then the requirements are somewhat stronger.

Section 4.10 has dealt mainly with the separation principle and its use for the design of observer based control systems. The main conclusion is that for linear systems, because of the principle of superposition, it is possible to design the observer and its corresponding full feedback control system independently of each other. The resulting system will have sets of observer and feedback eigenvalues at the design locations which can be placed independently of each other. This is of course under the assumption that the control object is controllable as well as observable.

## 4.12 Notes

### *4.12.1 Background for Observers*

The explicit use of embedded models for state estimation in control systems is by now a very old idea. In work published by Kalman and Bertram (1958) the use of open loop observers first occurs. Closed loop observers appeared in the early 1960's, both as deterministic observers and stochastic filters. Kalman filters were first mentioned in fundamental work on optimal filters by Kalman (1960) and Kalman and Bucy (1961). Kalman filters were originally full order, closed loop, stochastically optimal observers. A different approach was taken by Luenberger (1964,1966,1971) who concentrated on deterministic systems. His Ph. D. Thesis at Stanford from 1963 deals with a somewhat more general estimation problem than that treated by Kalman and Bucy which lead to an observer which has a reduced order with respect to the system with which it is used. It was also Luenberger who originated the name 'observer'. Thus 'Luenberger observer' is sometimes applied to reduced order observers.

## 4.13 Problems

### *Problem 4.1*

Given the continuous second order system:

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & -2 \\ 0.5 & -1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 2 \\ 2 \end{bmatrix}u,$$

$$y = [-1 \quad 1]\mathbf{x}.$$

a. Find the eigenvalues of the system.
b. Design a state controller for the system such that the closed loop system becomes two complex poles with the natural frequency 2 rad/sec and the damping ratio 0.707. Use the direct method of Example 4.3.
c. Calculate the closed loop system's unit step response $y(t)$. Calculate also the control signal $u(t)$.

### *Problem 4.2*

a. Repeat questions b. and c. from Problem 4.1 but with the matrix $\mathbf{B} = \begin{bmatrix} 4.1 \\ 2 \end{bmatrix}$.

b. Compare the results with those obtained in Problem 4.1 and comment on the differences.

### *Problem 4.3*

Given the system:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -2 & 2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u.$$

a. Is the system stable?
Is the system controllable?
Can the system be stabilized by linear state feedback?

b. If possible, determine the gain matrix for a linear state feedback controller by using Ackermann's formula. The closed loop system eigenvalues must be placed in the positions $\lambda = -1 \pm j$.

### *Problem 4.4*

Consider the following discrete time model for a linear third order system:

$$\mathbf{x}(k+1) = \begin{bmatrix} -2 & -3 & 5 \\ -0.875 & 0.5 & 1 \\ -1.875 & -0.5 & 3 \end{bmatrix} \mathbf{x}(k) + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u(k).$$

The sample period is $T = 0.2\,\text{sec}$.

a. Is the system stable?
Is the system controllable?

b. Transform the model to controller canonical form.

c. Design a linear feedback controller with the gain matrix **K** and with the *continuous time* eigenvalues

$$\lambda_{cont} = \begin{cases} -2 \\ -2 \pm j \end{cases}.$$

d. Repeat c. by using Ackermann's formula.

### *Problem 4.5*

The following linear model for a third order system is given:

$$\dot{\mathbf{x}} = \begin{bmatrix} -4 & -6 & 8 \\ -4 & -1 & 3 \\ -5 & -2 & 5 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 2 \\ -2 \\ -1 \end{bmatrix} u.$$

a. Show that the system is neither stable nor controllable.
b. Is the system stabilizable? (Use controllable subspace decomposition).
c. Design a state feedback controller such that the resulting closed loop system has the complex conjugate eigenvalue pair $\lambda = -2 \pm j2$.
d. Check the entire set of eigenvalues of the closed loop system.

### *Problem 4.6*

Consider the hydraulic position servo from Example 3.26. Let the state variables be the position, the velocity and the pressure difference.

a. Formulate a third order linear state equation for the system, insert the data from Example 3.26 and calculate the eigenvalues.
b. Design a state feedback controller with the gain matrix **K** and with the closed loop eigenvalues

$$\lambda_{cl} = \begin{cases} -20 \\ -12 \pm j12 \end{cases}.$$

c. Show that the system will have a stationary error for the reference $r = 0$ and a constant load force $f \neq 0$. Calculate the error for $f = 500$ N.
d. Design a state feedback controller with integration. Calculate $\mathbf{K}_1$ for the closed loop eigenvalues

$$\lambda_{cl1} = \begin{cases} -20 \\ -16 \\ -12 \pm j12 \end{cases}.$$

e. Calculate the stationary error for $f = 500$ N.
f. Calculate the unit step response of the system with and without integration. Calculate also the response to a 500 N step in the disturbance in the two cases.

### *Problem 4.7*

Reconsider the hydraulic servo of Problem 4.6 with the difference that a discrete time controller shall be designed.

a. Use the same system data as in Problem 4.6 a. and discretize the system with the sampling period $T = 2$ ms.
b. Design a discrete time state feedback controller with the gain matrix **K** and with the *continuous time* closed loop eigenvalues:

$$\lambda_{cl} = \begin{cases} -20 \\ -12 \pm j12 \end{cases}$$

c. Show that the system will have a stationary error for the reference $r(k) = 0$ and a constant load force $f(k) \neq 0$. Calculate the error for $f(k) = 500$ N.

d. Design a discrete time state feedback controller with integration. Calculate $\mathbf{K}_1$ for the continuous time closed loop eigenvalues

$$\lambda_{cl1} = \begin{cases} -20 \\ -16 \\ -12 \pm j12 \end{cases}.$$

e. Calculate the stationary error for $f(k) = 500$ N.

f. Calculate the unit step response of the system with and without integration. Calculate also the response to a 500 N step in the disturbance in the two cases.

***Problem 4.8***

a. A full order observer should be designed for the system in Problem 4.3 with

$$y = [0 \quad 3]\mathbf{x}.$$

Is this possible?

b. If the answer is yes, transform the system to observer canonical form and determine the observer gain matrix $\mathbf{L}$ such that the observer eigenvalues becomes

$$\lambda_{\mathbf{A_L}} = -4 \pm j4.$$

c. Repeat b. by use of Ackermann's formula.

d. Draw a block diagram with observer and the state controller from problem 4.3.

***Problem 4.9***

Given the following linear model for a third order system:

$$\dot{\mathbf{x}} = \begin{bmatrix} -4 & -6 & 8 \\ -4 & -1 & 3 \\ -5 & -2 & 5 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 2 \\ -2 \\ -1 \end{bmatrix}u,$$
$$y = [-1 \quad 0 \quad 1]\mathbf{x}.$$

a. Is the system stable?
Is it observable?

b. Is the system detectable? (Use observable subspace decomposition).

c. Design a full order observer such that the set of observer eigenvalues includes the complex conjugate pair $\lambda = -5 \pm j5$.

d. Check the entire set of eigenvalues of the observer.

### *Problem 4.10*

Consider the tank system in Example 2.9. Suppose that the level $H_2$ and the temperature $T_2$ can be measured. It is desired to design a reduced order observer for estimation of the two remaining states $H_1$ and $T_1$.

a. Use the matrices (2.102), (2.103), (2.104), (2.105) for the linearized system in Example 2.9 and carry out a similarity transformation which will transform the system to the form (4.169) which is required for the observer design in Sect. 4.9. This means that the state vector should be

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} H_2 \\ T_2 \\ H_1 \\ T_1 \end{bmatrix}.$$

b. Determine the observer matrices **L**, **M**, **N** and **P** such that the observer has the eigenvalues

$$\lambda_o = -0.4 \pm j0.5.$$

One may want to use the MATLAB `place` function in the design process.

c. Carry out a simulation (for instance using SIMULINK) on the linear system and check if the reduced order observer works, also if the observer is started in a 'wrong' state.

# Chapter 5
# Optimal Control

**Abstract** The principle of optimality in terms of minimization of a performance index is introduced. For continuous and discrete time state space system models a performance index is minimized. The result is applied in the case of a quadratic index for a linear system and the resulting time dependent and time independent Riccati equations are derived. The conditions for guaranteed stability of the steady-state Linear Quadratic Regulator (LQR) are presented. An eigenstructure assignment approach to the steady-state LQ regulator problem is developed.

## 5.1 Introduction to Optimal Control

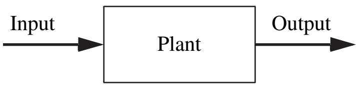
The design of a controller for the control of a process as shown in Fig. 5.1 consists of providing a control signal, the input, which will make the plant behave in a desired fashion, i.e., make the output change in a way described in a set of performance specifications. In classical control the performance specifications are given in terms of desired time domain and frequency domain measures, such as step response specifications (overshoot, rise time, settling time), frequency response specifications (bandwidth, resonance frequency, resonance damping) and relative stability in terms of phase and gain margins. Further, the specifications may be given in terms of disturbance rejection and noise suppression measures, specifying the desired frequency response of the sensitivity function and complementary sensitivity function.

As is well known in classical controller design, many of the above specifications are conflicting in the sense that they lead to impossible or conflicting requirements for the controller parameters. Typically, short rise time and high bandwidth require high gain controllers whereas small overshoot and good relative stability favor small gains in the controller. Classical controller design therefore often requires judicious and sometimes skillful selection of controller structure and parameters in order to find a reasonable compromise between conflicting performance specifications. This calls for time consuming ‘trial and error’ tuning on the part of the control engineer and does not lend itself well to automatic tuning. This is especially true with MIMO systems.

E. Hendricks et al., *Linear Systems Control*, DOI: 10.1007/978-3-540-78486-9_5,

© Springer-Verlag Berlin Heidelberg 2008

**Fig. 5.1** The control problem



In an effort to overcome some of these problems and in order to be able to design the 'best possible' controller, the methods presented here may be used. Basically, a measure of the quality of a controller is formulated in this chapter in terms of a *performance index*. This index is used to design the controller and depends on the control signal and the state vector. In this way the 'best' control signal is found that results in the minimum (or maximum) value of the index.

The job of the control engineer in Linear Quadratic Regulator (LQR) design is therefore not to determine control parameters directly, but to define the appropriate measure for controller quality, the performance index, and to minimize or maximize it. Several types of performance index will be introduced, leading to different kinds of controllers, but the basic idea is the same and will be introduced in the following section.

## 5.2 The General Optimal Control Problem

In this section the general optimal control problem is introduced and then solved in the following sections. Initially a general approach will be described which is suitable for nonlinear systems. Later the discussion will be specialized to linear systems.

Assume that an $n$-dimensional system with state vector $\mathbf{x}(t) \in \Re^n$ and input vector $\mathbf{u}(t) \in \Re^m$ is described by a general nonlinear, time varying state equation:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t). \tag{5.1}$$

Assume further, that the state vector has the value $\mathbf{x}_0$ at the initial time, i.e.,

$$\mathbf{x}(t_0) = \mathbf{x}_0. \tag{5.2}$$

It is desired to optimize the control over some time interval up to a final time, $t_1$. In addition an investigation will be made of optimal controllers both where time $t_1$ is given and also where there is a state or output requirement at the final time.

The system performance can be described by a *performance index*, which is a time integral depending on state and input vectors:

$$J(\mathbf{u}) = \Phi(\mathbf{x}(t_1), t_1) + \int_{t_0}^{t_1} L(\mathbf{x}(t), \mathbf{u}(t), t)dt. \tag{5.3}$$

The integrand, $L$, is a real-valued function of state and input vector and may also depend explicitly on time. It is often called the *cost function*. The function $\Phi$ is a real-valued function of the final state vector and the final time. The performance index reflects the quality of a controller and should be constructed in such a way that it is limited from below and such that the larger the index, the poorer the control. This may be achieved by requiring that the first term and the integrand in the second term in equation (5.3) be positive for all values of, $\mathbf{x}(t)$, $\mathbf{u}(t)$ and $t$.

The first term in equation (5.3) represents a constraint on the value of the state vector at the final or terminal time. The closer the final state vector is to some desired value, the smaller is the value of the performance index.

Once the cost function and the constraint have been defined, the objective is to findthe optimal controller, i.e., the value of the control signal $\mathbf{u}(t)$ for the time interval $t_0 \leq t \leq t_1$ which provides the minimum value of $J$ under the assumption that the state vector obeys the state equation (5.1).

***Example 5.1.*** **A Typical Performance Index**

A typical regulator problem involves forcing the plant to stay at a stationary point, $\mathbf{r}_0$. If the plant state deviates from $\mathbf{r}_0$ the object of the regulator is to make it return as fast as possible and with as little overshoot as possible. Looking at Fig. 5.2 the value of an output $y(t)$ deviates at time 0 from the desired stationary point $r_0$. The time development of this output $y(t)$ is influenced by the controller and it is desired to bring it back to $r_0$ as rapidly as possible. An optimal control problem can be formulated by stating that the best controller is one that minimizes the shaded area in Fig. 5.2. The shaded area is a time integral of the form:

$$\int_0^{\infty} |y(t) - r_0| dt. \tag{5.4}$$

This is the form of index mentioned above and an optimal control problem therefore consists in finding a control signal that will minimize the above

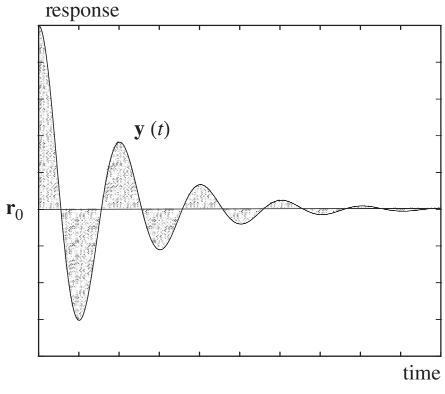


**Fig. 5.2** Sample regulator time response, here underdamped

performance index in (5.4) for the plant in question. As will be seen later the index has to be modified in part because in (5.4) there is no restriction on the size of the control signal and in part because the form of the index is inconvenient for calculations. These problems will be returned to after the solution of the general optimal control problem has been derived. ❐

## 5.3 The Basis of Optimal Control – Calculus of Variations

The calculus of variations is a general method for optimization of functions or functionals (functions of functions). Only a brief overview of this method will be given here. For a more in-depth treatment the reader is referred to more advanced control theory textbooks on the subject: see for example Bryson and Ho (1975).

A basic problem in the calculus of variations is the following:

A scalar integral which is a function of the time dependent vector $\mathbf{x}(t)$, its time derivative and the time is given:

$$J(\mathbf{x}) = \int_{t_0}^{t_1} F(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)dt, \tag{5.5}$$

where $F$ is a scalar function as is $J$ and $\mathbf{x}(t)$ is an $n$-dimensional vector whose elements are unconstrained functions of time.

The task is to determine that specific value of the vector $\mathbf{x}(t)$ which minimizes $J$ in time. Since $\mathbf{x}(t)$ is an $n$-dimensional vector, the task is actually to determine $n$ scalar time functions $x_1(t), x_2(t), \ldots, x_n(t)$ between the two time instants $t_0$ (the initial time) and $t_1$ (the final time).

The function $F$ is usually called a *loss function* and $J$ is called an optimization index or a *performance index*.

In order to find the required time functions it is necessary to know the *boundary conditions* for these functions.

Usually the initial condition is known or specified:

$$\mathbf{x}(t_0) = \mathbf{x}_0. \tag{5.6}$$

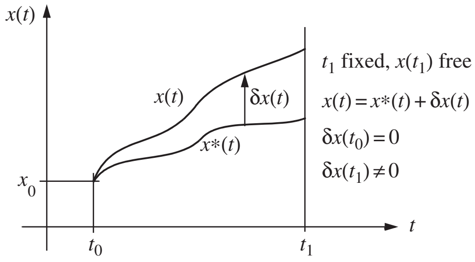
At the final time there are several possibilities:

1. $t_1$ is fixed and $\mathbf{x}(t_1)$ is fixed.
2. $t_1$ is fixed and $\mathbf{x}(t_1)$ is free.
3. $t_1$ is free and $\mathbf{x}(t_1)$ is fixed.
4. $\mathbf{x}(t)$ must satisfy certain constraint conditions at the final time.

To solve the problem it is convenient to introduce *variations*: perturbation functions close to the optimum function.

Suppose that the vector $\mathbf{x}^*(t)$ is the optimum vector that minimizes $J$. Figure 5.3 shows the situation in the case where $\mathbf{x}$ is a scalar. The optimum and a variational function are shown. For simplicity the boundary condition 2. above is selected. The optimum (minimum) value of $J$ can be written:

**Fig. 5.3** Time functions and variations



$$J^* = \int_{t_0}^{t^1} F(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t)dt. \tag{5.7}$$

By definition,

$$\Delta J = J - J^* > 0, \tag{5.8}$$

where

$$\begin{aligned}\Delta J &= \int_{t_0}^{t_1} F(\mathbf{x}(t), \dot{\mathbf{x}}(t), t)dt - \int_{t_0}^{t_1} F(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t)dt \\ &= \int_{t_0}^{t_1} F(\mathbf{x}^*(t) + \delta\mathbf{x}(t), \dot{\mathbf{x}}^*(t) + \dot{\delta}\mathbf{x}, t)dt - \int_{t_0}^{t_1} F(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t)dt.\end{aligned} \tag{5.9}$$

A Taylor series expansion of $F$ around the optimal solution yields:

$$\begin{aligned}F(\mathbf{x}(t), \dot{\mathbf{x}}(t), t) &= F(\mathbf{x}^*(t) + \delta\mathbf{x}(t), \dot{\mathbf{x}}^*(t) + \dot{\delta}\mathbf{x}(t), t) \\ &= F(\mathbf{x}^*(t), \dot{\mathbf{x}}^*(t), t) + \left.\frac{\partial F}{\partial \mathbf{x}}\right|_* \delta\mathbf{x}(t) + \left.\frac{\partial F}{\partial \dot{\mathbf{x}}}\right|_* \dot{\delta}\mathbf{x}(t) + \ldots .\end{aligned} \tag{5.10}$$

The *-notation on the partial derivatives means that the derivatives must be evaluated for $\mathbf{x}(t) = \mathbf{x}^*(t)$ and $\dot{\mathbf{x}}(t) = \dot{\mathbf{x}}^*(t)$. Note that only the first order expansion terms have been written out.

If this series expansion is inserted into equation (5.9) the result will be:

$$\Delta J = \int_{t_0}^{t_1} \left( \left.\frac{\partial F}{\partial \mathbf{x}}\right|_* \delta\mathbf{x}(t) + \left.\frac{\partial F}{\partial \dot{\mathbf{x}}}\right|_* \dot{\delta}\mathbf{x}(t) + \ldots \right) dt = \delta J + \ldots, \tag{5.11}$$

where $\delta J$ is called the *first variation* of $J$,

$$\delta J = \int_{t_0}^{t_1} \left( \left.\frac{\partial F}{\partial \mathbf{x}}\right|_* \delta\mathbf{x}(t) + \left.\frac{\partial F}{\partial \dot{\mathbf{x}}}\right|_* \dot{\delta}\mathbf{x}(t) \right) dt. \tag{5.12}$$

The last term of this equation is simplified using integration by parts:

$$\begin{aligned}\int_{t_0}^{t_1} \frac{\partial F}{\partial \dot{\mathbf{x}}}\bigg|_* \delta\dot{\mathbf{x}}(t)dt &= \frac{\partial F}{\partial \dot{\mathbf{x}}}\bigg|_* \delta\mathbf{x}(t)\big|_{t_0}^{t_1} - \int_{t_0}^{t_1} \frac{d}{dt}\left[\frac{\partial F}{\partial \dot{\mathbf{x}}}\bigg|_*\right]\delta\mathbf{x}(t)dt \\ &= \frac{\partial F}{\partial \dot{\mathbf{x}}}\bigg|_* \delta\mathbf{x}(t)\big|_{t_1} - \frac{\partial F}{\partial \dot{\mathbf{x}}}\bigg|_* \delta\mathbf{x}(t)\big|_{t_0} - \int_{t_0}^{t_1} \frac{d}{dt}\left[\frac{\partial F}{\partial \dot{\mathbf{x}}}\bigg|_*\right]\delta\mathbf{x}(t)dt.\end{aligned} \tag{5.13}$$

At the initial time $t = t_0$ the variation is zero, $\delta\mathbf{x}(t)|_{t_0} = \delta\mathbf{x}(t_0) = \mathbf{0}$, and therefore the second term in the last line of equation (5.13) disappears. If the remaining terms are inserted into (5.12), then the following equality will be satisfied,

$$\delta J = \int_{t_0}^{t_1} \left( \frac{\partial F}{\partial \mathbf{x}}\bigg|_* - \frac{d}{dt}\left[\frac{\partial F}{\partial \dot{\mathbf{x}}}\bigg|_*\right]\right)\delta\mathbf{x}(t)dt + \frac{\partial F}{\partial \dot{\mathbf{x}}}\bigg|_* \delta\mathbf{x}(t)|_{t_1} = \mathbf{0}, \tag{5.14}$$

because for the optimal solution, $\mathbf{x}(t) = \mathbf{x}^*(t)$, the variation $\delta J$ must be zero. Since (5.14) must hold for arbitrary $\delta\mathbf{x}(t)$ and $\delta\mathbf{x}(t_1)$ it is necessary to require that the integrand and the last term are zero:

$$\frac{\partial F(\mathbf{x}^*(t),\, \dot{\mathbf{x}}^*(t),\, t)}{\partial \mathbf{x}} - \frac{d}{dt}\left[\frac{\partial F(\mathbf{x}^*(t),\, \dot{\mathbf{x}}^*(t),\, t)}{\partial \dot{\mathbf{x}}}\right] = \mathbf{0} \tag{5.15}$$

and

$$\frac{\partial F(\mathbf{x}^*(t_1),\, \dot{\mathbf{x}}^*(t_1),\, t_1)}{\partial \dot{\mathbf{x}}} = \mathbf{0}. \tag{5.16}$$

Equation (5.15) is called the *Euler-Lagrange equation* and it must be satisfied by the optimal vector function $\mathbf{x}(t) = \mathbf{x}^*(t)$. The boundary condition (5.16) is called a *natural boundary condition*.

Note that not necessarily all functions which satisfy the Euler-Lagrange equation are optimal solutions. In other words, the condition is necessary but not sufficient. This fact has a parallel in the ordinary optimization of functions. It is necessary for an extremum that the function's derivative is zero but this condition is not sufficient. To ensure optimality it is also necessary to check the sign of the *second* derivative of the function. Similarly, one should investigate the *second variation* of (5.11). However this is cumbersome and it is thus omitted in many textbooks. This is also the case here.

The Euler-Lagrange equation is a very general result which forms an excellent basis for optimal control. However, two important modifications are necessary.

1. Constraints to be met by the optimal $\mathbf{x}(t)$ must be introduced.
2. A control input must be added.

The constraint(s) is (are) simply the system state equation(s). In optimal control the vector $\mathbf{x}(t)$ is the state vector in the state space model, equation (5.8):

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t),\, \mathbf{u}(t),\, t), \text{ with the initial condition, } \mathbf{x}(t_0) = \mathbf{x}_0, \tag{5.17}$$

where the control vector $\mathbf{u}(t)$ is $m$-dimensional.

Just as in static optimization (see Appendix A), the constraints are conveniently handled by introduction of *Lagrange multipliers*. In the dynamic case the multipliers are functions of time arranged as an $n$-dimensional vector usually denoted $\lambda(t)$.

The performance index will look a little different in the case of optimal control. To emphasize this, the change the symbol of the loss function is altered from $F$ to $L$ and the index becomes:

$$J(\mathbf{u}) = \int_{t_0}^{t_1} L(\mathbf{x}(t),\, \mathbf{u}(t),\, t)dt. \tag{5.18}$$

The index is of course still a function of $\mathbf{x}$ as well as of the newly introduced control vector $\mathbf{u}$. However in control the main task is to determine the *optimal input* $\mathbf{u}^*(t)$ and therefore it is more reasonable to add the extra argument $\mathbf{u}$ to the integrand. If $\mathbf{u}(t)$ is known, the state vector can be calculated from the state equation (5.17).

Now a useful trick will be utilized. The index is adjoined to (augmented with) the state equation arranged so that a zero is added to the performance index:

$$J(\mathbf{u}) = \int_{t_0}^{t_1} (L(\mathbf{x}(t),\, \mathbf{u}(t),\, t) + \lambda^T(t)(\mathbf{f}(\mathbf{x}(t),\, \mathbf{u}(t),\, t) - \dot{\mathbf{x}}(t)))dt. \tag{5.19}$$

This 'new' cost function (the integrand of (5.19)) is called $G$ here for reasons which will become clear a little later:

$$G = L(\mathbf{x}(t),\, \mathbf{u}(t),\, t) + \lambda^T(t)(\mathbf{f}(\mathbf{x}(t),\, \mathbf{u}(t),\, t) - \dot{\mathbf{x}}(t)). \tag{5.20}$$

The following augmented vectors are introduced:

$$\mathbf{z}(t) = \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{u}(t) \end{bmatrix}, \quad \dot{\mathbf{z}}(t) = \begin{bmatrix} \dot{\mathbf{x}}(t) \\ \dot{\mathbf{u}}(t) \end{bmatrix}, \tag{5.21}$$

where $\mathbf{z}(t)$ has dimension $n + m$.

This means that the index (5.19) becomes a function of $\mathbf{z}$ and $\dot{\mathbf{z}}$:

$$J(\mathbf{u}) = \int_{t_0}^{t_1} G(\mathbf{z}(t), \dot{\mathbf{z}}(t), t)dt. \tag{5.22}$$

The rationale behind these manipulations is that the index (5.22) has precisely the same form as (5.5). This means that the optimization problem with the *constraint* (5.17) has been reduced to the *unconstrained problem* of optimizing an expression like Eq. (5.5). Consequently the same Euler-Lagrange equation can be used to solve the augmented optimal control problem.

Equation (5.15) becomes (the $*$-notation is omitted for simplicity):

$$\frac{\partial G(\mathbf{z}(t), \dot{\mathbf{z}}(t), t)}{\partial \mathbf{z}} - \frac{d}{dt}\left[\frac{\partial G(\mathbf{z}(t), \dot{\mathbf{z}}(t), t)}{\partial \dot{\mathbf{z}}}\right] = \mathbf{0} \tag{5.23}$$

where the first term is

$$\frac{\partial G}{\partial \mathbf{z}} = \begin{bmatrix} \dfrac{\partial G}{\partial z_1} \\ \vdots \\ \dfrac{\partial G}{\partial z_n} \\ \dfrac{\partial G}{\partial z_{n+1}} \\ \vdots \\ \dfrac{\partial G}{\partial z_{n+m}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial G}{1} \\ \vdots \\ \dfrac{\partial G}{\partial x_n} \\ \dfrac{\partial G}{\partial u_1} \\ \vdots \\ \dfrac{\partial G}{\partial u_m} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial G}{\partial \mathbf{x}} \\ \dfrac{\partial G}{\partial \mathbf{u}} \end{bmatrix}. \tag{5.24}$$

If the second term is handled similarly, the Euler-Lagrange equation can be written:

$$\begin{bmatrix} \dfrac{\partial G}{\partial \mathbf{x}} \\ \dfrac{\partial G}{\partial \mathbf{u}} \end{bmatrix} - \frac{d}{dt}\begin{bmatrix} \dfrac{\partial G}{\partial \dot{\mathbf{x}}} \\ \dfrac{\partial G}{\partial \dot{\mathbf{u}}} \end{bmatrix} = \mathbf{0}. \tag{5.25}$$

According to (5.20) the four partial derivatives in equation (5.25) can be written:

$$\begin{aligned} \frac{\partial G}{\partial \mathbf{x}} &= \frac{\partial L}{\partial \mathbf{x}} + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right)^T \lambda, \\ \frac{\partial G}{\partial \mathbf{u}} &= \frac{\partial L}{\partial \mathbf{u}} + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right)^T \lambda, \\ \frac{\partial G}{\partial \dot{\mathbf{x}}} &= -\lambda, \\ \frac{\partial G}{\partial \dot{\mathbf{u}}} &= \mathbf{0}. \end{aligned} \tag{5.26}$$

With the expressions in equation (5.26) the Euler-Lagrange equation (5.25) becomes

$$\frac{\partial L}{\partial \mathbf{x}} + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right)^T \lambda + \dot{\lambda} = \mathbf{0}, \tag{5.27}$$

$$\frac{\partial L}{\partial \mathbf{u}} + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right)^T \lambda = \mathbf{0}. \tag{5.28}$$

Note that (5.27) contains $n$ first order differential equations whereas (5.28) is a set of $m$ algebraic equations. If to this the state equation (5.17) (the constraint) is added, one ends up with a total of $n$ first order differential equations and $m$ algebraic equations which must be solved simultaneously.

The natural boundary condition in equation (5.16) will be:

$$\left.\frac{\partial G}{\partial \dot{\mathbf{z}}}\right|_{t_1} = \begin{bmatrix} \dfrac{\partial G}{\partial \dot{\mathbf{x}}} \\ \dfrac{\partial G}{\partial \dot{\mathbf{u}}} \end{bmatrix}_{t_1} = \begin{bmatrix} -\lambda(t_1) \\ \mathbf{0} \end{bmatrix} = \mathbf{0}. \tag{5.29}$$

The usable parts of this expression are the $n$ conditions,

$$\lambda(t_1) = \mathbf{0}. \tag{5.30}$$

The remaining $n$ boundary conditions necessary for solving the $2n$ differential equations are the initial conditions for the state equation,

$$\mathbf{x}(t_0) = \mathbf{x}_0. \tag{5.31}$$

From (5.30) and (5.31) it may be seen that half of the boundary conditions are valid at the initial time $t_0$ and the rest are valid at the final time $t_1$. This is called a *two point boundary value problem*. It causes serious problems in working with optimal control problems since it prevents direct solution of the set of equations by analytical or numerical means.

So far the final value term in equation (5.3) has been omitted. To include this term in the investigation the following is noted:

$$\Phi(\mathbf{x}(t_1), t_1) = \Phi(\mathbf{x}(t_1), t_1) + \Phi(\mathbf{x}(t_0), t_0) - \Phi(\mathbf{x}(t_0), t_0). \tag{5.32}$$

Then the performance index of equation (5.3) can be written:

$$J(\mathbf{u}) = \int_{t_0}^{t_1} \frac{d}{dt}(\Phi(\mathbf{x}(t), t))dt + \int_{t_0}^{t_1} L(\mathbf{x}(t), \mathbf{u}(t), t)dt + \Phi(\mathbf{x}(t_0), t_0), \tag{5.33}$$

where the last term is a known constant and consequently it has no influence on the optimization problem. It can therefore be ignored.

The total differential of the scalar function $\Phi(\mathbf{x}(t), t)$ of the $n$ states is,

$$d\Phi(\mathbf{x}(t), t) = \frac{\partial \Phi}{\partial x_1} dx_1 + \frac{\partial \Phi}{\partial x_2} dx_2 + \ldots + \frac{\partial \Phi}{\partial x_n} dx_n. \tag{5.34}$$

The derivative with respect to time is then:

$$\frac{d\Phi(\mathbf{x}(t), t)}{dt} = \frac{\partial \Phi}{\partial x_1} \dot{x}_1 + \frac{\partial \Phi}{\partial x_2} \dot{x}_2 + \ldots + \frac{\partial \Phi}{\partial x_n} \dot{x}_n = \left(\frac{\partial \Phi}{\partial \mathbf{x}}\right)^T \dot{\mathbf{x}}. \tag{5.35}$$

Equation (5.35) is inserted into the performance index (5.33) and same term is added as in (5.20) (the last term is omitted as explained above),

$$\int_{t_0}^{t_1} \left( \left(\frac{\partial \Phi}{\partial \mathbf{x}}\right)^T \dot{\mathbf{x}} + L + \lambda^T (\mathbf{f} - \dot{\mathbf{x}}) \right) dt, \tag{5.36}$$

with the cost function,

$$G_\Phi = \left(\frac{\partial \Phi}{\partial \mathbf{x}}\right)^T \dot{\mathbf{x}} + L + \lambda^T (\mathbf{f} - \dot{\mathbf{x}}), \tag{5.37}$$

the Euler-Lagrange equation can be applied once more:

$$\begin{bmatrix} \dfrac{\partial G_\Phi}{\partial \mathbf{x}} \\ \dfrac{\partial G_\Phi}{\partial \mathbf{u}} \end{bmatrix} - \frac{d}{dt} \begin{bmatrix} \dfrac{\partial G_\Phi}{\partial \dot{\mathbf{x}}} \\ \dfrac{\partial G_\Phi}{\partial \dot{\mathbf{u}}} \end{bmatrix} = 0. \tag{5.38}$$

Calculating the partial derivatives leads to,

$$\begin{aligned} \frac{\partial G_\Phi}{\partial \mathbf{x}} &= \frac{\partial L}{\partial \mathbf{x}} + \frac{\partial^2 \Phi}{\partial \mathbf{x}^2} \dot{\mathbf{x}} + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right)^T \lambda, \\ \frac{\partial G_\Phi}{\partial \mathbf{u}} &= \frac{\partial L}{\partial \mathbf{u}} + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right)^T \lambda, \\ \frac{\partial G_\Phi}{\partial \dot{\mathbf{x}}} &= \frac{\partial \Phi}{\partial \mathbf{x}} - \lambda \;\Rightarrow\; \frac{d}{dt}\left(\frac{\partial G_\Phi}{\partial \dot{\mathbf{x}}}\right) = \frac{\partial^2 \Phi}{\partial \mathbf{x}^2} \dot{\mathbf{x}} - \dot{\lambda}, \\ \frac{\partial G_\Phi}{\partial \dot{\mathbf{u}}} &= 0 \;\Rightarrow \frac{d}{dt}\left(\frac{\partial G_\Phi}{\partial \dot{\mathbf{u}}}\right) = \mathbf{0}. \end{aligned} \tag{5.39}$$

If these expressions are used in equation (5.38) the Euler-Lagrange equations become:

$$\frac{\partial L}{\partial \mathbf{x}} + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right)^T \lambda + \dot{\lambda} = \mathbf{0}, \tag{5.40}$$

$$\frac{\partial L}{\partial \mathbf{u}} + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right)^T \lambda = \mathbf{0}, \tag{5.41}$$

which is exactly the same as achieved in Eqs. (5.27) and (5.28) *without* the final state term in the performance index.

The only new change appears in the natural boundary condition (5.16),

$$\left.\frac{\partial G_\Phi}{\partial \dot{\mathbf{x}}}\right|_{t_1} = 0 \;\Rightarrow\; \left.\left(\frac{\partial \Phi}{\partial \mathbf{x}} - \lambda\right)\right|_{t_1} = \mathbf{0}, \tag{5.42}$$

or

$$\lambda(t_1) = \frac{\partial \Phi(\mathbf{x}(t_1), t_1)}{\partial \mathbf{x}(t_1)}. \tag{5.43}$$

Note the difference from equation (5.30).

The last step in this derivation has the purpose of avoiding the term $\mathbf{f} - \dot{\mathbf{x}}$ in Eqs. (5.36) and (5.37). The *Hamilton function H* is introduced (it is not a function of $\dot{\mathbf{x}}$):

$$H(\mathbf{x}(t), \lambda(t), \mathbf{u}(t), t) = L(\mathbf{x}, \mathbf{u}, t) + \lambda^T(t) \cdot \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t), \tag{5.44}$$

which gives the performance index (5.19) the following appearance:

$$J(\mathbf{u}) = \Phi(\mathbf{x}(t_1), t_1) + \int_{t_0}^{t_1} [H(\mathbf{x}(t), \lambda(t), \mathbf{u}(t), t) - \lambda^T(t) \cdot \dot{\mathbf{x}}(t)] dt. \tag{5.45}$$

The partial derivatives of $H$ are:[†]

$$\begin{aligned}
\frac{\partial H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \lambda(t)} &= \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) = \dot{\mathbf{x}}(t), \\
\frac{\partial H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \mathbf{x}(t)} &= -\dot{\lambda}(t), \\
\frac{\partial H(\mathbf{x}(t), \mathbf{u}(t), \lambda(t), t)}{\partial \mathbf{u}(t)} &= 0.
\end{aligned} \tag{5.46}$$

[†] The gradient of a scalar function $f(\mathbf{x})$ is a column vector $\nabla f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1} \frac{\partial f}{\partial x_2} \cdots \frac{\partial f}{\partial x_n}\right]^T$. The gradient of $f$ with respect to $\mathbf{x}$ is denoted as $\frac{\partial f}{\partial \mathbf{x}}$ and $f_{\mathbf{x}}$. See Appendix B for further details on vector calculus.

Now each of these terms will be considered in turn to show what they represent physically.

Referring to the expression for $H$ in Eq. (5.44) it is seen that the first equation in (5.46) is *the equation of motion* of the system or *the state equation*. The $n$-dimensional Lagrange multiplier vector $\lambda$ is also called the *co-state vector* and therefore the second equation is called *the co-state equation*:

$$\frac{\partial H(\mathbf{x},\mathbf{u},\lambda,t)}{\partial \mathbf{x}} = \frac{\partial L(\mathbf{x},\mathbf{u},t)}{\partial \mathbf{x}} + \left(\frac{\partial \mathbf{f}(\mathbf{x},\mathbf{u},t)}{\partial \mathbf{u}}\right)^T \lambda = -\dot{\lambda}(t). \tag{5.47}$$

The third and last equation is called *the stationarity equation* and normally gives the relation between the control signal and the Lagrange multiplier $\lambda$:

$$\frac{\partial H(\mathbf{x},\mathbf{u},\lambda,t)}{\partial \mathbf{u}} = \frac{\partial L(\mathbf{x},\mathbf{u},t)}{\partial \mathbf{u}} + \left(\frac{\partial \mathbf{f}(\mathbf{x},\mathbf{u},t)}{\partial \mathbf{u}}\right)^T \lambda = 0. \tag{5.48}$$

The Hamiltonian $H$ plays a special role in the calculations. If $H$ is not an explicit function of time, i.e., if neither $L$ nor $\mathbf{f}$ explicitly depend on $t$, then the Hamiltonian is a constant over time. This can be demonstrated by differentiation. By the chain rule for differentiation,

$$\frac{dH}{dt} = H_{\mathbf{x}}^T\dot{\mathbf{x}} + H_{\mathbf{u}}^T\dot{\mathbf{u}} + H_{\lambda}^T\dot{\lambda} + \frac{\partial H}{\partial t}. \tag{5.49}$$

Introducing the results in (5.47) and (5.48) one obtains

$$\frac{dH}{dt} = -\dot{\lambda}^T\dot{\mathbf{x}} + \mathbf{0}\dot{\mathbf{u}} + \dot{\mathbf{x}}^T\dot{\lambda} + \frac{\partial H}{\partial t} = \frac{\partial H}{\partial t}. \tag{5.50}$$

This remaining term is zero if $H$ is not an explicit function of time.

## 5.4 The Linear Quadratic Regulator

In general the non-linear optimization problem discussed in the previous sections cannot be solved analytically and the optimal control signal will have to be found numerically. This of course severely limits the usefulness and convenience of the general optimization theory, which is why the controller design is normally based on a linearized state space model of the plant.

The plant which is to be investigated is therefore a linearized one but may be time varying. Here the states and inputs are the incremental ones though they will be written as $\mathbf{x}(t)$ and $\mathbf{u}(t)$ here for the sake of simplicity:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t). \tag{5.51}$$

In general the cost function is also normally restricted to have a certain particularly simple structure which will now be motivated.

### 5.4.1 The Quadratic Cost Function

In the time-domain the performance of a controller is judged by its ability to follow transient changes in an input, its ability to suppress disturbances and its ability to limit or eliminate stationary errors. For example the controller performs well if the step response rise time is short, the overshoot limited and the settling time to a small stationary error is short. As depicted in Fig. 5.4 one can arrive at this goal if the cross hatched area in the figure is minimized. If the input step is $r(t)$ and the response is termed $y(t)$, then this area is given by the time integral

$$\int_0^\infty |r(t) - y(t)|dt.$$

This particular performance index is a valid function to minimize because it is limited by zero from below. However, from a computational point of view it is easier to use a quadratic performance index of the form

$$\int_0^\infty (r(t) - y(t))^2 dt.$$

This has many of the same properties as the index above and also in a more general framework, where $y(t)$ is replaced by a general state vector, $\mathbf{x}(t)$, this expression is often related to the energy of the system, e.g. its kinetic and potential energy.

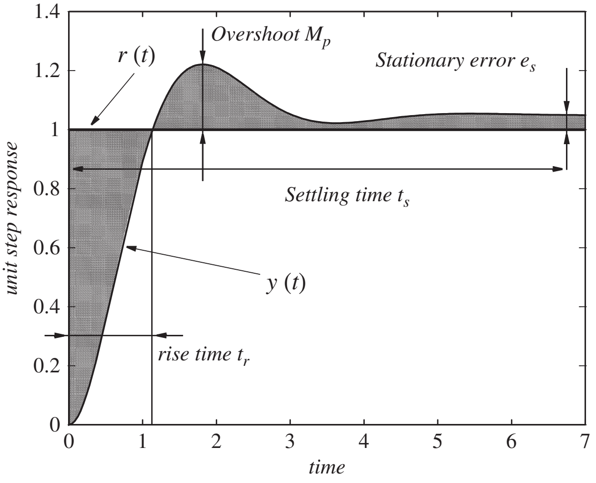


**Fig. 5.4** Typical transient response of a controller

From the integral above it is obvious that the optimal control signal is one that gives $y(t) = r(t)$ for all $t$. However, this can only be achieved if the system is made to change infinitely fast and therefore the control signal has to be infinite. Obviously, for any physical system the cost function has to be modified to take into account the fact that the control signal to the plant is limited in size and bandwidth. This can be achieved by adding a term to the cost function which is quadratic in the control signal $\mathbf{u}(t)$.

Initially, the regulator problem is considered, so the state is required to be close to a stationary operating point. Since the plant will be linearized around this operating point, the deviations in the state vector that should be added to the cost function are deviations from the stationary point. This is zero in the linearized model. This leads to an index of the form (written in terms of the incremental states and inputs):

$$J = \int_{t_0}^{t_1} [\mathbf{x}^T(t)\mathbf{R}_1\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}_2\mathbf{u}(t)]dt.$$

The integral is assumed to have upper bound $t_1 < \infty$, but later the special but important case of an infinite upper bound will be considered. Before the individual terms are discussed in detail, a final time state cost term is to be included as a quadratic term. This then gives a *quadratic performance index* of the form

$$J(\mathbf{u}) = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}(t_1)\mathbf{x}(t_1) + \frac{1}{2}\int_{t_0}^{t_1} [\mathbf{x}^T\mathbf{R}_1(t)\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}_2(t)\mathbf{u}(t)]dt. \tag{5.52}$$

The matrices $\mathbf{S}(t_1)$, $\mathbf{R}_1(t)$ and $\mathbf{R}_2(t)$ are called the weight matrices and determine how much deviations of $\mathbf{x}(t_1)$, $\mathbf{x}(t)$ and $\mathbf{u}(t)$ from their zeroes will add to the overall cost function. A necessary condition for $J$ to have a minimum is that it is bounded from below. Therefore that all terms in the index are non-negative for all values of $\mathbf{x}$ and $\mathbf{u}$. This will be satisfied if $\mathbf{S}(t_1)$ and $\mathbf{R}_1(t)$ are positive semidefinite for all values of $t$. Furthermore, since $\mathbf{u}(t)$ should be bounded for all $t$ it is necessary that $\mathbf{R}_2(t)$ have the stronger property of being positive definite, i.e.,

$$\begin{aligned} \mathbf{S}(t_1) &\geq 0, \quad \forall t, \\ \mathbf{R}_1(t) &\geq 0, \quad \forall t, \\ \mathbf{R}_2(t) &> 0, \quad \forall t. \end{aligned} \tag{5.53}$$

The three terms in (5.52) are quadratic forms. See Appendix B.4 for more properties of quadratic forms.

The weight matrices will determine the influence of individual components of the state vector or input vector relative to each other. For example if the element $[\mathbf{R}_1]_{ij}$ is large, then the corresponding product of state vector elements $x_i x_j$ will be penalized heavily in the cost function and the resulting optimal control law will tend to emphasize making that term small. In particular the relative sizes of the quadratic forms $\mathbf{x}^T(t)\mathbf{R}_1(t)\mathbf{x}(t)$ and $\mathbf{u}^T(t)\mathbf{R}_2(t)\mathbf{u}(t)$ will determine the speed of the control system. If $\mathbf{R}_1(t)$ is selected such that the first term is large compared to the second term the system states will tend to respond faster at the cost of increasing the control signal. In the opposite case the inputs are forced to remain small. As a consequence the response will be slow and the overall deviation of the state vector from stationary state (i.e., zero) will be larger. In this way the relative weights of the two matrices can be used to tune the response speed at the same time limiting of the size of the control signal.

The final state term takes into account the fact that it may not be possible to reach the state zero exactly. The weight matrix $\mathbf{S}(t_1)$ penalizes errors in the final state.

If it is known what the maximum sizes of the final initial states, continuing states and continuing inputs are, then the following general rule can be given for the selection of the weighting matrices:

$$[S(t)]_{ii} = \frac{1}{max([x_i(t_1)]^2)}, \tag{5.54}$$

$$[R_1(t)]_{ii} = \frac{1}{(t_1 - t_0) \cdot max([x_i(t)]^2)}, \tag{5.55}$$

$$[R_2(t)]_{jj} = \frac{1}{(t_1 - t_0) \cdot max([u_i(t)]^2)}, \tag{5.56}$$

where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. If the time is not important in the intended application then the time interval in the parenthesis, $(t_1 - t_0)$, may be set equal to 1. Cross product terms may be used in the weighting matrices if there is interaction among the input or state components.

## *5.4.2 Linear Quadratic Control*

In this section what is perhaps the most important modern LQR controller will be presented. This is a MIMO linear closed loop controller. The system is assumed to be linear but possibly time varying, i.e., it is described by the state equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t). \tag{5.57}$$

The controller is required to minimize a performance index of the form (5.52).

The final time is fixed but now the final state is allowed to deviate from zero. It is inserted into the performance index with a positive semi-definite weight matrix $\mathbf{S}(t_1)$,

$$\mathbf{S}(t_1) \geq 0.$$

The state and control signal weight matrices are assumed to be positive semi-definite and positive definite respectively:

$$\mathbf{R}_1(t) \geq 0 \text{ and } \mathbf{R}_2(t) > 0, \ \forall t.$$

The boundary condition is given by equation (5.43). Since $\Phi(\mathbf{x}(t_1), t_1) = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}(t_1)\mathbf{x}(t_1)$ this gives

$$\lambda(t_1) = \mathbf{S}(t_1)\mathbf{x}(t_1). \tag{5.58}$$

The Hamiltonian of the system is

$$H = \frac{1}{2}\mathbf{x}^T(t)\mathbf{R}_1(t)\mathbf{x}(t) + \frac{1}{2}\mathbf{u}^T(t)\mathbf{R}_2(t)\mathbf{u}(t) + \lambda^T(t)(\mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t)). \tag{5.59}$$

Hence the co-state equation becomes

$$\dot{\lambda}(t) = -\frac{\partial H}{\partial \mathbf{x}(t)} = -\mathbf{R}_1(t)\mathbf{x}(t)\mathbf{x}(t) - \mathbf{A}^T(t)\lambda(t) \tag{5.60}$$

and the stationarity equation is

$$\mathbf{0} = \frac{\partial H}{\partial \mathbf{u}(t)} = \mathbf{R}_2(t)\mathbf{u}(t) + \mathbf{B}^T(t)\lambda(t) \Longrightarrow \mathbf{u}(t) = -\mathbf{R}_2^{-1}(t)\mathbf{B}^T(t)\lambda(t). \tag{5.61}$$

The state equation and the co-state equation are two differential equations in the state and co-state variables with the initial condition that the state vector starts in $\mathbf{x}(t_0) = \mathbf{x}_0$ and that at time $t_1$ the co-state obeys equation (5.58). In general, solving these equations directly is difficult because of the two point boundary value problem. However in the linear-quadratic case it is possible to employ a trick which will make it possible to get around this problem.

If equation (5.61) is inserted into the state equation (5.57) then this equation and (5.60) can be expressed as

$$\begin{bmatrix} \dot{\mathbf{x}}(t) \\ \dot{\lambda}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(t) & -\mathbf{B}(t)\mathbf{R}_2^{-1}(t)\mathbf{B}^T(t) \\ -\mathbf{R}_1(t) & -\mathbf{A}^T(t) \end{bmatrix} \begin{bmatrix} \mathbf{x}(t) \\ \lambda(t) \end{bmatrix} = \mathbf{H}(t) \begin{bmatrix} \mathbf{x}(t) \\ \lambda(t) \end{bmatrix}. \tag{5.62}$$

This is called the *Hamilton equation* and the matrix $\mathbf{H}(t)$ is called the *Hamiltonian*.

The solution to this unforced $2n$-dimensional state equation is, according to (3.20),

$$\begin{bmatrix} \mathbf{x}(t) \\ \lambda(t) \end{bmatrix} = \phi(t, t_0) \begin{bmatrix} \mathbf{x}(t_0) \\ \lambda(t_0) \end{bmatrix} = \begin{bmatrix} \phi_1(t, t_0) & \phi_2(t, t_0) \\ \phi_3(t, t_0) & \phi_4(t, t_0) \end{bmatrix} \begin{bmatrix} \mathbf{x}(t_0) \\ \lambda(t_0) \end{bmatrix}, \tag{5.63}$$

where the state transition matrix has been partitioned into $4\, n \times n$ matrices. Applying the property (3.24), the solution can also be written

$$\begin{bmatrix} \mathbf{x}(t) \\ \lambda(t) \end{bmatrix} - \begin{bmatrix} \phi_1(t, t_1) & \phi_2(t, t_1) \\ \phi_3(t, t_1) & \phi_4(t, t_1) \end{bmatrix} \begin{bmatrix} \mathbf{x}(t_1) \\ \lambda(t_1) \end{bmatrix}. \tag{5.64}$$

Using (5.58) and eliminating $\mathbf{x}(t_1)$ from the two equations in (5.64) leads to

$$\lambda(t) = (\phi_3(t, t_1) + \phi_4(t, t_1)\mathbf{S}(t_1)) \cdot (\phi_1(t, t_1) + \phi_2(t, t_1)\mathbf{S}(t_1))^{-1}\mathbf{x}(t), \tag{5.65}$$

which can be written,

$$\lambda(t) = \mathbf{P}(t)\mathbf{x}(t). \tag{5.66}$$

The matrix $\mathbf{P}(t)$ is obviously a function of the constant final time $t_1$ as well as of $t$, so it would be more correct to write $\mathbf{P}$ as $\mathbf{P}(t,\ t_1)$. However, it is common practice to omit the final time as an argument as seen in Eq (5.66).

Equation (5.58) shows that

$$\mathbf{P}(t_1) = \mathbf{S}(t_1). \tag{5.67}$$

The control signal will be given by equation (5.61)

$$\mathbf{u}(t) = -\mathbf{R}_2^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t)\mathbf{x}(t). \tag{5.68}$$

Equation (5.68) shows that the control vector is derived from the state vector. In other words, a *closed loop control* has been established which is very convenient from an applications point of view.

The remaining problem is to determine the matrix $\mathbf{P}(t)$. This matrix must obey a differential equation that follows from differentiating (5.66) with respect to time,

$$\dot{\lambda}(t) = \dot{\mathbf{P}}(t)\mathbf{x}(t) + \mathbf{P}(t)\dot{\mathbf{x}}(t). \tag{5.69}$$

Inserting the state and co-state equations and using equation (5.66) yields

$$\begin{aligned} &- \mathbf{R}_1(t)\mathbf{x}(t) - \mathbf{A}^T(t)\mathbf{P}(t)\mathbf{x}(t) \\ &= \dot{\mathbf{P}}(t)\mathbf{x}(t) + \mathbf{P}(t)[\mathbf{A}(t)\mathbf{x}(t) - \mathbf{B}(t)\mathbf{R}_2^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t)\mathbf{x}(t)]. \end{aligned} \tag{5.70}$$

This equation has a solution for all $\mathbf{x}(t)$ if $\mathbf{P}(t)$ obeys the differential equation:

$$-\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{A}(t) + \mathbf{A}^T(t)\mathbf{P}(t) - \mathbf{P}(t)\mathbf{B}(t)\mathbf{R}_2^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t) + \mathbf{R}_1(t). \tag{5.71}$$

This important differential equation is known as the *Riccati equation*. The relevant boundary conditions are given by equation (5.67).

The Riccati equation is a coupled set of $n^2$ first order non-linear differential equations, defined on the interval $t_0 \le t \le t_1$ with $n^2$ boundary conditions at the final time $t_1$. This equation has to be solved backwards in time starting at time $t_1$. Under certain very general conditions the equation has one unique solution. For an analysis of this point the reader is referred to Kwakernaak and Sivan (1972) and Bryson and Ho (1975). $\mathbf{S}(t_1)$ is symmetric and because the Riccati equation is also symmetric, the solution will be symmetric for all values of $t$ (see Problem 5.17). This means that the number of equations in equation (5.71) is reduced to $n(n+1)/2$.

The conclusion from the above is that under fairly loose conditions the Riccati equation will have a real solution and from this a controller can be designed that will minimize the quadratic performance index. The existence of a solution to the Riccati equation does not require the system to be controllable. Even with the loss of full controllability the controller will attempt to minimize the performance index. However it is likely that it will be more capable if the system is controllable.

Observing that in general the Riccati equation has a unique solution, it can be concluded that the optimal controller for the linear quadratic regulator problem has a unique solution in the form of a control signal that is a state feedback controller:

$$\mathbf{u}(t) = -\mathbf{K}(t)\mathbf{x}(t). \tag{5.72}$$

The time-dependent feedback gain $\mathbf{K}(t)$ is called the *LQR* (*Linear Quadratic Regulator*) gain or optimal regulator gain and can be inferred from equation (5.68) to be:

$$\mathbf{K}(t) = \mathbf{R}_2^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t). \tag{5.73}$$

Figure 5.5 shows the state feedback block diagram for the LQ regulator. It is seen that in general the LQR gain will be time-dependent even when the system

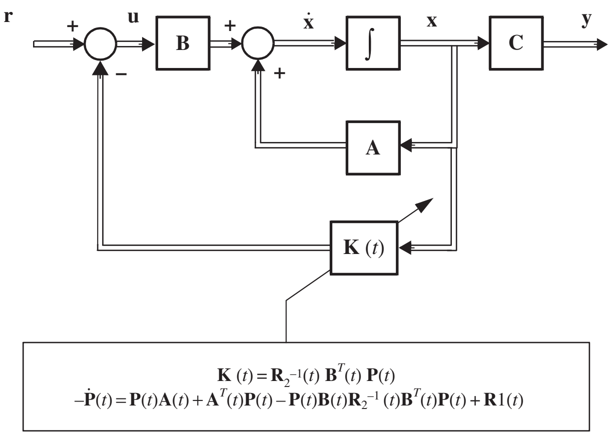

r
+
u
−
B
+
ẋ
+
∫
x
C
y
A
K (t)
K (t) = R₂⁻¹(t) Bᵀ(t) P(t)
−Ṗ(t) = P(t)A(t) + Aᵀ(t)P(t) − P(t)B(t)R₂⁻¹ (t)Bᵀ(t)P(t) + R1(t)

**Fig. 5.5** Closed loop linear quadratic regulator

is LTI and the cost function has constant weight matrices. It should also be noted that the LQR controller is a linear state feedback controller and that the LQR gain is solely dependent on parameters known in advance and can therefore be calculated off-line.

Having obtained the optimal control signal, the closed loop system becomes

$$\dot{\mathbf{x}}(t) = \mathbf{A}_c(t)\mathbf{x}(t) = (\mathbf{A}(t) - \mathbf{B}(t)\mathbf{K}(t))\mathbf{x}(t). \tag{5.74}$$

The optimal regulator's eigenfrequencies are thus determined in the same way as any state feedback controller and it has exactly the same structure.

Inserting the optimal gain into the Riccati equation it can be seen that it can be written in the so-called *Josephson stabilized form*:

$$-\dot{\mathbf{P}} = \mathbf{P}(\mathbf{A} - \mathbf{BK}) + (\mathbf{A} - \mathbf{BK})^T\mathbf{P} + \mathbf{K}^T\mathbf{R}_2\mathbf{K} + \mathbf{R}_1. \tag{5.75}$$

The value of the performance index can be evaluated based on the solution to the Riccati equation. First note that

$$\frac{d}{dt}(\mathbf{x}^T\mathbf{P}\mathbf{x}) = \dot{\mathbf{x}}^T\mathbf{P}\mathbf{x} + \mathbf{x}^T\dot{\mathbf{P}}\mathbf{x} + \mathbf{x}^T\mathbf{P}\dot{\mathbf{x}}. \tag{5.76}$$

If equations (5.57), (5.68) and (5.71) are inserted into (5.76), one finds that

$$\frac{d}{dt}(\mathbf{x}^T\mathbf{P}\mathbf{x}) = -\mathbf{x}^T\mathbf{R}_1\mathbf{x} - \mathbf{u}^T\mathbf{R}_2\mathbf{u}. \tag{5.77}$$

Inserting this into the index leads to

$$\begin{aligned} J(\mathbf{u}) &= \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}(t_1)\mathbf{x}(t_1) + -\frac{1}{2}\int_{t_0}^{t_1}\left[\frac{d}{dt}(\mathbf{x}(t)^T\mathbf{P}(t)\mathbf{x}(t))\right]dt \\ &= \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}(t_1)\mathbf{x}(t_1) - \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{P}(t_1)\mathbf{x}(t_1) + \frac{1}{2}\mathbf{x}^T(t_0)\mathbf{P}(t_0)\mathbf{x}(t_0). \end{aligned}$$

The first two terms disappear because of (5.67) and the optimum (minimum) value of the performance index can be found:

$$J_{min} = \frac{1}{2}\mathbf{x}^T(t_0)\mathbf{P}(t_0)\mathbf{x}(t_0). \tag{5.78}$$

***Example 5.2*. LQ Regulator for a First Order System**

To show the basic characteristics of LQR regulators a controller for a simple first order system will be designed first. The scalar state equation is,

$$\dot{x}(t) = ax(t) + bu(t),$$

where $a$ and $b$ are constants.

The performance index is the following quadratic function:

$$J = \frac{1}{2}s(t_1)x^2(t_1) + \frac{1}{2}\int_0^{t_1}(r_1x^2(t) + r_2u^2(t))dt.$$

The boundedness of the performance index dictates that $s(t_1) \geq 0$, $r_1 \geq 0$ and $r_2 > 0$. Without any loss of generality it may be assumed that the system starts at time 0 (since the system is time invariant). The system starts with the state variable $x(0) = x_0$. The Riccati equation for the control object is the following first order ordinary differential equation,

$$-\dot{p}(t) = 2ap(t) + r_1 - \frac{b^2}{r_2}p^2(t).$$

The solution to this equation can be inserted into equations (5.72) and (5.73) for the LQR gain to give the state feedback controller:

$$u(t) = -K(t)x(t) = -bp(t)x(t).$$

In this simple example the solution of the Riccati equation can be found analytically. Rearranging the terms in the equation and integrating gives:

$$\frac{dp}{dt} = \frac{b^2}{r_2}p^2(t) - 2ap(t) - r_1 \Rightarrow \int_{p(t)}^{p(t_1)} \frac{dp}{\frac{b^2}{r_2}p^2 - 2ap - r_1} = \int_t^{t_1} d\tau.$$

Integrating both sides of the equation on the right and rearranging the terms yields

$$p(t) = p_2 - \frac{p_2 - p_1}{1 - \dfrac{p(t_1) - p_1}{p(t_1) - p_2}e^{\frac{b^2}{r_2}(p_2 - p_1)(t_1 - t)}}. \tag{5.79}$$

Here

$$p_1 = \frac{a}{b^2} + \frac{1}{b}\sqrt{\frac{a^2}{b^2} + \frac{r_1}{r_2}}$$

and

$$p_2 = \frac{a}{b^2} - \frac{1}{b}\sqrt{\frac{a^2}{b^2} + \frac{r_1}{r_2}}.$$

Thus even in the simple scalar case the controller feedback becomes a relatively complicated time varying function. However, since the solution only depends on parameters known prior to controlling the process, the solution can be calculated off-line and stored in the form of a table, so the online computational load is limited.

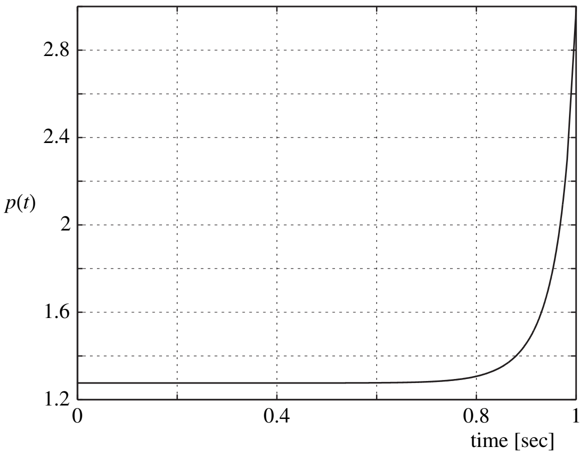


**Fig. 5.6** Solution to the Riccati equation in the case of a scalar system

Given that $a = 3, b = 3, r_1 = 7, r_2 = 1, s(t_1) = 3$ and choosing the final time $t_1 = 1$ the solution to the Riccati equation is obtained as shown on Fig. 5.6. The function ends at time $t_1 = 1$ at the final value $p(t_1) = s(t_1)$. If the time going backwards from $t_1$ to zero is considered, $p$ decreases to a constant value fairly rapidly. This stationary value is found from the equation by letting $t_1 - t$ grow 'large'. Since $p_2 - p_1 < 0$, the exponential function in (5.77) tends to zero for large values of $t_1 - t$ and consequently $p$ converges to the value

$$p = p_1 = \frac{a}{b^2} + \frac{1}{b}\sqrt{\frac{a^2}{b^2} + \frac{r_1}{r_2}} = 1.276.$$

The state $x(t)$ for four different values of $r_1$ and for $x_0 = 5$ is plotted in Fig. 5.7. The control signal is shown on Fig. 5.8. It is noted that the response becomes

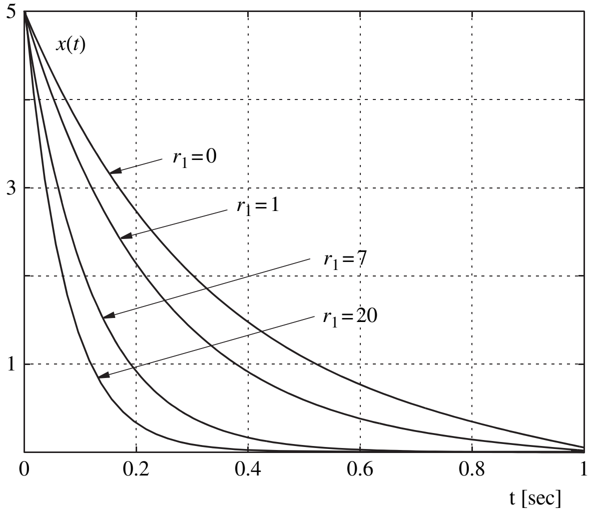


**Fig. 5.7** Response of the first order system for different weights

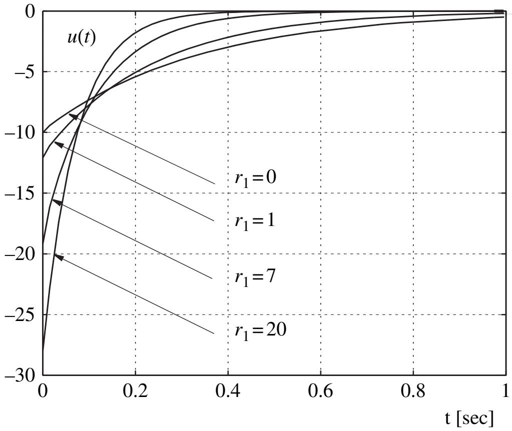


**Fig. 5.8** Control signal of the first order system for different state weights

faster the larger the value of $r_1$ and the price to be paid is that the control signal becomes larger. In the case $r_1 = 0$ the state is not taken into account at all in the performance index.

It should also be noted that the LQR gain is approximately proportional to $\sqrt{r_1/r_2}$ for large values of $t_1 - t$. This is generally true for LQR regulators and is a useful rule of thumb.

Since $a$ is positive, the open loop system is unstable. This is no problem when an LQR regulator is used. The closed loop system found with the LQR methodology will always be stable under the proper conditions. ❐

### *Example 5.3*. Closed Loop LQR for a Double Integrator

Consider now the double integrator described by the state equation:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} p(t) \\ v(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t).$$

The state vector here is comprised of the position $p(t)$ and the velocity $\nu(t)$ and the control input is the acceleration $u(t)$. The performance index which has to be minimized is the following:

$$J = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}(t_1)\mathbf{x}(t_1) + \frac{1}{2}\int_{t_0}^{t_1} (\mathbf{x}^T(t)\mathbf{R}_1\mathbf{x}(t) + r_2u^2(t))dt.$$

Here the weight matrices are selected to be

$$\mathbf{S}(t_1) = \begin{bmatrix} s_{11}(t_1) & s_{12}(t_1) \\ s_{12}(t_1) & s_{22}(t_1) \end{bmatrix},$$

$$\mathbf{R}_1 = \begin{bmatrix} r_p & 0 \\ 0 & r_v \end{bmatrix},$$

$$r_2 > 0.$$

In order for $\mathbf{S}(t_1)$ and $\mathbf{R}_1$ to be positive semi-definite, the parameters $r_p$, $r_v$, $r_2$ and the eigenvalues of $\mathbf{S}(t_1)$ have to be non-negative. Introducing the weight parameter matrices in the Riccati equation gives

$$\begin{aligned}-\dot{\mathbf{P}}(t) &= \mathbf{A}^T\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A} + \mathbf{R}_1 - \mathbf{P}(t)\mathbf{B}\mathbf{R}_2^{-1}\mathbf{B}^T\mathbf{P}(t) \\ &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\mathbf{P}(t) + \mathbf{P}(t)\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} r_p & 0 \\ 0 & r_v \end{bmatrix} - \mathbf{P}(t)\begin{bmatrix} 0 \\ 1 \end{bmatrix}\frac{1}{r_2}[0 \quad 1]\mathbf{P}(t).\end{aligned}$$

As $\mathbf{P}(t)$ is symmetric this leads to three equations for the elements of $\mathbf{P}(t) = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}$:

$$\begin{aligned}-\dot{p}_{11} &= r_p - \frac{1}{r_2}p_{12}^2, \\ -\dot{p}_{12} &= p_{11} - \frac{1}{r_2}p_{12}p_{22}, \\ -\dot{p}_{22} &= r_v + 2p_{12} - \frac{1}{r_2}p_{22}^2.\end{aligned}$$

These equations have to be solved backwards in time from an initial value at time $t_1$ of $\mathbf{S}(t_1)$. A numerical integration with the parameter values at $s_{11}(t_1) = s_{12}(t_1) = s_{22}(t_1) = 1$, $r_p = 3$, $r_v = 4$ and $r_2 = 1$ gives the solution shown in Fig. 5.9 (remember that $\mathbf{P}(t_1) = \mathbf{S}(t_1)$).

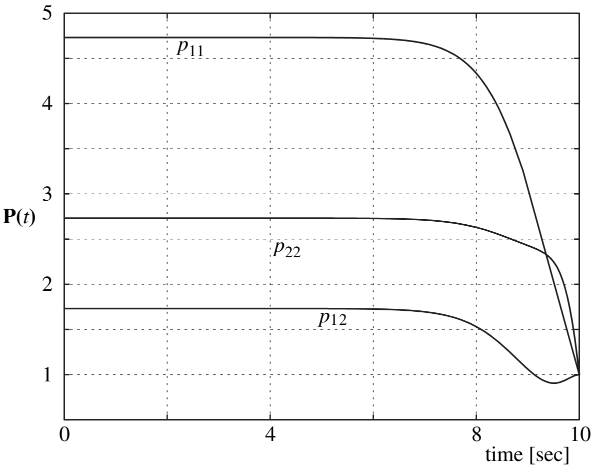
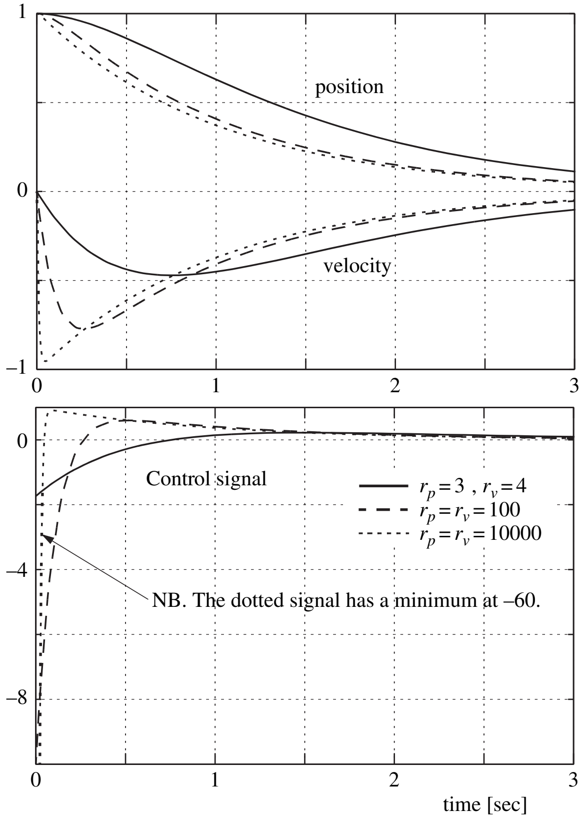


**Fig. 5.9** Solution of the Riccati equation for the double integrator

**Fig. 5.10** Response of the double integrator system with LQR regulator



It is seen from the figure that the three values of the matrix elements of $\mathbf{P}(t)$ approach a constant when $t_1 - t$ becomes sufficiently large. It will be seen later that this is a general quality of the Riccati equation that occurs under certain well-defined conditions. Having established the values of $\mathbf{P}(t)$ it is easy to calculate the control signal from equation (5.72) and the LQR gain in (5.73). Suppose the system starts at time $t = 0$ and has the stationary position $x_1(0) = 1$. Then the optimal controller will give a response as seen in Fig. 5.10. In the figure there are three different controls with different values of the $\mathbf{R}_1$ matrix. It is seen that the larger the values of the $\mathbf{R}_1$ matrix elements, the faster the response at the expense of a larger control signal. ❐

## 5.5 Steady State Linear Quadratic Regulator

The closed-loop LQR controller from Sect. 5.4.2 is the optimal controller that minimizes the performance index over a finite time interval $[t_0, t_1]$. As has been demonstrated this leads to a time varying LQR gain matrix which can be calculated off-line. In most cases it is more convenient to have a constant gain matrix and therefore attention is directed to the last example of the previous section, Fig. 5.9. It is seen that the solution to the Riccati equation becomes a matrix with constant elements over much of the time interval.

It would therefore be of interest to look at an optimal control problem with a performance index extending to infinity:

$$J(\mathbf{u}) = \lim_{t\to\infty}\left[\int_{t_0}^{t}(\mathbf{x}^T(t)\mathbf{R}_1\mathbf{x}(t) + \mathbf{u}(t)\mathbf{R}_2\mathbf{u}(t))dt\right]. \tag{5.80}$$

Assume further that the system is time invariant, i.e.,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{5.81}$$

so that all of the matrices are assumed to be constants and $\mathbf{R}_1 \geq 0$ and $\mathbf{R}_2 > 0$.

Since the problem with standard LQR controllers is to move the incremental states to the zero state in an optimal way, the state vector $\mathbf{x}(t)$ will approach the zero-vector as $t_1 \to \infty$ if the closed loop system is stable. It is therefore of no relevance to include a final state term here. This is the same as setting $\mathbf{S}(t_1) = \mathbf{0}$ in equation (5.52).

It is not difficult to show that the optimum value $\mathrm{J}_{\mathrm{min}}$ of the index has an upper bound for all values of the final time if the system is *stabilizable*. It is obvious that the index is monotonically non-decreasing and these facts prove that the index has a limiting value even for $t_1 \to \infty$.

The optimum value of the index is given by equation (5.78). Since the system is time invariant, this value $\mathrm{J}_{\mathrm{min}}$ must be independent of the initial time $t_0$ which means that the matrix $\mathbf{P}$ must be constant. This implies that $\dot{\mathbf{P}} = \mathbf{0}$ and the Riccati equation reduces to

$$\mathbf{0} = \mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{R}_1 - \mathbf{P}\mathbf{B}\mathbf{R}_2^{-1}\mathbf{B}^T\mathbf{P}. \tag{5.82}$$

This is a set of coupled nonlinear (quadratic) algebraic equations. It is common practice to call this equation the *Algebraic Riccati Equation* (ARE) although it is no longer a differential equation. The limiting constant solution is denoted as $\mathbf{P}_\infty$.

Equation (5.82) may have multiple solutions, but it can be shown that only one of them is positive semi-definite (provided that the system is stabilizable) and that particular solution leads to the minimum value of the performance index:

$$J_{min} = \frac{1}{2}\mathbf{x}_0^T\mathbf{P}_\infty\mathbf{x}_0. \tag{5.83}$$

The optimal steady state gain matrix is found as before as

$$\mathbf{K}_\infty = \mathbf{R}_2^{-1}\mathbf{B}^T\mathbf{P}_\infty \tag{5.84}$$

and the control signal becomes

$$\mathbf{u}(t) = -\mathbf{K}_{\infty}\mathbf{x}(t). \tag{5.85}$$

A very important property of LQR regulators is that the closed-loop system is stable under certain conditions. The following theorem is stated without proof and the interested reader should refer to Lewis (1986) and Bryson and Ho (1975) for details:

**Steady State Continuous LQR Regulator Theorem**:

For the system described by the time invariant state equation,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{5.86}$$

subject to the following performance index,

$$J(\mathbf{u}) = \int_0^{t_1} (\mathbf{x}^T(t)\mathbf{R}_1\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}_2\mathbf{u}(t))dt, \tag{5.87}$$

then the following holds true:

If the system in equation (5.86) is stabilizable and the matrix pair $(\mathbf{A}, \sqrt{\mathbf{R}_1})$ is detectable then the algebraic Riccati equation has one and only one solution which is positive definite. This solution, $\mathbf{P}_{\infty}$, leads to the minimum value (5.83) for the performance index.

If the system in equation (5.86) is stabilizable and the matrix pair $(\mathbf{A}, \sqrt{\mathbf{R}_1})$ is detectable then the resulting state feedback law,

$$\mathbf{u}(t) = -\mathbf{K}_{\infty}\mathbf{x}(t) = -\mathbf{R}_2^{-1}\mathbf{B}^T\mathbf{P}_{\infty}\mathbf{x}(t), \tag{5.88}$$

gives an asymptotically stable closed loop system.

This is probably one of the most important results in modern control theory and one that has far-reaching consequences for the design of optimal control systems. Basically, the theorem ensures that under very broad conditions (stabilizability and detectability), which can easily be tested for LTI systems, the state-feedback in equation (5.88) will give a stable control system.

***Example 5.4*. Steady State LQR for the First Order System**

The first order system in Example 5.2 has the state equation,

$$\dot{x}(t) = ax(t) + bu(t),$$

which is controllable (and therefore also stabilizable) if $b \neq 0$. The steady-state performance index is:

$$J = \frac{1}{2}\int_0^{\infty} (r_1x^2(t) + r_2u^2(t))dt.$$

Here, $r_2 > 0$ and if $r_1 > 0$ the 'matrix' pair $(a, \sqrt{r_1})$ is observable (and detectable of course). Then the steady-state optimal control law can be found by solving the ARE,

$$0 = 2ap_\infty + r_1 - \frac{b^2}{r_2} p_\infty^2.$$

This equation has one positive solution,

$$p_\infty = \frac{1}{b^2}\left(a + \sqrt{a^2 + \frac{r_1 b^2}{r_2}}\right).$$

Comparing with Example 5.2, it is clear that this is exactly the same as obtained for large $t_1 - t$. The feedback control law is then:

$$u(t) = -bp_\infty x(t) = -\left(\frac{a}{b} + \sqrt{\left(\frac{a}{b}\right)^2 + \frac{r_1}{r_2}}\right) x(t).$$

The control law is seen to give large control signals for large values of $r_1 \,/\, r_2$ and small control signals for small values of $r_1 \,/\, r_2$. The LQR gain is approximately proportional to $\sqrt{(r_1/r_2)}$ as noted earlier. The pole of the closed loop system is easily calculated to be

$$s = -\frac{1}{b}\sqrt{\frac{a^2}{b^2} + \frac{r_1}{r_2}}.$$

Again the general feature of the control law is to give a faster system for large values of $r_1/r_2$.

If controllability is lost, i.e., $b = 0$, control is impossible since the control signal cannot influence the system. The LQR calculation does its best to cope with the situation by making the control signal infinite. The closed loop state equation reduces to $\dot{x}(t) = ax(t)$, so stability depends on the sign of $a$. If $r_1 \to 0$ the system is not observable and detectability may also have been lost. In this case the control signal becomes

$$u(t) = -\left(\frac{a}{b} + \left|\frac{a}{b}\right|\right) x(t).$$

The closed loop state equation is consequently

$$\dot{x}(t) = -\frac{b}{|b|}|a| x(t).$$

and the stability now depends on the sign of $b$. ❐

***Example 5.5*. Steady State LQR for the Double Integrator**

The next example here is for the second order system from Example 5.3. This is a double integrator with the state equation:

$$\mathbf{x}(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t).$$

The controller has to minimize the performance index,

$$J = \int_0^\infty (\mathbf{x}^T(t)\mathbf{R}_1\mathbf{x}(t) + r_2 u^2(t))dt.$$

The state weight matrix is chosen to be diagonal: $\mathbf{R}_1 = \begin{bmatrix} r_p & 0 \\ 0 & r_v \end{bmatrix}$. $r_p$ and $r_v$ are the position and velocity weights respectively. With the values above the algebraic Riccati equation results in the following three coupled equations:

$$\begin{aligned} 0 &= r_p - \frac{1}{r_2} p_{12}^2, \\ 0 &= p_{11} - \frac{1}{r_2} p_{12} p_{22}, \quad . \\ 0 &= 2p_{12} + r_v - \frac{1}{r_2} p_{22}^2 \end{aligned}$$

Here, $\mathbf{P}_\infty = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}$ is the solution to the ARE. This algebraic equation is easily solved and gives the unique positive definite solution:

$$\begin{aligned} p_{11} &= \sqrt{r_p}\sqrt{2r_2\sqrt{r_p r_2} + r_v r_2}, \\ p_{12} &= \sqrt{r_p r_2}, \\ p_{22} &= \sqrt{2r_2\sqrt{r_p r_2} + r_v r_2}. \end{aligned}$$

The controller becomes a state feedback controller with the control signal:

$$\begin{aligned} u(t) &= -\mathbf{K}_\infty \mathbf{x}(t) = (-\mathbf{R}_2^{-1}\mathbf{B}^T\mathbf{P}_\infty \mathbf{x}(t)) \\ &= -\left[ \sqrt{\frac{r_p}{r_2}} \sqrt{2\sqrt{\frac{r_p}{r_2}} + \frac{r_v}{r_2}} \right] \mathbf{x}(t). \end{aligned}$$

It is seen that the control gain increases as the weights of the states are increased, i.e., $r_p$ and $r_v$ are increased compared to the control signal weight. The closed loop characteristic equation can be found as

$$s^2 + \sqrt{2\sqrt{\frac{r_p}{r_2}} + \frac{r_v}{r_2}}s + \sqrt{\frac{r_p}{r_2}} = 0.$$

The corresponding closed loop eigenfrequency $\omega_n$ and damping ratio $\zeta$ can be seen to be:

$$\omega_n = \sqrt[4]{\frac{r_p}{r_2}},$$

$$\zeta = \frac{1}{\sqrt{2}}\sqrt{1 + \frac{r_v}{2\sqrt{r_2}\sqrt{r_p}}}.$$

Now let $r_2 = 1$. It is then clear that if the weight of the velocity is zero ($r_v = 0$) the closed loop poles have a damping ratio of $1/(\sqrt{2}) \approx 0.71$. The root-locus as a function of weights is seen in Fig. 5.11, where $r_p$ varies from 0 till 10. For this particular system it is seen that the damping ratio of the closed loop poles is always guaranteed to be greater than 0.71 for positive weight matrix elements. So for this type of second order system the steady-state LQ regulator gives a well-damped system. ❐

***Example 5.6.* Steady State LQR for the Two Link Robot**

The robot control problem in Example 4.14 is now repeated with the steady state (continuous time) LQ regulator. The linear system model is given by:

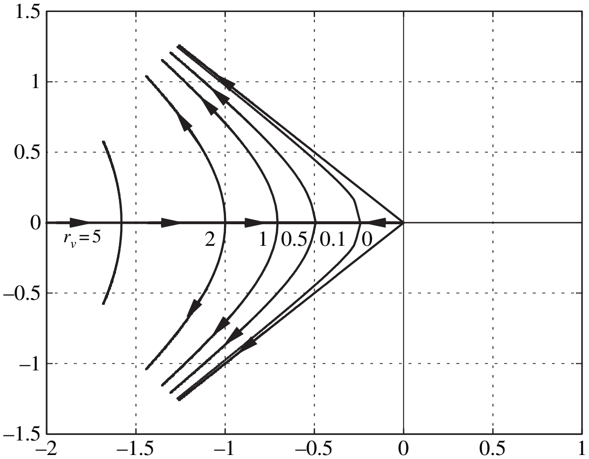


**Fig. 5.11** Root-locus for Newtonian system for variation of $r_p$

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 17.832 & 0 & -3.0024 & 0 \\ 0 & 0 & 0 & 1 \\ -30.063 & 0 & 10.456 & 0 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 & 0 \\ 1.2514 & -2.4337 \\ 0 & 0 \\ -2.4337 & 6.5512 \end{bmatrix} \mathbf{u}(t)$$

$$= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t).$$

In order to eliminate stationary errors, output error integration will be introduced and the system augmented with two error integral states that obey the state equation

$$\dot{\mathbf{x}}_i(t) = \mathbf{r}(t) - \mathbf{y}(t) = \mathbf{e}(t).$$

The error is here the difference of a set of reference positions, $\mathbf{r}(t)$, and the actual outputs, $\mathbf{y}(t)$. The resulting augmented system has six states and therefore the weight matrix for the states becomes

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 500 & 250 \\ 0 & 0 & 0 & 0 & 250 & 500 \end{bmatrix}.$$

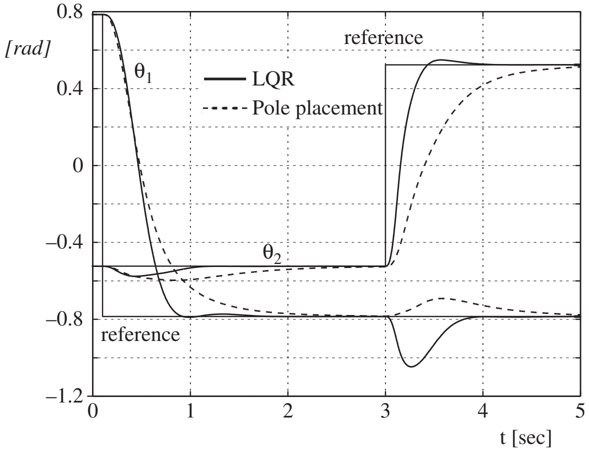
The error integrals are weighted harder to make sure that they will be eliminated quickly. The selection of weight matrix elements is done by trial and error, as is common practice. However, the overall guideline is quite simple. If a given state vector element upon simulation is responding too slowly, the corresponding state weight matrix diagonal element has to be increased. Similarly for the relative strength of the control signals. Since cross coupling between the two links is a major problem, the integral error cross coupling terms are also weighted harder. The maximum torques that the actuators can exert on the links are not available. If they were then a good relative weight for the control signals would be

$$\mathbf{R}_2 = \rho \begin{bmatrix} \frac{1}{max(u_1^2)} & 0 \\ 0 & \frac{1}{max(u_2^2)} \end{bmatrix}.$$

The two control signals are weighted such that the lower link torque is allowed to be relatively larger, since it carries a greater load. A good selection seems to be

$$\mathbf{R}_2 = 5 \cdot 10^{-4} \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}.$$

**Fig. 5.12** Two-link robot position



As in Example 4.14 the robot starts from initial position (45°, −30°), makes a step at $t = 0.1$ sec to (−45°, −30°) and finally at $t = 3$ sec makes another step to a final position (−45°, 30°). The resulting state responses are depicted in Fig. 5.12. The control signals are shown in Fig 5.13. For comparison a pole placement controller is also shown that is designed to have roughly the same maximum control signal inputs.

As is seen the LQR controller works well even on a non-linear system such as the two link robot. Compared to the pole placement controller it is much faster in response with roughly the same control signal. However the cross coupling from link 2 to link 1 is substantial and it is likely that finer tuning could decouple this interaction. ❐

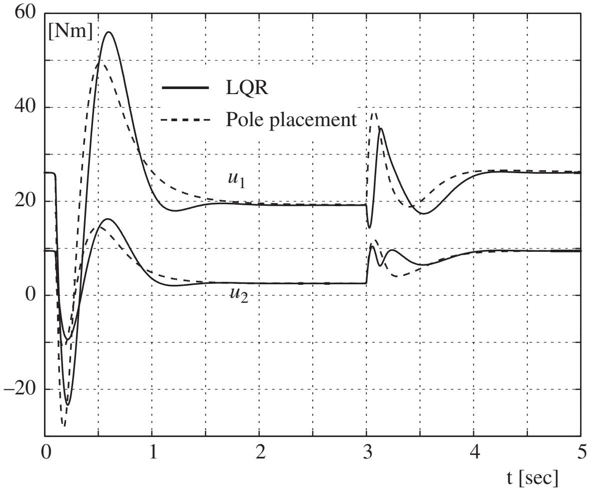


**Fig. 5.13** Control signals for the two link robot

### 5.5.1 Robustness of LQR Control

LQR regulators have some particularly useful characteristics as regards their gain margin, phase margin and tolerance of modelling error. For time varying control objects this robustness is dependent on the characteristics of the system in question and what control is being required of it. For LTI control objects however it is possible to be much more concrete. This subsection documents the stability characteristics of LQR controllers when applied to LTI systems.

The approach to the robustness of LQR regulators here is based on Lyapunov's direct method which was reviewed in Sect. 3.7.5. Consider a Lyapunov candidate function, $V(\mathbf{x}) = \mathbf{x}^T\mathbf{P}\mathbf{x}$, for a LQR regulator. It is known from previous sections that $\mathbf{P} > \mathbf{0}$ (is positive definite). This means that it must be so that the time derivative of $V$ is neative for stability, i.e., $\dot{V} < 0$ for $\mathbf{x} \neq \mathbf{0}$. $\dot{V}$ can be found by differentiation of the candidate Lyapunov function. Assume that the $\mathbf{B}$ matrix of the control object is incorrect or has been changed to a different one for some reason or another to $\mathbf{B}_\Delta$:

$$\begin{aligned}\dot{V} &= \mathbf{x}^T\mathbf{P}\dot{\mathbf{x}} + \dot{\mathbf{x}}^T\mathbf{P}\mathbf{x}\\ &= \mathbf{x}^T\mathbf{P}(\mathbf{A} - \mathbf{B}_\Delta\mathbf{K}_\infty)\mathbf{x} + \mathbf{x}^T\mathbf{P}(\mathbf{A}^T - \mathbf{K}_\infty^T\mathbf{B}_\Delta^T)\mathbf{P}\mathbf{x}\\ &= \mathbf{x}^T(\mathbf{P}\mathbf{A} + \mathbf{A}^T\mathbf{P})\mathbf{x} - \mathbf{x}^T(\mathbf{P}\mathbf{B}_\Delta\mathbf{K}_\infty + \mathbf{K}^T\mathbf{B}_\Delta^T\mathbf{P})\mathbf{x},\end{aligned} \tag{5.89}$$

where $\mathbf{K}_\infty$ is the optimal steady state LQR gain and the $\mathbf{x}$ time derivatives have been replaced by the closed loop state equation, $\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}_\Delta\mathbf{K}_\infty)\mathbf{x}$, including the modified input matrix, $\mathbf{B}_\Delta$. The input to the state equation can be ignored because it does not influence the stability of the system. The term in the parenthesis on the left in the last line above can be replaced by $-\mathbf{R}_1 + \mathbf{P}\mathbf{B}\mathbf{K}_\infty$ since the system is known to answer to the steady state Riccati equation (5.82).

The quantities in the parentheses in Eq. (5.89) are the argument of a quadratic form, thus

$$\begin{aligned}\dot{V} &= \mathbf{x}^T(-\mathbf{R}_1 + \mathbf{P}\mathbf{B}\mathbf{K}_\infty)\mathbf{x} - \mathbf{x}^T(2\mathbf{P}\mathbf{B}_\Delta\mathbf{K}_\infty)\mathbf{x}\\ &= \mathbf{x}^T(-\mathbf{R}_1 + \mathbf{P}\mathbf{B}\mathbf{K}_\infty - 2\mathbf{P}\mathbf{B}_\Delta\mathbf{K}_\infty)\mathbf{x}\\ &= \mathbf{x}^t(-\mathbf{R}_1 + \mathbf{P}(\mathbf{B} - 2\mathbf{B}_\Delta)\mathbf{K}_\infty)\mathbf{x}.\end{aligned} \tag{5.90}$$

Now let $\mathbf{B}_\Delta = \mathbf{B}\Delta$ where $\Delta$ is a diagonal matrix of arbitrary gains. Then the condition for stability, $\dot{V} < 0$, requires that

$$\mathbf{R}_1 - \mathbf{P}\mathbf{B}(\mathbf{I} - 2\Delta)\mathbf{K}_\infty > \mathbf{0}. \tag{5.91}$$

If $\Delta$ is allowed to vary then for stability, $\Delta_{ii} > \frac{1}{2}$, $i = 1,2, \ldots, n$, and an LQR regulator allows for a one half gain reduction. On the other hand it is allowed that $\Delta_{ii} \rightarrow \infty$ and an infinite gain margin is predicted for a gain increase, given an accurate control object model.

The phase margin can be investigated approximately by letting the diagonal elements of $\Delta$ be phases, $\Delta_{ii} = e^{j\theta_i}$. In this case it is clear that it is permitted that the real part of the matrix elements of $\Delta$ be greater than ½ for all $|\, \theta_i \,| < 60^{\circ}$. Thus there is a $60^{\circ}$ phase margin for all inputs.

It is thus seen that LQR control of LTI systems can be realized with guarenteed large gain and phase margins given accurate control object information. This inherent robustness is one of the main reasons that LQR control is currently so wide spread in so many different control applications. Moreover it is the most important benchmark to which any other multivariable controller is compared when considering a particular control application.

## *5.5.2 LQR Design: Eigenstructure Assignment Approach*

Now a different way of designing a steady-state LQ regulator will be given. In certain cases this method can be used to an advantage.

A LTI control object will be assumed to have the usual linearized form:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{5.92}$$

with the steady-state performance index:

$$J = \frac{1}{2}\int_0^{\infty} (\mathbf{x}^T(t)\mathbf{R}_1\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}_2\mathbf{u}(t))\; dt. \tag{5.93}$$

It is also assumed that the system $(\mathbf{A}, \mathbf{B})$ is stabilizable and that $(\mathbf{A}, \sqrt{\mathbf{R}_1})$ is detectable.

It is now to be shown that the steady-state optimal state feedback gain $\mathbf{K}_{\infty}$ can be found directly from the Hamiltonian system (5.62) as

$$\begin{bmatrix} \dot{\mathbf{x}}(t) \\ \dot{\lambda}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A} & -\mathbf{B}\mathbf{R}_2^{-1}\mathbf{B}^T \\ -\mathbf{R}_1 & -\mathbf{A}^T \end{bmatrix} \begin{bmatrix} \mathbf{x}(t) \\ \lambda(t) \end{bmatrix} = \mathbf{H} \begin{bmatrix} \mathbf{x}(t) \\ \lambda(t) \end{bmatrix}. \tag{5.94}$$

Moreover it will be shown that the optimal closed loop system,

$$\dot{\mathbf{x}}(t) = (\mathbf{A} - \mathbf{B}\mathbf{K}_{\infty})\mathbf{x}(t), \tag{9.95}$$

has eigenvalues that are the stable eigenvalues of $\mathbf{H}$. Above $\mathbf{K}_{\infty}$ is the state feedback gain in equation (5.88).

The derivation of the state feedback gain $\mathbf{K}_{\infty}$ starts with the observation that the eigenvalues of $\mathbf{H}$ are symmetric on either side of the imaginary axis. To show this start by noting that

$$\mathbf{J}\mathbf{H}^T\mathbf{J} = \mathbf{H},$$

where $\mathbf{J} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix}$. If μ is an eigenvalue for **H** there exists an eigenvector **v** such that

$$\mathbf{H}\mathbf{v} = \mu\mathbf{v},$$

since $\mathbf{J}^T = -\mathbf{J}$,

$$\mathbf{J}\mathbf{H}^T\mathbf{J}\mathbf{v} = \mu\mathbf{v} \Leftrightarrow (\mathbf{J}\mathbf{v})^T\mathbf{H} = -\mu(\mathbf{J}\mathbf{v})^T. \tag{5.96}$$

So it can be observed that −μ is also an eigenvalue with (left) eigenvector $(\mathbf{J}\mathbf{v})^T$. Consequently it can be concluded that **H** has $2n$ eigenvalues, $n$ of which are stable and $n$ of which are unstable.

It will now be demonstrated that the $n$ stable eigenvalues are also the closed loop eigenvalues of the system in equation (5.94). Since the system is stabilizable and detectable the closed loop system has $n$ stable eigenvalues. Let $\mu_i$ be one of these and assume that only the mode corresponding to this eigenvalue, $\mu_i$, is excited. If $\mathbf{X}_i$ is the corresponding eigenvector the state equation can be integrated to give

$$\mathbf{x}(t) = \mathbf{X}_i e^{\mu_i t}. \tag{5.97}$$

Since $\mathbf{u}(t)$ and $\lambda(t)$ depend linearly on $\mathbf{x}(t)$ one also has that

$$\begin{aligned} \mathbf{u}(t) &= \mathbf{U}_i e^{\mu_i t}, \\ \lambda(t) &= \Lambda_i e^{\mu_i t}. \end{aligned} \tag{5.98}$$

For the Hamiltonian system this gives

$$\begin{bmatrix} \dot{\mathbf{x}}(t) \\ \dot{\lambda}(t) \end{bmatrix} = \mu_i \begin{bmatrix} \mathbf{x}(t) \\ \lambda(t) \end{bmatrix} = \mathbf{H} \begin{bmatrix} \mathbf{x}(t) \\ \lambda(t) \end{bmatrix} \Leftrightarrow \mathbf{H} \begin{bmatrix} \mathbf{X}_i \\ \Lambda_i \end{bmatrix} = \mu_i \begin{bmatrix} \mathbf{X}_i \\ \Lambda_i \end{bmatrix}. \tag{5.99}$$

So $\mu_i$ is also an eigenvalue of **H**. Since there are exactly $n$ stable eigenvalues in **H** it can be deduced that these are the $n$ closed loop poles of the stable LQR state feedback system.

The feedback gain $\mathbf{K}_\infty$ can be determined from the eigenstructure of the Hamilton matrix in the following way. From Eqs. (5.61) and (5.88) one has:

$$\mathbf{U}_i = -\mathbf{R}_2^{-1}\mathbf{B}^T\Lambda_i = -\mathbf{K}_\infty\mathbf{X}_i.$$

Since this is valid for all the $n$ eigenvectors $\mathbf{X}_i$ this may be written in matrix form as

$$\mathbf{K}_\infty\mathbf{X} = \mathbf{R}_2^{-1}\mathbf{B}^T\Lambda. \tag{5.100}$$

Here the system modal matrix is $\mathbf{X} = [\mathbf{X}_1\, \mathbf{X}_2 \ldots \mathbf{X}_n]$ and $\Lambda = [\Lambda_1\, \Lambda_2 \ldots \Lambda_n]$. For a non-defective system, i.e., a system without repeated eigenvalues, the modal matrix is non-singular, so it may be inverted it to obtain

$$\mathbf{K}_\infty = \mathbf{R}_2^{-1}\mathbf{B}^T \Lambda \mathbf{X}^{-1}. \tag{5.101}$$

So for a non-defective system the steady-state optimal controller can be found from the eigenstructure of $\mathbf{H}$. The method involves finding the eigenvectors of $\mathbf{H}$ for its $n$ stable eigenvalues. Comparing equations (5.101) and (5.88) one sees that the solution of the algebraic Riccati equation is

$$\mathbf{P}_\infty = \Lambda \mathbf{X}^{-1}. \tag{5.102}$$

It is also clear that if the state weight matrix $\mathbf{R}_1$ is zero the Hamilton matrix is

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} & -\mathbf{B}\mathbf{R}_2^{-1}\mathbf{B}^T \\ \mathbf{0} & -\mathbf{A}^T \end{bmatrix}.$$

So its eigenvalues are the open-loop eigenvalues in $\mathbf{A}$ combined with the eigenvalues of $-\mathbf{A}$. Thus, if there is no state weight then the closed loop poles are the $n$ stable poles in $\mathbf{A}$ and $-\mathbf{A}$. Therefore without state weight the poles of a stable system cannot be moved with a steady state LQ regulator.

***Example 5.7*. Optimal Eigenstructure Assignment for the Double Integrator**

The double integrator system from Example 5.6 has the state equation:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u(t),$$

with the same weight matrices as before, $\mathbf{R}_1 = \begin{bmatrix} r_p & 0 \\ 0 & r_v \end{bmatrix}$ and $r_2 = 1$. The Hamiltonian matrix can be obtained as

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} & -\mathbf{B}\mathbf{B}_2^{-1}\mathbf{B}^T \\ -\mathbf{R}_1 & -\mathbf{A}^T \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ -r_p & 0 & 0 & 0 \\ 0 & -r_v & -1 & 0 \end{bmatrix}.$$

This matrix has the characteristic polynomial,

$$\lambda^4 - r_v\lambda^2 + r_p = 0,$$

which has four solutions, two in each half plane,

$$\lambda = \pm\sqrt{\frac{r_v}{2} \pm \sqrt{\frac{r_v}{2} - r_p}}.$$

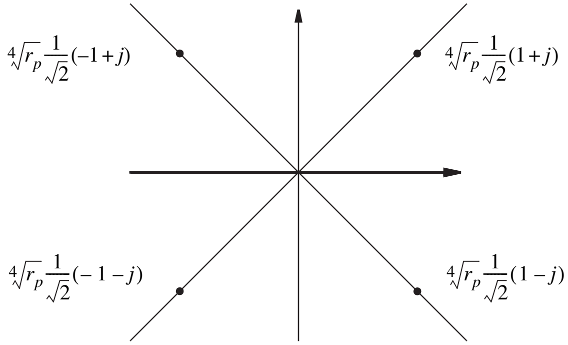

$\sqrt[4]{r_p}\frac{1}{\sqrt{2}}(-1+j)$
$\sqrt[4]{r_p}\frac{1}{\sqrt{2}}(1+j)$
$\sqrt[4]{r_p}\frac{1}{\sqrt{2}}(-1-j)$
$\sqrt[4]{r_p}\frac{1}{\sqrt{2}}(1-j)$

**Fig. 5.14** Eigenvalues of the double integrator Hamiltonian

The two stable poles are the same as were found in Example 5.6. For $r_p$ going to zero, the four poles become $(0, 0, \sqrt{r_v}, -\sqrt{r_v})$, with marginal stability because the system loses detectability.

For $r_\nu$ going to zero, the system retains detectability and the four closed loop poles become

$$\lambda = \pm\sqrt{\pm\sqrt{-r_p}} = \pm 4\sqrt{r_p}e^{\pm j\frac{\pi}{4}}.$$

The four poles are arranged as seen in Fig. 5.14.

The eigenvectors corresponding to the stable eigenvalues can be used to calculate the state feedback gain. This is most easily calculated numerically, in for example MATLAB. ❐

## 5.6 Discrete Time Optimal Control

In many practical applications the optimal controller is implemented in a computer in digital form. Therefore a transformation of the results from continuous time to a form applicable to digital implementation is often necessary. This transformation can be carried out in two distinctly different ways. One is to design the controller in continuous time and then to derive an approximate discrete version of it. A popular method is the Tustin approximation to the discretization or any of a number of other discretization techniques. See for instance a thorough treatment in Franklin, Powell and Workman (1990).

The alternative method to design a digital controller is to discretize the continuous time model and perform the design in discrete time. In Chap. 2 it was described how a discrete time model can be derived from a continuous time model by integrating over a sample period of duration $T$. In Chap. 4 it was shown how make a pole placement controller in discrete time by placing the poles inside the unit circle. It will be assumed here that a discrete time model has been found for the system in the form of a linearized LTI state space model:

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k). \tag{5.103}$$

A linear quadratic controller in discrete time will be introduced here and the discrete time equivalent optimal state feedback controller and the discrete Riccati equation derived. First however, the problem of finding the discrete cost function must be addressed.

## 5.6.1 Discretization of the Performance Index

The continuous time quadratic performance index for the LQ regulator is:

$$J = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}(t_1)\mathbf{x}(t_1) + \frac{1}{2}\int_0^{t_1} (\mathbf{x}^T(t)\mathbf{R}_1\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}_2\mathbf{u}(t))dt. \tag{5.104}$$

The cost function is also assumed time invariant and therefore the initial time is defined as zero, which is not a limitation. It will now be assumed that the final time is a whole number of samples, i.e., $t_1 = NT$ and therefore the cost function may be written:

$$J = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}(t_1)\mathbf{x}(t_1) + \frac{1}{2}\sum_{k=0}^{N-1}\int_{kT}^{(k+1)T} [\mathbf{x}^T(t)\mathbf{R}_1\mathbf{x}(t) + \mathbf{u}^T(t)\mathbf{R}_2\mathbf{u}(t)]dt.$$

To a first approximation the state vector may be assumed constant between samples and therefore the cost function can be approximated by the expression

$$J = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}(t_1)\mathbf{x}(t_1) + \frac{1}{2}\sum_{k=0}^{N-1}[\mathbf{x}^T(k)\mathbf{R}_1\mathbf{x}(k) + \mathbf{u}^T(k)\mathbf{R}_2\mathbf{u}(k)] \cdot T. \tag{5.105}$$

It is seen that the continuous time weight matrices are transformed into an equivalent discrete time version by the following conversion:

$$\begin{aligned} \mathbf{S}_d(N) &= \mathbf{S}(t_1), \\ \mathbf{R}_{1d} &= \mathbf{R}_1 T, \\ \mathbf{R}_{2d} &= \mathbf{R}_2 T. \end{aligned} \tag{5.106}$$

The discrete time cost function is then

$$J_d = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}_d(N)\mathbf{x}(t_1) + \frac{1}{2}\sum_{k=0}^{N-1}[\mathbf{x}^T(k)\mathbf{R}_{1d}\mathbf{x}(k) + \mathbf{u}^T(k)\mathbf{R}_{2d}\mathbf{u}(k)]. \tag{5.107}$$

This is the cost function that will form the basis for the derivation of the discrete time linear quadratic regulator. As usual the requirement for a lower bound on the cost function is that the weight matrices have the following properties:

$$\mathbf{S}_d(N) \geq 0, \qquad \mathbf{R}_{1d} \geq 0, \qquad \mathbf{R}_{2d} > 0.$$

### 5.6.2 Discrete Time State Feedback

With the state equation in (5.103) and the cost function in (5.107) a vector of Lagrange multipliers $\lambda \in \Re^n$ is introduced and the Hamiltonian becomes:

$$\begin{aligned} H(\mathbf{x}(k),&\mathbf{u}(k),\lambda(k+1)) \\ &= \frac{1}{2}(\mathbf{x}^T(k)\mathbf{R}_{1d}\mathbf{x}(k) + \mathbf{u}^T(k)\mathbf{R}_{2d}\mathbf{u}(k)) \\ &\qquad + \lambda^T(k+1)(\mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k)). \end{aligned} \tag{5.108}$$

Note that the Lagrange multiplier is indexed with the time $k + 1$ rather than $k$. This is only for notational convenience and is just a convention.

The co-state equation gives

$$\lambda(k) = \frac{\partial H}{\partial \mathbf{x}(k)} = \mathbf{R}_{1d}\mathbf{x}(k) + \mathbf{F}^T\lambda(k+1). \tag{5.109}$$

The stationarity equation leads to

$$0 = \frac{\partial H}{\partial \mathbf{u}(k)} = \mathbf{R}_{2d}\mathbf{u}(k) + \mathbf{G}^T\lambda(k+1).$$

Solving for $\mathbf{u}(k)$ gives

$$\mathbf{u}(k) = -\mathbf{R}_{2d}^{-1}\mathbf{G}^T\lambda(k+1). \tag{5.110}$$

Inserting this in the state equation gives

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) - \mathbf{G}\mathbf{R}_{2d}^{-1}\mathbf{G}^T\lambda(k+1). \tag{5.111}$$

The boundary condition at final time sample $N$ is

$$\lambda(N) = \frac{\partial}{\partial \mathbf{x}(N)}\Phi(\mathbf{x}(N)) = \mathbf{S}_d(N)\mathbf{x}(N). \tag{5.112}$$

Suppose that (analogous to the continuous time case) this relation is valid for all time samples, i.e., that there exists a matrix $\mathbf{P}_d(k)$ such that

$$\lambda(k) = \mathbf{P}_d(k)\mathbf{x}(k) \text{ with } \mathbf{P}_d(\mathbf{N}) = \mathbf{S}_d(\mathbf{N}). \tag{5.113}$$

As in the continuous time case this expression may be used to modify the equations in (5.109) and (5.111) to obtain:

$$\mathbf{P}_d(k)\mathbf{x}(k) = \mathbf{R}_{1d}\mathbf{x}(k) + \mathbf{F}^T\mathbf{P}_d(k+1)\mathbf{x}(k+1),$$
$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) - \mathbf{G}\mathbf{R}_{2d}^{-1}\mathbf{G}^T\mathbf{P}_d(k+1)\mathbf{x}(k+1).$$

Solving the second equation with respect to **x**($k$ + 1) and inserting the result in the first equation above,

$$\mathbf{P}_d(k)\mathbf{x}(k) = \mathbf{R}_{1d}\mathbf{x}(k) + \mathbf{F}^T\mathbf{P}_d(k+1)[\mathbf{I} + \mathbf{G}\mathbf{R}_{2d}^{-1}\mathbf{G}^T\mathbf{P}_d(k+1)]^{-1}\mathbf{F}\mathbf{x}(k).$$

If this is to be true for all values of **x**($k$) then the *discrete Riccati equation* is valid:

$$\mathbf{P}_d(k) = \mathbf{R}_{1d} + \mathbf{F}^T\mathbf{P}_d(k+1)[\mathbf{I} + \mathbf{G}\mathbf{R}_{2d}^{-1}\mathbf{G}^T\mathbf{P}_d(k+1)]^{-1}\mathbf{F}. \tag{5.114}$$

The discrete Riccati equation is solved backwards in time from sample $N$ where the boundary value, $\mathbf{P}_d(N) = \mathbf{S}_d(N)$, is given. This version of the Riccati equation is most useful for hand calculations of smaller systems.

From equation (5.110) one has

$$\mathbf{u}(k) = -\mathbf{R}_{2d}^{-1}\mathbf{G}^T\mathbf{P}_d(k+1)\mathbf{x}(k+1).$$

Now this is not a very convenient form of the controller since it requires future state variables. It may be rewritten in a more convenient compact form by inserting the state equation in the control law to obtain

$$\mathbf{u}(k) = -\mathbf{R}_{2d}^{-1}\mathbf{G}^T\mathbf{P}_d(k+1)(\mathbf{F}\mathbf{x}(k) + \mathbf{G}\mathbf{u}(k)).$$

Solving for the control signal a state feedback law is obtained in the form

$$\mathbf{u}(k) = -\mathbf{K}_d(k)\mathbf{x}(k). \tag{5.115}$$

Here the discrete LQR gain is given by

$$\mathbf{K}_d(k) = [\mathbf{R}_{2d} + \mathbf{G}^T\mathbf{P}_d(k+1)\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{P}_d(k+1)\mathbf{F}. \tag{5.116}$$

The Riccati equation may be solved using equation (5.114) but a more compact form is found by the application of the matrix inversion lemma,[†]

[†] The matrix inversion lemma may be found in most linear algebra textbooks and it supports the identity (valid for $det(A_{11}) \neq 0)$ : $(\mathbf{A}_{11}^{-1} + \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21})^{-1} = \mathbf{A}_{11} - \mathbf{A}_{11}\mathbf{A}_{12}(\mathbf{A}_{21}\mathbf{A}_{11}\mathbf{A}_{12} + \mathbf{A}_{22}^{-1})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}$.

$$\mathbf{P}_d(k) = \mathbf{R}_{1d} + \\ + \mathbf{F}^T[\mathbf{P}_d(k+1) - \mathbf{P}_d(k+1)\mathbf{G}[\mathbf{R}_{2d} + \mathbf{G}^T\mathbf{P}_d(k+1)\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{P}_d(k+1)]\mathbf{F}. \tag{5.117}$$

If the matrix $\mathbf{P}_d(k+1)$ can be inverted this may further be simplified by yet another application of the inversion lemma:

$$\mathbf{P}_d(k) = \mathbf{R}_{1d} + \mathbf{F}^T[\mathbf{P}_d^{-1}(k+1) + \mathbf{G}\mathbf{R}_{2d}^{-1}\mathbf{G}^T]^{-1}\mathbf{F}. \tag{5.118}$$

This version of the Riccati equation is most useful for machine calculations because of its compact matrix form.

The closed loop state equation is found by inserting the state feedback signal (equation (5.115)) into the state equation in equation (5.103):

$$\mathbf{x}(k+1) = [\mathbf{I} - \mathbf{G}(\mathbf{R}_{2d} + \mathbf{G}^T\mathbf{P}_d(k+1)\mathbf{G})^{-1}\mathbf{G}^T\mathbf{P}_d(k+1)]\mathbf{F}\mathbf{x}](k).$$

### *5.6.3 Steady State Discrete Optimal Control*

The discrete optimal controller above has the disadvantage that it is time varying. Even for plants that are time-invariant and therefore for the control of the plant over $N$ samples, $N \times m$ control signals have to be calculated and stored ready for control signal calculation. Therefore as in the continuous time case, it is useful and convenient to see if a constant sub-optimal LQR gain can be derived. The discrete time Riccati equation may for large $N$ or as $k \to -\infty$ converge to a constant or it may not. If the Riccati equation solution converges to a constant then this the limiting solution is denoted $\mathbf{P}_{d\infty}$. Obviously if $\mathbf{P}_{d\infty}$ exists it will be a solution to the discrete time *algebraic Riccati equation*, which is found by setting $\mathbf{P}_{d\infty} = \mathbf{P}_d(k+1) = \mathbf{P}_d(k)$,

$$\mathbf{P}_{d\infty} = \mathbf{R}_{1d} + \mathbf{F}^T[\mathbf{P}_{d\infty} - \mathbf{P}_{d\infty}\mathbf{G}(\mathbf{R}_{2d} + \mathbf{G}^T\mathbf{P}_{d\infty}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{P}_{d\infty}]\mathbf{F}. \tag{5.119}$$

This equation is in general a set of at most $n^2$ non-linear algebraic equations which may have symmetric, positive semi-definite solutions, but which may also have negative definite and even complex solutions. Therefore even though the limiting solution is a solution to the algebraic Riccati equation, it may have many other solutions and none of these may be positive semi-definite.

If the algebraic Riccati equation has a positive semi definite solution then the steady-statefeedback controller can be written in the form:

$$\mathbf{u}(k) = -\mathbf{K}_{d\infty}\mathbf{x}(k) = -[\mathbf{R}_{2d} + \mathbf{G}^T\mathbf{P}_{d\infty}\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{P}_{d\infty}\mathbf{F}\mathbf{x}(k). \tag{5.120}$$

As in the continuous case necessary and sufficient conditions will be stated for the steady-state optimal LQ regulator to give stable control. The result is given without proof. The reader is referred to Lewis (1986) and Bryson and Ho (1975) for more details.

The following holds true for an LQR regulator:

### Steady State Discrete LQR Regulator Theorem

If the system to be controlled, described by equation (5.103), is stabilizable then there is a bounded limiting solution to the Riccati equation and the limiting solution, $\mathbf{P}_{d\infty}$, is also a positive semi-definite solution to the algebraic Riccati equation in (5.119).

If furthermore the matrix pair $(\mathbf{F}, \sqrt{\mathbf{R}_{1d}})$ is observable, then there is one and only one positive semi-definite solution to the algebraic Riccati equation, which is then also the unique limiting solution to the Riccati equation independent of the final time boundary value, $\mathbf{P}_d(N)$. With the above conditions the closed loop system formed with the control law in equation (5.120) is asymptotically stable.

***Example 5.8*. Discrete LQR for the First Order System**

Consider a system with a discrete time state space equation given by:

$$x(k+1) = x(k) + Tu(k).$$

This corresponds to a discrete time integrator. A steady state discrete time LQ regulator for this system can be designed for the cost function:

$$J = \sum_{i=1}^{\infty} [x^2(i) + \rho u^2(i)],$$

by using the discrete time Riccati equation (5.117) with $R_{1d} = 1$ and $R_{2d} = \rho$. This gives:

$$p = 1 + 1 \cdot \left[p - \frac{p^2T^2}{\rho + T^2 p}\right] \cdot 1,$$

which reduces to the quadratic equation,

$$p^2 - p - \frac{\rho}{T^2} = 0.$$

The positive solution for this equation is

$$p = \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\frac{\rho}{T^2}},$$

and the optimal feedback gain is obtained from equation (5.120) as

$$K = [\rho + T^2 p]^{-1} T(p \cdot 1) = \frac{Tp}{\rho + T^2 p}.$$

Inserting the ARE solution gives for $K$,

$$K = \frac{2}{T + \sqrt{T^2 + 4\rho}}. \tag{5.121}$$

Note that a regulator which reduces the output of the system to zero as fast as possible is found when $\rho \rightarrow 0$, $K = 1 / T$, in which case the input power is allowed to be as large as necessary. A deadbeat regulator for the system gives the same result. The characteristic equation for the system becomes

$$\lambda - (1 - TK) = 0, \tag{5.122}$$

from which it is clear that $K = 1 / T$ gives the required deadbeat regulator gain. ❒

***Example 5.9.*** **Discrete LQR for the Double Integrator**

The double integrator system from Example 5.6 has a discrete-time model of the form

$$\mathbf{x}(k+1) = \mathbf{F}\mathbf{x}(k) + \mathbf{G}u(k) = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \mathbf{x}(k) + \begin{bmatrix} \frac{1}{2}T^2 \\ T \end{bmatrix} u(k).$$

The cost function weight matrices are chosen to be

$$\mathbf{R}_{1d} = \begin{bmatrix} r_p & 0 \\ 0 & r_v \end{bmatrix} \quad \text{and} \quad \mathbf{R}_{2d} = \rho.$$

For $r_p$ positive, the system with output matrix $\sqrt{\mathbf{R}_{1d}}$ and system matrix $\mathbf{F}$ is observable and therefore the ARE solution will be positive definite and hence regular. Therefore one may use the simpler form of the ARE which follows from equation (5.118). The ARE then becomes

$$\mathbf{P}_{d\infty} = \mathbf{R}_{1d} + \mathbf{F}^T[\mathbf{P}_{d\infty}^{-1} + \mathbf{G}\mathbf{R}_{2d}^{-1}\mathbf{G}^T]^{-1}\mathbf{F}.$$

Even for this relatively simple system the ARE is a set of three coupled quadratic algebraic equations for the elements of $\mathbf{P}_{d\infty}$. A numerical study where $r_p = 3$, $r_v = 4$ and $\rho = 1$ gives a closed loop system with a step response as depicted in Fig. 5.15. The system starts at rest at time zero and the LQ regulator drives

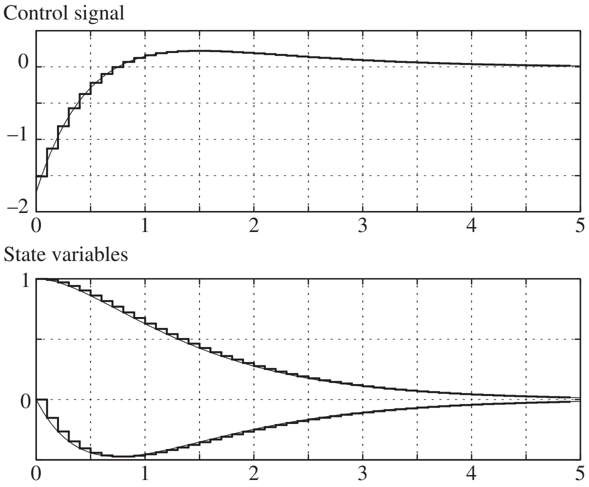


**Fig. 5.15** Response of a double integrator with a discrete time LQR regulator

the state to zero. The continuous time controller response with the same weight matrix values is shown for comparison. ❐

***Example 5.10.* Discrete LQR Control of the Two Link Robot**

A discrete time LQ controller is desired for the two link robot in Example 5.6. The design is based on a linearized discrete time model as given in Example 4.14:

$$\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t) + \mathbf{G}\mathbf{u}(t),$$

where

$$\mathbf{F} = \begin{bmatrix} 1003.6 & 20.024 & -0.60095 & -0.0040071 \\ 357.19 & 1003.6 & -60.152 & -0.60118 \\ -6.0182 & -0.040106 & 1002.1 & 20.014 \\ -602.38 & -6.0182 & 209.36 & 1002.1 \end{bmatrix} \cdot 10^{-3},$$

$$\mathbf{G} = \begin{bmatrix} 0.25048 & -0.48716 \\ 25.068 & -48.758 \\ -0.48716 & 1.3112 \\ -48.758 & 131.21 \end{bmatrix} \cdot 10^{-3}.$$

The sample time is $T = 0.02$ sec. If one attempts to use the same weight matrices as in the corresponding continuous time case in Example 5.6, it turns out that the control signal becomes much too large, so the weight on the control signal has to be increased. After some adjustments the following weights are used:

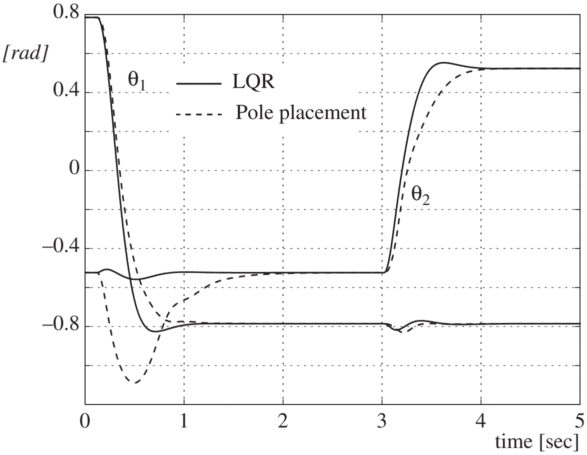


**Fig. 5.16** Robot link response with discrete time LQ regulator and error integration

$$\mathbf{R}_{1d} = \begin{bmatrix} 500 & 0 & 0 & 0 & 0 & 0 \\ 0 & 500 & 0 & 0 & 0 & 0 \\ 0 & 0 & 500 & 0 & 0 & 0 \\ 0 & 0 & 0 & 500 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1000 & 250 \\ 0 & 0 & 0 & 0 & 250 & 1000 \end{bmatrix},$$

$$\mathbf{R}_{2d} = 0.1 \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}.$$

The response is shown in Fig. 5.16 and is compared to the corresponding discrete-time pole placement controller from Chap. 4. In Fig. 5.17 the corresponding control signal is given. It is seen that the response is much faster for

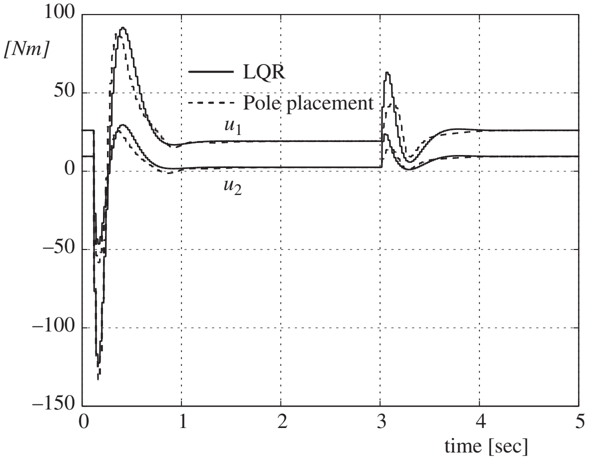


**Fig. 5.17** Robot control signal (motor torques)

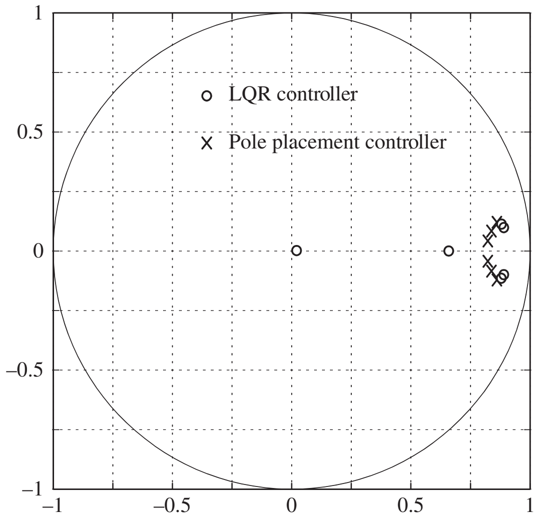


**Fig. 5.18** Closed loop poles for pole placement and LQ controllers

the LQ regulator than for the pole placement controller, even though the maximum control torque is the same. Also the cross coupling between the two links is substantially reduced with the LQ regulator.

The closed loop poles are shown for the pole placement and LQ controller in Fig. 5.18. It is seen that the LQ regulator gives poles that are faster than the pole placement controller but also have somewhat the same pole lengths.

The step response of the LQR is significantly better than for the pole placement. However this is with full state feedback, i.e., a controller that needs the measurement of position and speed of both links. It is interesting to see what the effect of using a pole placement observer is. This reduces the measurements to position only. The pole placement observer will be tuned the same as in Example 4.22. The pole position is such that the observer is ten times faster than the corresponding regulator. The combined controller including analog-digital converters and zero order hold networks is shown in Fig. 4.47. The controller gains are the same as before in the state feedback case:

$$\mathbf{K}_1 = [\mathbf{K}, -\mathbf{K}_i] - \begin{bmatrix} 768.47 & 78.09 & 2312.02 & 22.55 & -65.50 & -20.85 \\ 266.74 & 27.48 & 169.99 & 16.02 & -23.15 & -15.77 \end{bmatrix}.$$

The observer will be 10 times faster than the closed loop poles and the observer gain matrix is

$$\mathbf{L} = \begin{bmatrix} 1.942 & 6.53 \cdot 10^{-4} \\ 47.39 & -0.0162 \\ -9.51 \cdot 10^{-3} & 2.032 \\ -0.963 & 51.87 \end{bmatrix}.$$

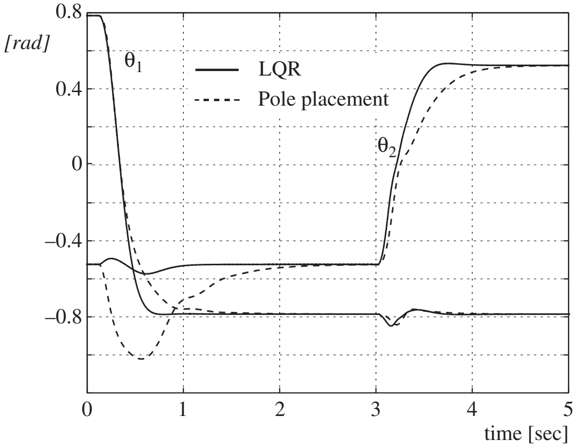


**Fig. 5.19** Observer based LQ controller for the two link robot

The resulting step response is given in Fig. 5.19. It is seen that the response is similar to the state feedback case and that is substantially better than the pole placement result. The corresponding control signal is shown in Fig. 5.20.

In order to investigate how the controller handles noise, because the observer gains have to be rather high, quantization noise is introduced as depicted in Fig. 4.51 on p. 339. In Fig. 5.21 the resulting response is shown for a 10 bit converter resolution. It is seen that the control signal is so noisy that it is probably useless in practice. If a 12 bit conversion is used the response is much more acceptable however there could be a problem with the observer amplification of noise, when such a fast observer is used. In the following chapters the question of noise will be dealt with systematically and observers designed to minimize the influence of noise in state estimation. ❐

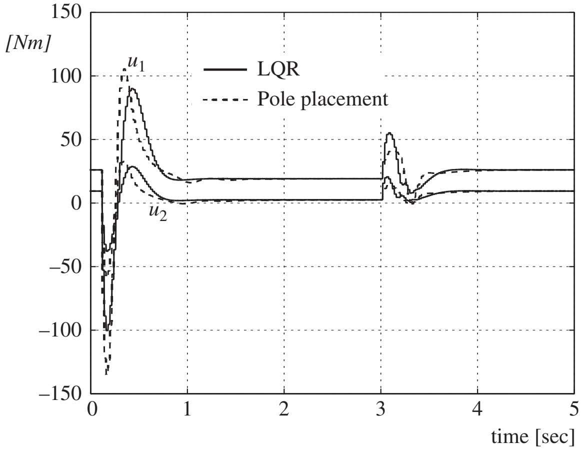


**Fig. 5.20** Observer based LQ control signal for the two-link robot (motor torques)

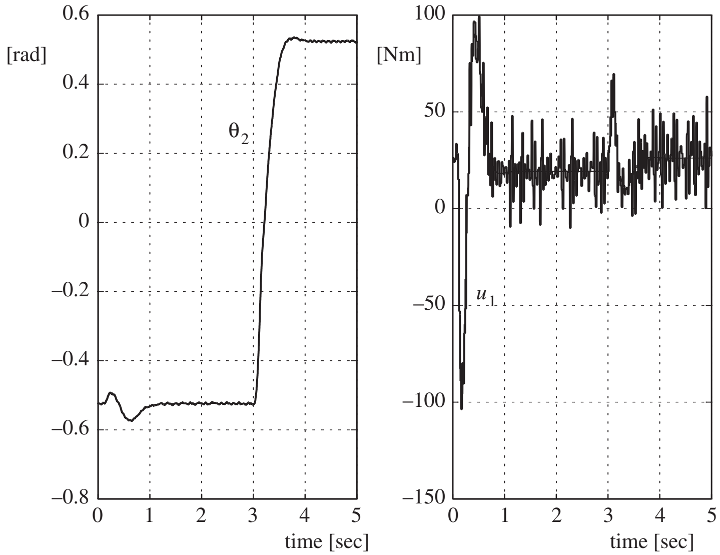


**Fig. 5.21** Response with fast observer (10 bit quantization)

## 5.7 Summary

This chapter has presented optimal control based on the formulation of a general optimization problem for a general cost function and a general nonlinear system equation. The main topics of the chapter are:

1. The solution of the general optimization problem.
2. The introduction of the quadratic cost function and the formulation of the linear quadratic regulator problem.
3. The solution of the LQR problem and the introduction of the Riccati equation.
4. Presentation of the requirement for a limiting solution of the Riccati equation and the stability of the steady-state LQ regulator.

This last point, i.e., the theorem on p. 318, is probably the most useful result in modern control theory for practical applications, since it guarantees the stability of a steady state optimal controller in the form of a constant-gain state feedback controller,

$$\mathbf{u}(t) = -\mathbf{K}_{\infty}\mathbf{x}(t) = -\mathbf{R}_2^{-1}\mathbf{B}^T\mathbf{P}_{\infty}(t). \tag{5.123}$$

Here, $\mathbf{P}_{\infty}$ is the solution to the Algebraic Riccati equation (ARE) in equation (5.82) on p. 317. This state-feedback controller is stable for a linear system that is stabilizable and detectable. A similar result holds for discrete time systems.

The ARE is a nonlinear matrix equation for the elements of the symmetric matrix, $\mathbf{P}_\infty$. There exist numerous numerical solvers for this, e.g. in MATLAB the function:

$$\texttt{[K, P, e]} = \texttt{lqr(A, B, R}_1\texttt{, R}_2\texttt{, N)}$$

gives the state feedback gain, the ARE solution and the resulting closed loop poles for a continuous time system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t),$$

with a performance index of the form

$$J = \int_0^\infty (\mathbf{x}^T\mathbf{R}_1\mathbf{x} + \mathbf{u}^T\mathbf{R}_2\mathbf{u} + 2\mathbf{x}^T\mathbf{N}\mathbf{u})dt.$$

Where also cross-products of states and control signals can be penalized. A similar routine is offered for discrete time problems.

While the gains in an optimal controller are arrived at using special methods, its structure is exactly like that of any state feedback regulator. Thus its eigenfrequencies are determined in the usual way by solving the equation

$$det(\lambda\mathbf{I} - \mathbf{A} + \mathbf{B}\mathbf{K}) = 0.$$

The LQR state feedback controller can be combined with an observer and in the subsequent chapters an optimal way of designing the observer will be discussed and finally combined with the LQR regulator to form an optimal output state feedback. This however requires the study of stochastic processes, since one has to handle process and measuring noise in an optimal way.

This chapter is concluded by emphasizing that:

For time-invariant, linear systems steady-state LQR state feedback leads to stable closed loop control provided the system is stabilizable and the matrix-pair ($\mathbf{A}$, $\sqrt{\mathbf{R}_1}$) is detectable.

## 5.8 Notes

### *5.8.1 The Calculus of Variations*

A standard problem in differential calculus is to determine the point at which a given curve is a maximum or minimum. A related problem (and a much more difficult one) is to determine a curve such that a functional (for example an integral along the curve) is maximized or minimized. In some cases it is also required that the optimizing function satisfy some auxiliary conditions or constraints, that the interval or region of definition can vary, that the curve is expressed in some parametric fashion and/or that the curve is a surface. All such problems fall into the domain of the calculus of variations and which is an important tool in modern physics and technology.

The calculus of variations emerged first as a collection of disparate problems and solutions in the period from 1680 to 1700. As might be expected, the first of these problems emerged with the invention of calculus itself by Isaac Newton and Gottfried Wilhelm Leibniz, published in the period from 1667 to 1687. The Leibniz form of calculus was based on a type of geometrical analysis combined with an algebra of differentials. This method of infinitesimal analysis about a point, together with Leibniz's suggestive notation, $y' = \frac{dy}{dx}$, made it possible for continental mathematicians to attack the problems of the calculus of variations more conveniently than was possible in England with Newton's fluxion notation: $\dot{y}$. Thus continental mathematicians were able to contribute most significantly to the initial development of the calculus of variations. In particular this is true of the Bernoulli family: Johann, Jakob and Daniel. It is of interest to note that the initial development of the subject was also closely related to the problems of mechanics as well as the fundamental interest in analysis (advanced calculus) itself.

One of the first problems considered was to derive the equation of the cycloid. A cycloid is the curve traced by a point on a circle which rolls without slipping on a straight line. This problem was proposed by Johann Bernoulli in 1696 and was solved by Newton, Leibniz, the Marquis de l'Hospital and the three members of the Bernoulli family mentioned above. Its solution was a differential equation in the arc length of the curve.

Leonard Euler extracted from various solutions offered by the Bernoulli family a unified approach to integral variational problems and published his results between 1732 and 1741. The unified approach found was what is now called the Euler-Lagrange equation: the fundamental necessary condition for the integral variational problem. Euler also observed that his derivation was not necessarily dependent on using any particular coordinate system, a fact which is of great importance in physical applications.

As mentioned in an earlier note, Euler recognized the promise of a young Frenchman, Joseph Louis Lagrange, and made sure of his employment as director of the Berlin Academy after Euler. Lagrange honored this trust by publishing in 1760 a method of deriving the Euler-Lagrange equation using the method of variations which is currently used for this purpose. Euler himself adopted this method and coined the name which is currently used to describe the entire discipline: 'calculus of variations'. Lagrange continued his researches in this area and this work culminated in 1788 with the publication of his analytical mechanics, the Lagrange equations of motion and the principles of virtual work and least action.

Beginning from a result proved by Adrian-Marie Legendre in 1786, Carl Jacobi investigated the question of when solutions of the single-integral variational problem actually lead to a minimum. Legendre was able to show that in addition to the Euler-Lagrange equation, a necessary condition is that the second derivative of the integrand with respect to the first derivative of the minimizing function is positive (in the domain of definition) as well as several other important results. He was also able to provide a simple geometrical picture of his theory in terms of the collection of extremals of the variational problem.

In the period from 1800 to 1850 considerable effort was expended in extending the single variable theory to multiple integrals. The most successful assault on these problems was published by Frederic Sarrus in a paper published in 1846. Later this work was extended by Augustine-Louis Cauchy, who also developed an acceptable theory of limits, and Francois Moigno. Sarrus' name is little remembered now but his contributions are considered to be some of the most original and important of his century. The background for this work was an overriding interest in the differential geometry of surfaces.

Karl Weierstrass was responsible for introducting rigor to study of the calculus of variations, the now well-known $\varepsilon - \delta$ terminology and the parametric theory of the first and second variations. The main advantage of the parametric approach is its ability to handle geometric applications, in particular the finding of geodesics (the shortest lines on surfaces). Weierstrass was slow in publishing his work so that many of his results were made public in his lectures at the University of Berlin from 1864 to about 1900. Among the students who attended Weierstrass' lectures were Sonya Kovalevski and David Hilbert. Hilbert and his students began the modern era in the calculus of variations with the introduction of topology in order to develop a macro-analysis. In addition Hilbert made large contributions to mathematical physics in general and quantum mechanics in particular with his introduction of Hilbert space.

## 5.9 Problems

### *Problem 5.1*

The Lagrangian for a conservative physical system is its kinetic energy minus its potential energy,

$$L = K.E. - P.E.$$

It can be shown that this quantity or its integral is a minimum for motion of the system over any time interval.

Consider a harmonic oscillator whose position is $x$, velocity is $\nu$, spring constant is $k$ and which has a mass, $m$.

a. Write down expressions for its kinetic and potential energy and for its Lagrangian.
b. Use the Euler-Lagrange equation to derive the equation of motion for the harmonic oscillator when it has no external driving force. What is the equation of motion when there is an external driving force, $F_{ext}$?

If the harmonic oscillator is disturbed from a zero equilibrium position then it will execute a harmonic movement.

c. How will the movement look in phase space, i.e., when the position of the oscillator is plotted against its velocity? What is the name of this figure? Give an equation for it.

### *Problem 5.2*

The Lagrangian for a conservative physical system is its kinetic energy minus its potential energy,

$$L = K.E. - P.E.$$

It can be shown that this quantity or its integral is a minimum for motion of the system over any time interval.

A ball is thrown upward with an initial velocity $\nu_0$ at an angle of $\theta_0$ degrees. The following integral is a minimum for its movement:

$$J = \int_{t_0}^{t_1} L\,dt.$$

The mass of the ball is $m$ and the acceleration of gravity is $g$.

a. Write down an expression for the Lagrangian of the ball in two dimensions ($x$ and $y$ coordinates).
b. Use the Euler-Lagrange equation to write down the equation of motion of the ball.
c. Integrate the equations of motion found to find expression which give the motion of the ball as functions of time. Is this what is to be expected?

### *Problem 5.3*

Assume a linear system which is described by the equation,

$$y(x,u) = ax + bu + c,$$

where $a$, $b$ and $c$ are constants, $x$ is the state and $u$ is the input. What is required is to find an input which minimizes the performance index:

$$L(x,u) = r_1x^2 + r_2u^2,$$

where $r_1 \geq 0$ and $r_2 > 0$ are constants.

a. Find the required input using the Euler-Lagrange equations.

   The problem has been formulated as a scalar problem. Now let $\mathbf{A}$ ($n \times n$ matrix), $\mathbf{B}$ ($m \times m$ matrix), $\mathbf{C}$ ($n \times n$ matrix), $\mathbf{R}_2 \geq 0$ ($n \times n$ matrix) and $\mathbf{R}_2 > 0$ ($m \times m$ matrix) be constant matrices and the state and input be vectors.

b. What is the index that must be minimized?
c. What is the input that minimizes this more general performance index?

### *Problem 5.4*

Consider the problem of finding the curve which has the minimum arc length between two points. The arc length between to arbitrary points ($t_0$, $a$) and ($t_1$, $b$) is

$$J = \int_a^b 1\ dt = \int_a^b \sqrt{1+\dot{x}^2}dt,$$

since $ds = \sqrt{1+((dx)/(dt))^2}dt = \sqrt{1+\dot{x}^2}dt.$

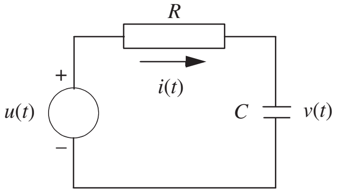
a. By considering this as a control problem, find the general equation for the curve.
b. Find the particular equation for the curve which goes through the two end points specified above.

### *Problem 5.5*

Use the information given in Problem 5.4 to find the shortest distance from a point to a line. The initial condition for the state is that $x(0) = 0$. The target line is given by the equation $y(t) = -st + y_{int}$, where $s$ and $y_{int}$ are constants.

### *Problem 5.6*

Consider the simple electrical circuit below. It is desired to find the input voltage which can charge the capacitor from a voltage $v(t = 0) = v_0$ to $v(t_f) = v_f$ while at the same time minimizing the energy dissipated in the resistor.



a. Write down the state equation of the system.
b. Write down the index which is to be minimized.
c. Write down the Hamiltonian of the system.
d. Find the time dependent input voltage which minimizes the energy dissapated in the resistor.

### Problem 5.7

A simplified state space model for an aircraft is given below:

$$\dot{v} = -k_1 v^2 - b_1 v^2 \sin(\alpha) + b_2 T,$$
$$\dot{h} = -K + b_3 v^2 \sin(\alpha),$$

where $v$ is the speed, $h$ is the height, $\alpha$ is the elevator angle, $T$ is the engine thrust, and the other quantities are positive constants.

A LQR regulator is to be designed for the linearized system where the control variables are the elevator angle and the engine thrust.

a. Linearize the model around the nominal operating point ($v_0$, $h_0$) and find the corresponding input trim variables ($\alpha_0$, $T_0$).
b. Are the linearized model's parameters dependent on $h_0$? Are they dependent on $v_0$?
c. Consider the incremental speed and height as the output of the system. Set up a quadratic expression for the calculation of the stationary optimal state feedback system. The relative weights for the speed and height are to be $\sigma_1$ and $\sigma_2$ while for the two inputs they are $\gamma_1$ and $\gamma_2$. The relative weight of the two states to the two inputs is be $\rho$. Call the matrix elements for the system $a_{11}$, $a_{12}$, etc. and $b_{11}$, $b_{12}$, etc. Do not attempt to solve for the matrix elements of the $\mathbf{P}_\infty$ matrix.

### Problem 5.8

A plant consists of a D.C. motor which has and angular velocity $\omega(t)$ which is driven by an input voltage $V_a(t)$. The system is described by the scalar state equation

$$\dot{\omega}(t) = -\alpha\omega(t) + \beta V_a(t),$$

where $\alpha$ and $\beta$ are positive constants. See Example 2.3 for the identification of the constants $\alpha$ and $\beta$.

a. Design an LQR regulator for this system which can minimize the performance index

$$J = \lim_{t\to\infty}\left[\int_0^t (\omega^2 + \rho V_a^2)\, dt\right].$$

### Problem 5.9

Consider the double integrator system $\mathbf{x} = [x_1, x_2]^T$:

$$\dot{x}_1 = x_2,$$

$$\dot{x}_2 = u,$$

with the performance index:

$$J = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{x}(t_1) + \int_{t_0}^{t_1} \left(\mathbf{x}^T(t)\mathbf{x}(t) +^2 (t)\right) \, dt.$$

a. Derive the Riccati equation as three coupled scalar differential equations and write the gain matrix in terms of the solutions to the Riccati equation. A solution to the Riccati equation above can be found numerically in Matlab.
b. Find an analytic expression for the solution to the steady-state Riccati equation for $t_1 \to \infty$ and find the optimal steady-state feedback gain.

### *Problem 5.10*

A D.C. motor position control system for an electronic throttle control is described by the state differential equation

$$\begin{bmatrix} \dot{\theta} \\ \dot{\omega} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -\alpha \end{bmatrix} \begin{bmatrix} \theta \\ \omega \end{bmatrix} + \begin{bmatrix} 0 \\ \beta \end{bmatrix} V_a,$$

where $\theta$ is the angular position, $\omega$ is the angular velocity, $\alpha$ is a viscous damping coefficient, $\beta$ is the D.C. motor velocity constant from the anchor voltage $V_a$. Consider the problem of regulating this system about its zero state where the controlled variable is the position, i.e.,

$$\mathbf{y} = \mathbf{C}\mathbf{x}, \qquad \mathbf{C} = [1 \quad 0].$$

a. Find and optimal steady state control for the system. Let the input weight-parameter for the LQR index be $\mathbf{R}_2 = [\mathrm{r}_2]$.
b. Sketch the movement of the poles of the closed loop system as a function of the weight parameter $\rho$.
c. If the motor damping parameter $\alpha$ is (is not) equal to zero, i.e., $[A]_{22} = -\alpha = 0$ ($-\alpha \neq 0$) then what does the root curve look like? Explain by inspection, do not calculate.
d. Use the numerical values $\alpha = 7.14$ rad/sec, $\beta = 286.3$ rad/(Vsec$^2$). Find the value of $\mathrm{r}_2$ such that the closed loop response has a 50 m sec response time. Determine by computation or simulation the response of the closed loop system given the initial conditions: $\theta(0) = 0.3$ rad and $\omega(0) = 0$ rad/sec.

### *Problem 5.11*

In a soft drinks depot one wishes to regulate the warehouse stock as accurately as possible. One counts the stock and makes a new order at a certain time every day. The new supply is received from the supplier the day after the order is given. If the available supply for day number $i$ is $x(i)$ and the order is $u(i)$ then it is true that

$$x(i+1) = x(i) + u(i-1).$$

a. Determine the systems state equations. The states are to be called $x_1$ and $x_2$.

Because of a heat wave the sales suddenly increase and then stays constant at a new value.

b. Find a regulator which can match the supply from the supplier to the new sales in at most 2 days. The regulator must at the same time bring the available supply back to the original level, $x_0$.
c. Find a state feedback which can minimize the criteria

$$J = \lim_{t\to\infty}\left[\left|\int_0^t [(x(i) - x_0)^2 + \rho u^2(i-1)]dt\right|\right].$$

d. In what limit of the parameter ρ does the LQR regulator become identical with the regular found under point c? What does this imply about the system input level?

### *Problem 5.12*

This exercise deals with a nonlinear optical fiber pulling process at NKT A/S (Nordic Cable and Wire Company, Inc.), Denmark. In an earlier project a linear control for this system was made, Hendricks et al. (1985). It has become clear that because of newer products which require a variation of the fiber diameter and high accuracy that a digital control of the process would be advantageous. where $D_p$ is the preform diameter (about 10–15 mm) and $V_p$ is the preform advance velocity. Both $D_p$ and $V_p$ are approximately constant.

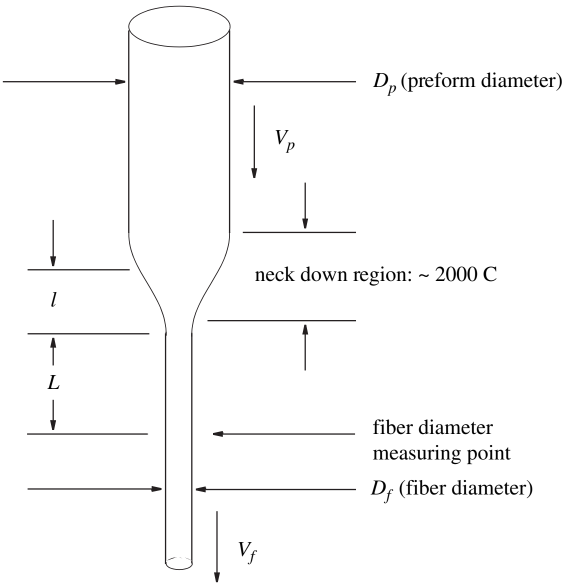
In the Fig. 5.22 a schematic drawing of the pulling process is shown. Optical fibers are drawn from a quarts rod or preform. This is done by heating the preform to a high temperature (about 2000°C) and pulling a fiber from the semi-molten zone of the preform called the neck down region. The control object is to control the diameter of the fiber, $D_f$, (usually 0.125–1 mm) by changing the drawing speed, $V_f$, (usually 0.2–1 m/s).

The fiber is formed over a distance which is $l$ but its diameter can be measured (optically) only after it has traveled a further distance $L$. The process is such that it obeys the mass conservation law:

$$D_f^2 V_f = D_p^2 V_p.$$

The dynamic process model has been identified earlier and is in the form of a transfer function for the process

**Fig. 5.22** Schematic drawing of an optical fiber pulling process



$$\frac{d_f(s)}{\nu_f(s)} = \frac{-K_v e^{-s\tau_d}}{1 + \tau_\nu s},$$

where s is the Laplace operator, $d_f$ is the incremental fiber diameter, $v_f$ is the incremental fiber pulling velocity, $\tau_d$ is the measurement time delay ($= L / V_f$), $\tau_v$ is the process relaxation time ($= l/V_f$), $L = 2.5$ cm, $l = 5.0$ cm and $K_v$ is the process amplification. $\tau_d$, $\tau_v$ and $K_v$ all depend on the large signal pulling velocity, $V_f$. The process gain is given by the expression,

$$K_v = \frac{dD_f}{2\,V_f},$$

where $k$ is a constant.

a. Write down an differential equation which describes the dynamics of the system when it is assumed that $\tau_d$, $\tau_v$ and $K_v$ are constant.
b. What is the time constant and eigenfrequency of the system?
c. The tolerance for $D_f$ is $\pm 1\%$. Show using the mass conservation law that $V_f$ is approximately constant in time.

Now the system is to be sampled in such a way that n samples are to be taken in the least of the characteristic times $\tau_d$ or $\tau_v$, i. e., $T =$ (least time constant)$/n$, where $T$ is the sampling period and $n$ is a whole number. As both $\tau_d$ and $\tau_v$ are scalars with respect to $1/V_f$, this leads to the state difference equation,

$$d(k+1) = a\ d(k) + (1-a)\ v(k-n),$$

where $a$ is a constant and $d(k) = d_f(k)$.

d. Show that the difference equation above is correct and give expressions for $a$ and $v$ in term of the original system parameters and constants. Also select a reasonable sampling time (or $n$) for the system and explain this choice.
e. Write down the state equations of this reduced system.
f. Design a steady state optimal regulator for the system.
g. What are the advantages of using a non-constant sampling time in the system description? Name the technical disadvantages of such a choice.

### *Problem 5.13*

Consider the damped harmonic oscillator given by the state equation $\mathbf{x} = [x_1, x_2]^T$:

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -2\varsigma\omega_n^2 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u(t),$$

with the performance index

$$J = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{x}(t_1) + \int_0^{t_1}(\mathbf{x}^T(t)\mathbf{x}(t) + ru^2(t))dt.$$

a. Derive the Riccati equation as three coupled scalar differential equations and write the gain matrix in terms of the solutions to the Riccati equation. A solution to the Riccati equation above can be found numerically in Matlab.
b. Find an analytic expression for the solution to the steady-state Riccati equation for $t_1 \to \infty$ and find the optimal steady-state feedback gain.

### *Problem 5.14*

A system is described by a bilinear state equation of the form

$$\dot{x}(t) = ax(t) + dx(t)u(t) + bu(t).$$

$x(t)$ and $u(t)$ are scalars and $a$, $b$ and $d$ are real constants. The cost function is defined by

$$J = \frac{1}{2}s(T)x^2(T) + \frac{1}{2}\int_0^T (qx^2(t) + ru^2(t))dt.$$

Find a set of two coupled differential equations that describe the closed loop system. $u(t)$ has to be eliminated from the equations.

***Problem 5.15***

A plant is given by

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u(t).$$

a. Verify the that plant is unstable and reachable.
b. Compute the open-loop optimal controller over the time interval [0, 1].

***Problem 5.16***

A discrete-time first order system is described by the state equation

$$x(k+1) = 0.74x(k) + 0.26\ u(k).$$

Design a time varying feedback controller that minimizes the cost function

$$J = \sum_{k=0}^{4}(x^2(k) + 0.2\ u^2(k)).$$

For $k = 0, \ldots, 4$ calculate the solution to the time varying Riccati equation, the value of the feedback gain and the value of the control signal, if $x(0) = 1$. Why is $P_d(4) = 0$?

***Problem 5.17***

Show that the matrix function,

$$\mathbf{P}(t) = \int_{t_0}^{t} e^{\mathbf{A}(t-t')}\mathbf{B}\mathbf{R}_2^{-1}\mathbf{B}^T e^{\mathbf{A}^T(t-t')}dt',$$

is the solution to the Lyapunov differential equation,

$$\dot{\mathbf{P}}(t) = \mathbf{A}\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A}^T + \mathbf{B}\mathbf{R}_2^{-1}\mathbf{B}^T,$$

with the initial condition $\mathbf{P}(t_0) = \mathbf{0}$.

a. Show that the unique solution to the Lyapunov differential equation is symmetric.
b. Show that the solution to the Riccati equation is symmetric if $\mathbf{P}(t_1)$ is chosen symmetric.

# Chapter 6
# Noise in Dynamic Systems

**Abstract** The purpose of this chapter is to present a brief review of the salient points of the theory of stochastic processes which are relevant to the study of stochastic optimal control and observer systems. Starting with a brief review of the main properties of random variables, this chapter goes forward to a detailed description of the main random process which is used as a model for noise in technological systems: Gaussian white noise. Both time domain and frequency domain descriptions of this important noise model are given. The main result of these considerations is the time dependent Lyapunov equation which is a compact way to express how white noise propagates through linear dynamic systems. Both continuous time and discrete time versions of this equation are given.

## 6.1 Introduction

Up to this point it has been assumed that control objects can be disturbed from a desired operating point by some sort of signal. In general it has been assumed tacitly that this signal is of a deterministic nature with zero mean, for example a sine wave. It has however never been specified what such a disturbance might be exactly. It is also clear that up to this point the entire description and analysis of state space systems has been model based: in general for deterministic linear systems. What is required now is a simple, mathematically tractable and sufficiently general model of system disturbances, both for internal process- and measurement-disturbances. One of the most useful disturbance models which can be used in a reasonably efficient manner is a random process or a filtered random process. The main reason for this is its mathematical simplicity and a secondly because of its proved scientific and technical utility. Such a random process in a deterministic system is generally called noise. What is needed here is a model for noise and a description of how it influences and propagates through dynamic systems.

It is well known that a number of different types of random processes or noise are important in a number of different systems. Electrical resistors are known to give Johnson noise which is proportional to the absolute temperature

E. Hendricks et al., *Linear Systems Control*, DOI: 10.1007/978-3-540-78486-9_6,
© Springer-Verlag Berlin Heidelberg 2008

at which they operate. Shot and 1/f noise influence strongly the operation of transistor amplifiers and must be taken into account when such amplifiers are used to amplify low level signals. Brownian motion or a random walk takes place when pollen particles are immersed in a liquid solution. Chemical mixing processes often have a large random component due to turbulence effects. Mechanical friction is thought to be due to random attractions between different areas of the surfaces which are in contact. Thus it is well established that random processes play a very important part in common technical systems. Many of these disturbances may be fairly accurately modelled as various kinds of random processes.

Unfortunately the theoretical background for treating random processes in dynamic systems is very complex for a number of reasons:

1. Advanced and difficult to use mathematical tools are required to handle signals which are often not differentiable or integrable in the ordinary sense.
2. It is difficult to do experiments with random processes: often this requires special equipment and long measurement times.
3. Nonlinearities introduce a very large perturbation of the linear theory and thus many technical systems are difficult to model adequately in a simple way.

For these reasons and others the treatment of noise in these notes has to be specialized to a particularly simple case: linear systems and Gaussian distributed noise. This also requires that this chapter will be constructed mainly on a heuristic basis. The mathematical formulations will be intuitively appealing, formally correct for only one class of systems and disturbances but rigorously incorrect. The results obtained however will be useful for many practical systems: experience with real systems over many years suggests this conclusion.

As it will take some time to review the necessary background for the treatment intended here, it is important here to define exactly the problem which will be treated. It is desired to be able to construct optimal feedback control systems and observers for the system:

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{B}_v(t)\mathbf{v}_1(t), \tag{6.1}$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{v}_2(t), \tag{6.2}$$

where $\mathbf{v}_1(t)$ and $\mathbf{v}_2(t)$ are random processes. $\mathbf{v}_1(t)$ is process noise and $\mathbf{v}_2(t)$ is measurement noise. $\mathbf{A}(t)$, $\mathbf{B}(t)$ and $\mathbf{C}(t)$ are the ordinary dynamic, input and output matrices. $\mathbf{B}_v(t)$ is a scaling and mixing matrix for the process noise input(s). It will turn out to be possible to construct optimum estimators and controllers for this simple model. Optimality is sought in the sense of minimum energy: minimum control energy and maximum signal to noise power ratio.

It is assumed here that the reader has a basic knowledge of the theory of probability. The background will be reviewed briefly however so that the reader who does not have this background will be aware of the material which should be learned. Alternatively, the reader who has the proper background should be able to review the material presented and be able to work constructively with it on a short time scale.

### 6.1.1 Random Variables

A continuous random variable, $X$, is a variable which can take on a continuous range of particular values, $x$, at random after no particular pattern. Such a variable cannot be described in any other way than by specifying its Probability Distribution Function (P.D.F.), $F(x)$, or its probability density function (p.d.f.), $f(x)$. As is the conventional practice, random variables themselves will be written as capital letters while small letters will be used for particular values of that variable. Thus for example

1. $x$ is a particular value of the random or stochastic variable $X$ and
2. $y$ is a particular value of the random or stochastic variable $Y$, etc.

The Probability Distribution Function (P.D.F.) (often shorted to distribution function) is defined by the expression:

$$F(x) = Pr(X \leq x), \quad -\infty < x < \infty, \tag{6.3}$$

where '$Pr$' means the probability of. Thus Eq. (6.3) may be read: the probability distribution function is the probability that the random variable $X$ is less than or equal to a specified particular value $x$. The P.D.F. has the following characteristics:

1. $F(x)$ is monotone increasing.
2. $\lim\limits_{x\to-\infty} F(x) = 0, \lim\limits_{x\to\infty} F(x) = 1.$
3. $F(x)$ is continuous from the right.

The probability density function (p.d.f.) is defined by the expression:

$$f(x) = \lim_{\Delta x\to 0} \frac{Pr(x \leq X \leq x + \Delta x)}{\Delta x} = \lim_{\Delta x\to 0} \frac{Pr(X \leq x + \Delta x) - Pr(X \leq x)}{\Delta x}. \tag{6.4}$$

Equation (6.4) in its first form is read: the probability density function is the probability that the random variable $X$ is between $x$ and $x$ plus $\Delta x$, divided by $\Delta x$ as $\Delta x$ goes to zero. The p.d.f. has the following properties:

1. $f(x)$ is non-negative.
2. $\int\limits_{-\infty}^{\infty} f(\chi)d\chi = 1.$
3. $Pr(X \in A) = \int_A f(\chi)d\chi, A \subset \Re.$

The P.D.F. and p.d.f. are related to each other in a way which is easy to express mathematically, one is the derivative of the other:

$$F(x) = \int_{-\infty}^{x} f(\chi)d\chi \tag{6.5}$$

and

$$f(x) = \frac{dF(x)}{dx}. \tag{6.6}$$

It is simple to give a geometrical interpretation of the meaning of the p.d.f. The product $f(x)\,dx$ is an element of area that represents an element of probability. This is the probability that in a single trial that $X$ falls in the infinitesmal interval $(x, x + \Delta x)$. Since the range of $x$ is divided up into elements of length $dx$ and the element of probability, $f(x)\,dx$, is the probability of falling into one of these elements, the density $f(x)$ is actually a function which shows the relative probability of $X$ falling into the interval $dx$ at $x$.

***Example 6.1.*** **Uniform Distribution**

A scalar stochastic variable which is uniformly distributed over an interval $a \le x \le b$ has a probability density function

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if a } \le x \le b \\ 0, & \text{if x } < a,\ x > b \end{cases}. \tag{6.7}$$

That this distribution function has property 2 of the p.d.f. can be shown by the calculation,

$$\int_{-\infty}^{\infty} f(\chi)d\chi = \int_{a}^{b} \frac{1}{b-a} d\chi = 1. \tag{6.8}$$

The probability distribution function for this distribution can be found using Eq. (6.5) on the density function in Eq. (6.7):

$$F(x) = \int_{-\infty}^{x} f(\chi)d\chi = \frac{1}{b-a} \chi|_{a}^{x} = \frac{x-a}{b-a}, \quad a \le x \le b, \tag{6.9}$$

which is monotone increasing. In addition $F(x)$ is equal to zero for $x < a$ and equal to 1 for $x > b$. This is of course as required to satisfy properties 1 and 2 of the P.D.F.

The Fig. 6.1 (top) shows an example of a uniform distributed random variable and how its distribution function is constructed. On the left of the figure a collection of randomly distributed numbers in the interval $0 \le X \le 1$ is shown. They have been plotted as a function of an arbitrary sample number, $n$, and 50 numbers have been selected. The second part of the figure shows the same numbers plotted on a vertical scale so that their uniformity in the interval above can be judged. At the far right in the top figure the distribution function for the numbers is shown. ❐

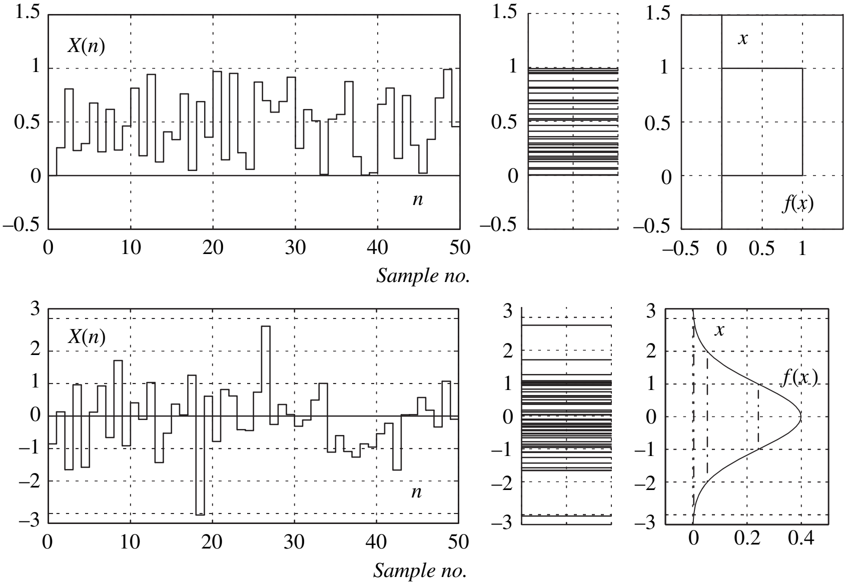


**Fig. 6.1** *Top*: a set of uniformly distributed random numbers, the same numbers plotted on a vertical scale and the corresponding probability density function. *Bottom*: a set of Gaussian distributed numbers, the same numbers plotted on a vertical scale and the corresponding probability density function

### *Example 6.2*. Gaussian Distribution

One of the most widely used and important distribution functions for both the theoretical and practical reasons is the Gaussian or normal distribution. This has to do with its relative simplicity and wide applicability. The form of the density function is given by the equation:

$$f(x) = He^{-h^2(x-a)^2}. \tag{6.10}$$

The three constants appear to be independent here but in reality the density function is defined by only two parameters. This can be shown by using property 3 of the p.d.f.:

$$\int_{-\infty}^{\infty} f(\chi)d\chi = H\int_{-\infty}^{\infty} e^{-h^2(x-a)^2}d\chi = H\int_{-\infty}^{\infty} e^{-h^2u^2}du = H\frac{\sqrt{\pi}}{h} = 1, \tag{6.11}$$

from which one can conclude that $H = h/(\sqrt{\pi})$. This gives the two parameter version of the normal probability density function,

$$f(x) = \frac{h}{\sqrt{\pi}}e^{-h^2(x-a)^2}. \tag{6.12}$$

The geometric interpretation of the parameters $h$ and $a$ can be found in Example 6.4.

An example of a set of Gaussian distributed numbers is shown in Fig. 6.1 (bottom) on the left. The middle figure is vertical scale which shows the same numbers so that their distribution is clear: note the strong concentration around the value zero. At the right is the distribution function. ❐

It is interesting to note that the concepts above can be immediately applied to a variable (random or deterministic) which changes in time. To do this consider a signal $y(t)$ as shown in Fig. 6.2. The probability that the amplitude $Y(t)$ of the given signal can be found between $y$ and $y + \Delta y$ is given by the relative probability,

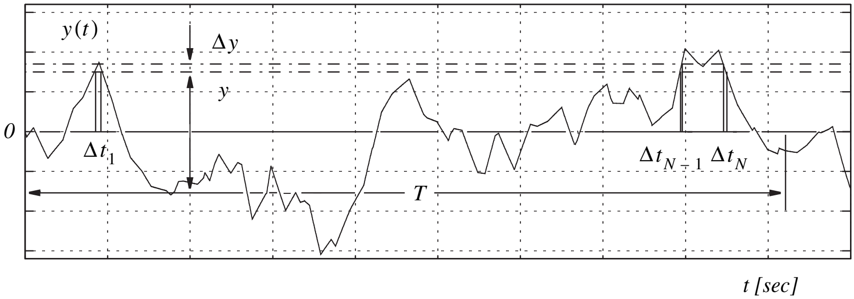


**Fig. 6.2** An example of how a probability density function can be found for a random signal. See text

$$\Delta Pr(y) = Pr(y < Y < y + \Delta y) = \lim_{T \to \infty} \left( \frac{\sum_i \Delta t_i}{T} \right), \tag{6.13}$$

where $\sum_i \Delta t_i$ is the total time spent by $y(t)$ in the interval $y$ to $y + \Delta y$ during the time interval $T$. Thus the probability density function gives the relative frequency of occurrence of the various instantaneous values of $y(t)$ in $(y, y + \Delta y)$ and hence

$$f(y) = \lim_{\substack{\Delta y \to 0 \\ T \to \infty}} \left( \frac{\sum_i \Delta t_i}{T \Delta y} \right). \tag{6.14}$$

It is quite possible to build analog circuits which can carry out this operation on real signals: they are called amplitude splitters and can be used to generate approximate p.d.f.'s. of arbitrary signals. The time dependent random processes which are the main subject of this chapter can be considered to be described by p.d.f.'s generated using Eq. (6.14), see Sect. 6.3.

### Discrete Random Variables

A discrete random variable is one which can only assume discrete values. The sample space for a discrete random variable can be discrete, continuous or a mixture of the two. A dice has for example a discrete sample space. The distribution and density functions above can be defined for discrete random variables in the same way as for continuous variables except that the integrations used for continuous variables must be replaced by summations. If the values of the discrete stochastic variable $X$ are denoted by $x_i$ then the probability distribution function, P.D.F. $F(x)$ can be written,

$$F(x) = \sum_{i=1}^{N} Pr(X = x_i)\ \upsilon(x - x_i), \tag{6.15}$$

where $\upsilon(\ldots)$ is the unit step function defined by

$$\upsilon(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}, \tag{6.16}$$

and $N$ is infinite for some systems. The probability density function, p.d.f. $f(x)$ corresponding to the P.D.F. in Eq. (6.15) is

$$f(x) = \sum_{i=1}^{N} Pr(X = x_i)\ \delta(x - x_i). \tag{6.17}$$

Clearly Eq. (6.17) has been obtained by differentiating Eq. (6.15) as suggested by the continuous case definition, Eq. (6.6).

## 6.2 Expectation (Average) Values of a Random Variable

The expectation, mean or average value of a random variable $X$ can be computed as the sum or integral of all the values which a random variable may assume weighted by the probability that particular value will be taken. The sum is used for a random variable that takes on discrete values while the integral is for a random variable which can assume a continuous range of values. The probability that $X$ takes on a value in a small interval $dx$ centered at $x$ is given by the probability density function $f(x)$. Therefore the expectation value of $X$ can be written,

$$E\{X\} = m_X = \langle X \rangle = \bar{X} = \int_{-\infty}^{\infty} \chi f(\chi) d\chi. \tag{6.18}$$

This is also called the mean value of $X$, the average value of $X$ or the first moment of $X$. In the limit as the number of observations of $X$ tends towards infinity, the average value of $X$ will tend to the value given above in the right hand expression.

It is possible to find the mean value of functions of a random variable in the same way as the mean value of the variable. If $Y$ is a random variable such that $Y = g(X)$ then $Y$ is also a random variable with a distribution function which can be derived from the distribution function for $X$. It is however not necessary to find this distribution function in order to find the mean value of $g(X)$. It is only necessary to calculate

$$E\{Y\} = E\{g(X)\} = \langle g(X)\rangle = \overline{g(X)} = \int_{-\infty}^{\infty} g(\chi)f(\chi)d\chi. \tag{6.19}$$

One of the most important functions of a random variable which one would be interested in finding is the mean value of the square of the variable, i.e., $g(X) = X^2$. This gives

$$E\{X^2\} = \langle X^2\rangle = \overline{X^2} = \int_{-\infty}^{\infty} \chi^2 f(\chi)d\chi. \tag{6.20}$$

This quantity is also called the second moment of $X$. The root-mean-square (rms) value of $X$ is the square root of $E\{X^2\}$:

$$X_{rms} = \sqrt{E\{X^2\}} = \sqrt{\int_{-\infty}^{\infty} \chi^2 f(\chi)d\chi}. \tag{6.21}$$

The variance of a random variable is the mean squared deviation of that variable from its mean value. It can be calculated as

$$\sigma_X^2 = \int_{-\infty}^{\infty} (\chi - E\{\chi\})^2 f(\chi)d\chi = E\{X^2\} - (E\{X\})^2. \tag{6.22}$$

The square root of the variance, $\sigma_X$, is the standard deviation of the random variable. The rms value of a random variable which has zero mean is the standard deviation.

***Example 6.3*. Expectation Values for the Uniform Distribution**

The mean value of a uniformly distributed random variable can be found from

$$E\{X\} = \int_{-\infty}^{\infty} \chi\, f(\chi)d\chi = \int_{a}^{b} \chi \frac{1}{b-a} d\chi = \frac{b+a}{2}. \tag{6.23}$$

The second moment is

$$E\{X^2\} = \int_{a}^{b} \chi^2 f(\chi)d\chi = \int_{a}^{b} \chi^2 \frac{1}{b-a} d\chi \;=\; \frac{b^2 + a^2 + ab}{3}. \tag{6.24}$$

The standard deviation is easily computed using the results of Eqs. (6.23) and (6.24) as

$$\sigma_X^2 = \langle X^2 \rangle - \langle X \rangle^2 = \frac{(b-a)^2}{12}. \tag{6.25}$$

For the uniformly distributed numbers in Fig. 6.1, p. 355 (top), $a = 0$, $b = 1$, thus the mean value is $1/2$ and the standard deviation is $1/\sqrt{12}$. ❐

***Example 6.4*. Parameters of the Normal Probability Density Function**

The two parameter version of the normal probability density function was found earlier in Example 6.2:

$$f(x) = \frac{h}{\sqrt{\pi}} e^{-h^2(x-a)^2}. \tag{6.26}$$

To find the physical interpretation of the parameters $h$ and $a$ it is useful to compute the mean value of the variable $X$:

$$\bar{X} = \int_{-\infty}^{\infty} \chi\, f(\chi)\; d\chi = \frac{ah}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-h^2u^2} d\chi = a. \tag{6.27}$$

Thus $a$ can be interpreted as being the mean value of the variable $X$.

The variance or square of the standard deviation of the normally distributed variable is

$$\sigma^2 = \frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} (\chi - a)^2\; e^{h^2(\chi-a)^2} d\chi = \frac{h}{\sqrt{\pi}} \int_{-\infty}^{\infty} u^2 e^{-h^2u^2} du = \frac{1}{2h^2}, \tag{6.28}$$

from which $h = 1/(\sigma\sqrt{2})$. The probability density function of a normally distributed variable with mean, $a$, and standard deviation $\sigma$ is thus

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}, \tag{6.29}$$

which is the well known bell curve centered around the mean value, $a$. The probability density function for the set of random numbers shown in Fig. 6.1, p. 355 (bottom) is shown on the right of the figure. For the numbers in the figure the mean value is zero and the standard deviation is 1. The dotted and dashed lines show the levels for $\pm 1\sigma$, $\pm 2\sigma$ and $\pm 3\sigma$.

For a normally distributed variable with mean value $a$, the probability distribution function is from property 1 of the p.d.f.:

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(\chi-a)^2}{2\sigma^2}} d\chi. \tag{6.30}$$

This equation cannot be integrated analytically but may be found in tables. ❐

### *Example 6.5*. Characteristics of the Gaussian Distribution

It is well to become acquainted with details of the Gaussian distribution and density functions. As related above, the Gaussian density function is given by the well known bell shaped curve given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \tag{6.31}$$

where the mean is given by $m = E\{X\}$ and its variance is $\sigma^2 = E\{(X - m)^2\}$ or $N(m, \sigma^2)$. It is completely specified by only these two parameters and this is one of the main reasons for its usefulness: its simplicity.

There are some properties of a Gaussian density function which are very useful for rough calculations. The probability that a Gaussian distributed variable, $X(t)$ (which takes on different values at different times, $t \geq t_0$, $t_0$ arbitrary), will be between $m - a$ and $m + a$ is given by the integral

$$Pr((m-a) \leq X(t) \leq (m+a)) = \int_{(m-a)}^{(m+a)} f(\chi)d\chi = erf\left\{\frac{a}{\sqrt{2}\sigma}\right\}, \tag{6.32}$$

where $erf(x) = 2F(x) - 1$, for $\sigma = 1$. This means that the following relations can be easily derived

$$Pr(|X(t) - m| \leq \sigma) = 0.6827, \tag{6.33}$$

$$Pr(|X(t) - m| \leq 2\sigma) = 0.9545, \tag{6.34}$$

$$Pr(|X(t) - m| \leq 3\sigma) = 0.9973. \tag{6.35}$$

This gives a simple and useful way to judge the value of σ when one is observing a realization of a Gaussian process, for example on an oscilloscope screen or a recorded time series. The expression in Eq. (6.35) shows that such a signal must remain within a $\pm 3\sigma$ window of its mean 99.73% of the time. Thus when observed over a reasonably long period compared to the smallest time period (1/ frequency) in the signal, it is likely that the signal will show its $\pm 3\sigma$ limits and then σ can easily be found.

It is also useful to know the amplitude of the probability density function at the different σ points above (i.e., distances from the maximum). The maximum amplitude of the bell shaped curve is

$$h_{max}(m) = f(x)|_{x=m} = \frac{1}{\sqrt{2\pi}\sigma}. \tag{6.36}$$

Assuming that $\sigma = 1$, at a distance of $\pm 1\sigma$ from $x = m$ (the maximum of the curve) its amplitude is 0.607 of its maximum height, $1/(2\pi)^{1/2} = 0.399$

or 0.242. At a distance of ±2σ it is 0.135 of its height or 0.0539. At a distance of ±3σ it is 0.011 of its height or 0.0044. At a distance of ±4σ it is about $5 \cdot 10^{-4}$ of its maximum height or $2 \cdot 10^{-4}$. See Fig. 6.3 below. ❐

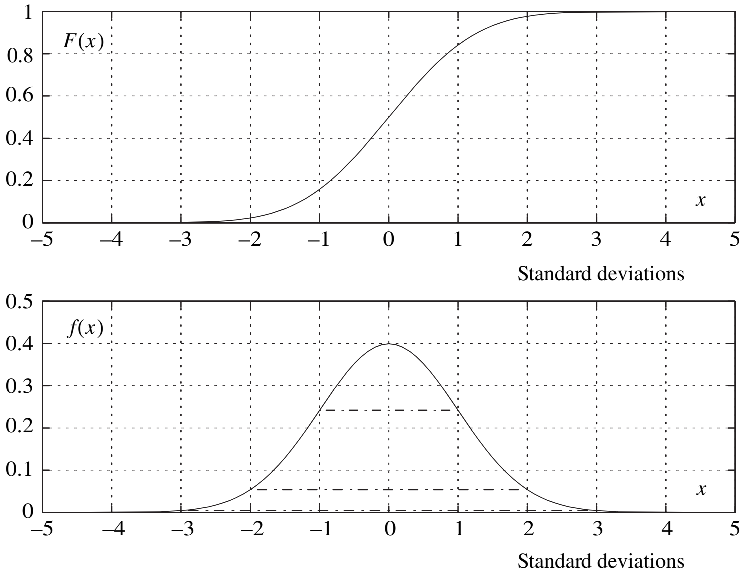


**Fig. 6.3** The P.D.F. (*top*) and p.d.f. (*bottom*) for a Gaussian distributed set of random numbers. The horizontal scale is in standard deviations. The *dotted* and *dashed lines* centered around $x = 0$ show the ±1σ, ±2σ , and ±3σ levels respectively

## *6.2.1 Average Value of Discrete Random Variables*

The expectation or average value of a discrete random variable can be easily defined on the basis of the definition of the discrete P.D.F. and p.d.f. It is easy to see that for a discrete random variable $X$,

$$E\{X\} = \sum_{i=1}^{N} x_i \, Pr(X = x_i). \tag{6.37}$$

For a function a random variable $g(X)$ this implies that

$$E\{g(X)\} = \sum_{i=1}^{N} g(x_i) \, Pr(X = x_i). \tag{6.38}$$

The generalization of the continuous to the discrete case is thus immediate.

### 6.2.2 Characteristic Functions

It is interesting to note that the results of Sect. 6.2 can be generalized to find the expectation value of arbitrary weight functions such as

$$E\{g(X,t)\} = \langle g(X,t)\rangle = \int_{-\infty}^{\infty} g(\chi,t)f(\chi)\ d\chi, \tag{6.39}$$

where $t$ is a parameter. In particular this generalization works for time functions. A particularly useful example is that represented by characteristic functions. This allows the calculation of the probability density function of functions of time.

If $g(X,t) = e^{jXt}$ in Eq. (6.39) then

$$\Phi(t) = E\{e^{jXt}\} = \langle e^{jXt}\rangle = \int_{-\infty}^{\infty} e^{j\chi t}f(\chi)\ d\chi. \tag{6.40}$$

This average is called the characteristic function of the probability density function $f(x)$ and is the Fourier transform of the probability density function. Clearly Eq. (6.40) is one member of a Fourier transform pair. The inverse Fourier transform is

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(\tau)e^{-j\tau x}d\tau, \tag{6.41}$$

where $f(x)$ is obviously the probability density function.

One of the most important properties of characteristic functions is that there is an exact one-to-one correspondence between P.D.F.s (or equivalently, p.d.f.s) and characteristic functions. This implies that if two distribution functions have the same characteristic functions then it may be immediately concluded that the distribution functions are the same. This fact has important consequences and uses in the theory of random variables and stochastic process. This is because Fourier transforms often make it possible to operate with algebraic instead of integral or differential equations when working with linear systems.

***Example 6.6.*** **Characteristic Function for a Uniformly Distributed Variable**

If the stochastic variable $X$ is uniformly distributed over the interval, $a \le x \le b$, then its characteristic function is given by

$$\Phi(t) = \frac{1}{b-a}\int_{-\infty}^{\infty} e^{j\chi t}d\chi = \begin{cases} 1, & \text{if } \ t\ =\ 0 \\ \frac{e^{\,jbt}\ -\ e^{\,jat}}{jt(b-a)}, & \text{if } \ t\ \neq 0 \end{cases}. \tag{6.42}$$

For a stochastic variable uniformly distributed over the symmetrical interval $-c \le X \le c$ this becomes

$$\Phi(t) = \begin{cases} 1, & \text{if } \ t = 0 \\ \frac{\sin(ct)}{ct}, & \text{if } \ t \neq 0 \end{cases}. \tag{6.43}$$

❐

One of the main uses of the characteristic function is to find the moments of a given random variable or distribution function. Differentiating equation (6.10) $r$ times with respect to $t$ gives

$$\frac{d^r}{dt^r}\Phi(t) = \int_{-\infty}^{\infty} (j\chi)^r e^{j\chi t} f(\chi)d\chi = E\{(jX)^r e^{jXt}\}. \tag{6.44}$$

For $t = 0$ this becomes

$$\frac{d^r}{dt^r}\Phi(t) = j^r E\{X^r\}. \tag{6.45}$$

Thus

$$\langle X^r \rangle = \frac{1}{j^r}\Phi^{(r)}(0). \tag{6.46}$$

***Example 6.7*. Characteristic Function for a Normally Distributed Variable**

The characteristic function of a normally distributed stochastic variable, $X$, with mean, $\mu$, and standard deviation, $\sigma$, (or $X \in N(\mu, \sigma^2)$) can be calculated from the expression:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{j\chi t} e^{-\frac{(\chi-\mu)^2}{2\sigma^2}} d\chi \ = \ e^{j\mu t - \frac{1}{2}\sigma^2 t^2}. \tag{6.47}$$

In order to find the moments of the distribution it is necessary to differentiate this expression with respect to $t$ to find

$$\frac{d\Phi(t)}{dt} = (j\mu - \sigma^2 t)e^{j\mu t - \frac{1}{2}\sigma^2 t^2} \tag{6.48}$$

and

$$\frac{d^2\Phi(t)}{dt^2} = \left[j^2\mu^2 - j\mu\sigma^2 - \sigma^2 - \sigma^2 t\frac{d}{dt}\right]e^{j\mu t - \frac{1}{2}\sigma^2 t^2}, \tag{6.49}$$

for the first two moments. From the equations above, it is clear that

$$m = E\{X\} = \frac{1}{j}\dot{\Phi}(0) = \mu \tag{6.50}$$

and

$$E\{X^2\} = \frac{1}{j^2}\ddot{\Phi}(0) = \mu^2 + \sigma^2. \tag{6.51}$$

An interesting case to consider is that for the sum of $N$ independent stochastic variables which are all normally distributed: $X_i \in N(\mu_i, \sigma_i{}^2)$, $i = 1, 2, \ldots, N$. From Eq. (6.47), the sum, $Y = X_1 + X_2 + \ldots + X_N$, has the characteristic function

$$\Phi(t) = \exp\left\{j(\mu_1 + \mu_2 + \ldots + \mu_N) - \frac{1}{2}(\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_N^2)\, t^2\right\}. \tag{6.52}$$

This shows that the sum of the stochastic variables has a distribution function which is $N(\mu_1 + \mu_2 + \ldots + \mu_N\,, \sigma_1{}^2 + \sigma_2{}^2 + \ldots + \sigma_N{}^2)$. This is a useful result which plays an important part in the analysis of multivariable noisy and coupled systems. ❐

It is sometimes easier to determine the characteristic function and then to transform it to obtain a probability density function than to determine the probability density directly. From the definition of $\Phi(t)$, Eq. (6.40), it is the expected value of $e^{jXt}$, where $X$ is the amplitude and $t$ is a parameter. The characteristic function of any function, $g(z)$, is the expected value of $e^{jtg(z)}$ (Cramér, 1945). Using this fact it is possible to determine the characteristic function and thus the probability density function of a deterministic periodic signal. The mean value of the periodic function is then according to Cramér,

$$\Phi(u) = \frac{1}{T}\int_0^T e^{juh(\tau)}d\tau, \tag{6.53}$$

where $h(t)$ is a periodic function with period $T$. Using this expression in Eq. (6.41) (thus carrying out the inverse Fourier transform) gives the probability density function of the deterministic signal:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\frac{1}{T}\int_0^T e^{juh(\tau)}d\tau\right]e^{-jux}\, du. \tag{6.54}$$

This result can be used to calculate the probability density function of common deterministic periodic signals.

***Example 6.8*. Probability Density Functions of Periodic Waveforms**

1. Sine Wave: $h(t) = A\,\sin(\omega_0 t), \omega_0 = 2\pi/T$.

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\frac{1}{T}\int_0^T e^{juA\,\sin(\omega_0\tau)}d\tau\right]e^{-jux}du \tag{6.55}$$

$$= \begin{cases} \frac{1}{\pi}(A^2 - x^2)^{-\frac{1}{2}}, & |x| \le A \\ 0, & |x| > A \end{cases}. \tag{6.56}$$

$f(x)$ is independent of the frequency and phase of the sine wave considered. Notice that the probability density function becomes very large close to the top of the sine wave. See Fig. 6.4, p. 365.

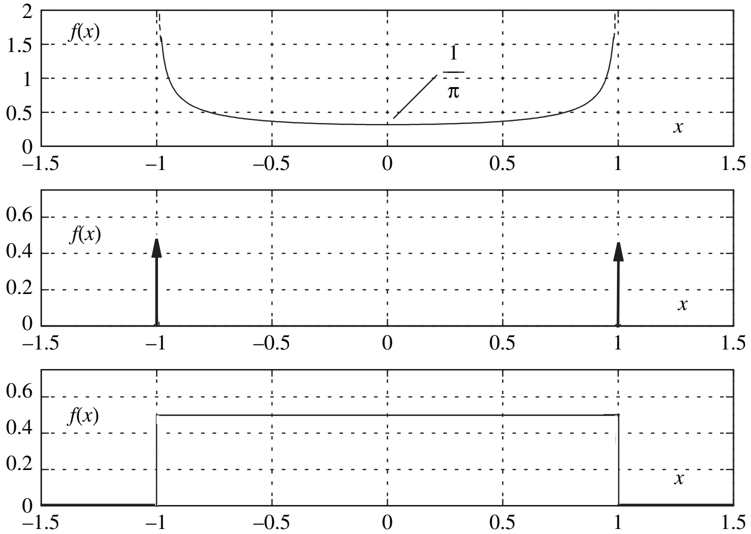


**Fig. 6.4** p.d.f.'s for sine, square and triangle waves of Example 6.8

2. Triangle Wave. The time function which describes a triangle wave ramping up and down between 0 and $A$ is

$$h(t) = \begin{cases} \frac{2A}{T}\left(t + \frac{T}{2}\right), & -\frac{T}{2} < t < 0 \\ \frac{2A}{T}\left(\frac{T}{2} - t\right), & 0 < t < \frac{T}{2} \end{cases}, \tag{6.57}$$

where $T$ is the period and $A$ is the maximum amplitude. From this formulation

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\frac{2e^{jAu}}{T}\int_0^{\frac{T}{2}} e^{-j2A\frac{u}{T}\tau}d\tau\right]e^{-jux}du \tag{6.58}$$

$$= \begin{cases} A^{-1}, & 0 \le x \le A \\ 0, & x > A \end{cases}. \tag{6.59}$$

See Fig. 6.4, p. 365.

3. Square Wave. A square wave stepping between $-A$ and $A$ with a period of $T$ can be expressed as

$$h(t) = \begin{cases} A, & 0 \le t \le \frac{T}{2} \\ -A, & \frac{T}{2} < t < T \end{cases} \tag{6.60}$$

The probability density function is then

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[ \frac{1}{T} \int_{0}^{\frac{T}{2}} e^{juA} dt + \frac{1}{T} \int_{\frac{T}{2}}^{T} e^{-juA} d\tau \right] e^{-jux} du \tag{6.61}$$

$$= \frac{1}{2} [\delta(x - A) + \delta(x + A)]. \tag{6.62}$$

This density function is two delta functions each centered at the maximum amplitude of the square wave. This is reasonable as the transition times assumed for the square wave edges are zero. See Fig. 6.4, p. 365. ❒

### 6.2.3 Joint Probability Distribution and Density Functions

It is often necessary to consider the simultaneous occurrence of several random variables in connection with control or other dynamic systems. This section is concerned with the background for describing and analyzing such systems. For simplicity only the occurrence of two random variables will be considered in this section. The generalization to larger numbers of variables is immediate.

In order to describe the statistical properties of multivariable stochastic systems it is useful to extend the concept of probability distribution and density functions to higher dimensions. This can be done by defining the joint probability distribution function,

$$F_2(x, y) = Pr(X \le x \text{ and } Y \le y), \tag{6.63}$$

where $X$ and $Y$ are two different random variables. The corresponding probability density function is

$$f_2(x, y) = \frac{\partial^2 F_2(x, y)}{\partial x \partial y}. \tag{6.64}$$

The individual separate distribution and density functions can be derived from the corresponding joint distributions. It is possible to collapse the joint P.D.F. and p.d.f. by either setting the upper limit in the definition to infinity (P.D.F.) or integrating over the dimension (p.d.f.) which is to be removed. The resulting P.D.F.s and p.d.f.s are called the marginal distribution or density functions. In the case of the two dimensional system considered here:

Distribution function:

$$F(x) = F_2(x, \infty), F(y) = F_2(\infty, y). \tag{6.65}$$

Density function:

$$f(x) = \int_{-\infty}^{\infty} f_2(x, \gamma) d\gamma \ , \ f(y) = \int_{-\infty}^{\infty} f_2(\chi, y) d\chi. \tag{6.66}$$

If $X$ and $Y$ are independent, the event $(X \leq x)$ is independent of the event $(Y \leq y)$ and the probability for joint occurrence of these events is the product of the probabilities of the separate events. This implies that

$$F_2(x, y) = Pr(X \leq x \text{ and } Y \leq y) \tag{6.67}$$

$$= Pr(X \leq x) Pr(Y \leq y) \tag{6.68}$$

$$= F_X(x) F_Y(y) \ \Leftrightarrow \ f_2(x, y) = f_X(x) f_Y(y). \tag{6.69}$$

In the same way as for scalar random variables it is possible to calculate expectation values for joint P.D.F.s and p.d.f.s. This is done using the same methodology as shown above. For example, to calculate the expectation value of a function $g(X, Y)$ of two vector random variables, the following integral has to be evaluated

$$E\{g(X, Y)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\chi, \gamma) f(\chi, \gamma) \ d\chi d\gamma. \tag{6.70}$$

An important special case of this is the covariance matrix which is the expectation of the product of the deviations of two random variables from their means

$$P_{XY} = E\{(X - E\{X\})(Y - E\{Y\})\} \tag{6.71}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\chi - E\{X\})(\gamma - E\{Y\}) \ d\chi d\gamma \tag{6.72}$$

$$= E\{XY\} - E\{X\} E\{Y\}. \tag{6.73}$$

The term $E\{XY\}$ is the second moment of $X$ and $Y$. The covariance matrix measures the absolute level of correlation of $X$ and $Y$. Often it is convenient to normalize the covariance with the standard deviations of $X$ and $Y$ in order to obtain information about the relative magnitude of variable correlation. This results in the so called correlation coefficient, $\rho$:

$$\rho = \frac{E\{XY\} - E\{X\} E\{Y\}}{\sigma_X \sigma_Y}. \tag{6.74}$$

If $X$ and $Y$ are independent, ρ is zero. If $X$ and $Y$ are completely correlated then $\rho = +1$ or if anti-correlated, –1. Attempts to make a linear approximation of $Y$ using $X$ results in an error in the approximation of $\sigma^2(1 - \rho^{\,2})$. Thus ρ is a measure of the degree of linear dependence between $X$ and $Y$.

It is of interest to find the expectation value of collections of more than two random variables. In particular sums and products of random variables are often required. Let $X_1, X_2, \ldots, X_n$ be $n$ random variables. The expectation value of the sum of the variables is equal to the sum of the expectations

$$E\{X_1 + X_2 + \ldots + X_n\} = E\{X_1\} + E\{X_2\} + \ldots + E\{X_n\}. \tag{6.75}$$

It is immaterial here if the n variables are independent or not. If however they are independent then the expectation value of the product is equal to the product of the expectation values:

$$E\{X_1 \cdot X_2 \cdot \ldots \cdot X_n\} = E\{X_1\} \cdot E\{X_2\} \cdot \ldots \cdot E\{X_n\}. \tag{6.76}$$

It is also true that the variance of the sum of random variables is equal to the sum of the variances if the variables are independent,

$$X = \sum_{i=1}^{n} X_i \;\Rightarrow\; \sigma_X{}^2 = \sum_{i=1}^{n} \sigma_{X_i}{}^2, \tag{6.77}$$

for independent $X_i$, $n = 1, 2, \ldots$ i,..., $n$.

There are some interesting and useful results that apply to random variables which are independent and identically distributed. Among these are the strong law of large numbers and the central limit theorem.

### Law of Large Numbers

This law states that if the random variables, $X_1, X_2, \ldots, X_n$, are independent and identically distributed, each with mean, $m$, then

$$Pr\left\{\lim_{n\to\infty} \frac{X_1 + X_2 + \ldots + X_n}{n} = m\right\} = 1. \tag{6.78}$$

This means that in the limit as the number of variables becomes large, the mean value of each of the random variables will converge in the mean to the same overall average value independent of the separate distribution type. This version of the law of large numbers is sometimes called the strong law of large numbers.

### Central Limit Theorem

If the random variables, $X_1$, $X_2$, …, $X_n$, are independent and identically distributed with mean $m$ and variance $\sigma^2$ then

$$\lim_{n\to\infty} Pr\left\{\frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}} \leq x\right\} = F(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{\chi^2}{2}}\, d\chi. \quad (6.79)$$

This equation has the interpretation that if the $n$ random variables are independent and identically distributed then their common distribution will converge to a normal distribution as the number of variables becomes large. In other words sums of large numbers of random variables give rise to normal distributions independent of the form of their original distributions. This is a very surprising conclusion which proves to be of good use in the analysis of noise in dynamic systems.

It can be shown that even small numbers of distributions of sums of identically distributed stochastic variables (say 3 or 4) are very close to normal distributions. This means that in any practical physical systems, even simple ones, one can expect to see approximative normal distributions as the overall effect or output. This again explains the use of normal distributions as models for uncertainty and/or noise in static and dynamic systems.

Both the law of large numbers and the central limit theorem can be proved using more advanced methods than are of interest in these notes. Other forms of both theorems can also be proved but these formulations are beyond the scope of this simple presentation.

### *Example 6.9.* A Cylindrical Joint Density Function

Consider a cylindrical joint probability density function which is specified by

$$f(x,y) = \begin{cases} \frac{1}{a}, & \textit{for } x^2 + y^2 \leq 1 \\ 0, & \textit{otherwise} \end{cases}, \quad (6.80)$$

where $a$ is a constant. See Fig. 6.5 below. The marginal density function for $y$ is then

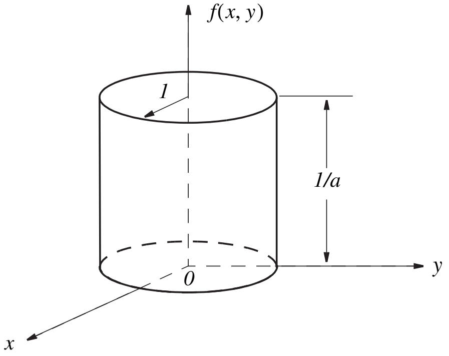


**Fig. 6.5** A cylindrical joint density function of two random variables

$$f(y) = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} f(\chi, \gamma)\, d\chi = \frac{2}{a}\sqrt{1-y^2}, \tag{6.81}$$

for $-1 \leq y \leq 1$ and $f(y) = 0$ elsewhere. Integrated over all $y$ this expression yields the result,

$$\frac{2}{a}\frac{\sqrt{\pi}}{2}\frac{\Gamma\left(\frac{1}{2}\right)}{\Gamma(1)} = \frac{2}{a}\frac{\sqrt{\pi}}{2}\frac{\sqrt{\pi}}{1} = 1, \tag{6.82}$$

which shows that the constant $a = \pi$. ❒

***Example 6.10.* Multivariable Uniformly Distributed Variables**

For a random variable $X_1$, uniformly distributed in one dimension (see example 6.1) between $a_1$ and $b_1$, the following probability distribution and density functions can easily be constructed.

$$F(x_1) = \begin{cases} 0, & \text{if } x_1 < a_1 \\ \dfrac{x_1 - a_1}{b_1 - a_1}, & \text{if } a_1 \leq x_1 \leq b_1, \\ 1, & \text{if } b_1 < x_1 \end{cases} \tag{6.83}$$

$$f(x_1) = \begin{cases} 0, & \text{if } x_1 < a_1 \text{ or } x_1 > b_1 \\ \dfrac{1}{b_1 - a_1}, & \text{if } a_1 \leq x_1 \leq b_1 \end{cases}. \tag{6.84}$$

The corresponding $n$ dimensional uniform distribution and density functions can be written:

$$F(x) = \begin{cases} 0, & \text{if } x_i < a_i \text{ for one i} \\ \prod\limits_{i=1}^{n} \dfrac{x_i - a_i}{b_i - a_i}, & \text{if } a_i \leq x_i \leq b_i \text{ for all i}, \\ 1, & \text{if } x_i > b_i \text{ for one i} \end{cases} \tag{6.85}$$

$$f(x) = \begin{cases} 0, & \text{if } x_i < a_i \text{ for one i} \\ \prod\limits_{i=1}^{n} \dfrac{1}{b_i - a_i}, & \text{if } a_i \leq x_i \leq b_i \text{ for all i}, \\ 0, & \text{if } x_i > b_i \text{ for one i} \end{cases} \tag{6.86}$$

To determine the marginal distribution function for all variables but $x_1$, one only has to let $x_1 \to \infty$. To determine the corresponding marginal density function one only has to integrate equation (6.86) over say, all $x_j$ for a selected $j$. The example is from Meditch (1969). ❒

**Joint Distribution and Density Function for Discrete Variables**

The development of distribution and density functions for multiple discrete variables is immediate given their definitions for single variables. The joint distribution function for the stochastic variables $X$ (with $N$ possible values of $x$) and $Y$ (with $M$ possible values of $y$) is

$$F_2(x, y) = \sum_{i=1}^{N} \sum_{j=1}^{M} Pr(X = x_i \text{ and } Y = y_j)\ \upsilon(x - x_i)\upsilon(y - y_j). \quad (6.87)$$

The joint p.d.f. for $X$ and $Y$ is accordingly

$$f_2(x, y) = \sum_{i=1}^{N} \sum_{j=1}^{M} Pr(X = x_i \text{ and } Y = y_j)\ \delta(x - x_i)\delta(y - y_j). \quad (6.88)$$

Moments of such distributions can be found by generalizing Eq. (6.37) and (6.38).

## 6.3 Random Processes

What is desired in this book is to use simple stochastic processes as mathematical models for disturbances and noise phenomena in multivariable control systems. The theory of stochastic processes deals with the description of phenomena which are subject to probabilistic laws in which time or some other variable is a parameter. While a rigorous treatment of this subject is well beyond the scope of this book, it is possible to present formally, in a heuristic manner, those parts of the theory which are necessary to the study of disturbances in control objects and control systems. This is the only goal which is intended here. For a more rigorous treatment the reader is encouraged to seek more fundamental treatments in the references.

### *6.3.1 Random Processes*

A stochastic or random process may be thought of as a collection or ensemble of functions of time indexed by a parameter $t$ all of whose values lie in some appropriate index set $I$, $\{X(t), t \in I\}$. $x(t)$ can be either a scalar or a vector. Most often in control system technology it is necessary for $x$ to be a vector.

The index set can be any abstract set but will here be either continuous or discrete time intervals.

In the continuous time case the index set is intervals of the time axis such as $I = \{t : 0 \leq t \leq T\}$ or more compactly $I = \{t : t \geq t_0\}$. Such a process is called a continuous time random or stochastic process.

In the discrete case the set is of discrete time instants $I = \{t_k : k = 0, 1, \ldots\}$, where $t_k < t_k$+*1*, which are not necessarily uniformly spaced (but most often will be so in what follows). To simplify the notation it is convenient to specify the index set as $I = \{k : k = 0, 1, \ldots\}$. In this case one can speak of a discrete time stochastic process.

A scalar stochastic process can be thought of as a collection or ensemble of time functions over the index set. The ensemble may contain either a non-denumerable or countable number of elements. Only the first case is of interest here. A random *n*th dimensional vector process is a collection of *n* scalar random processes.

A typical example of an ensemble for a random process is shown in Fig. 6.6. The ensemble is composed of *N* identical and identically prepared systems for which the some of the time series, realizations or sample functions are indicated.

If one considers only the time $t_1$ then a random variable, $X(t_1)$, is being considered. In the ensemble picture, P.D.F. and p.d.f. concepts are directly applicable. In fact the ensemble shown is for low pass filtered noise.

Since both time and ensemble concepts are used for a single random process, the stochastic process is in reality a function of two variables, one of which is from the time index set, *I*, and the other being of the sample space index. This can be indicated by the notation $\{X(\omega, t), \omega \in \Omega, t \in I\}$, where $\Omega$ is an appropriately defined sample space for the system under consideration. If *X* is a scalar process the for a fixed value of *t*, $X(\ldots, t)$ is a scalar valued function on the sample space, a random variable. If $\omega$ is fixed then $X(\omega, \ldots)$ is a scalar function

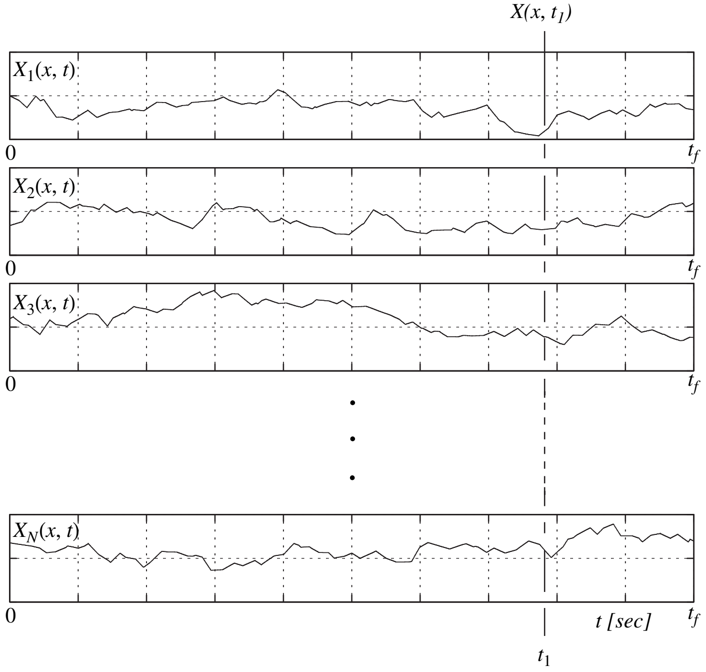


**Fig. 6.6** Example of an ensemble of a random process

of time which is called one possible realization or sample function of the process. If on the other hand, $X$ is a vector process for a fixed value of $t$, $\mathbf{X}(t)$ is a vector valued function on the sample space, a random vector. In the vector case, if ω is fixed then $\mathbf{X}(\omega, \ldots)$, is a vector valued function of time which is again called one possible realization or sample function of the process.

### *Example 6.11.* A Scalar Stochastic Process

A simple scalar stochastic process $\{X(t), t \geq 0\}$ can be described by $X(t) = A\sin(t)$, where $A$ is a continuous random variable which is uniformly distributed between +1 and –1,

$$f(A) = \begin{cases} \frac{1}{2}, & -1 \leq A \leq 1 \\ 0, & \text{otherwise} \end{cases}. \tag{6.89}$$

The ensemble of $X$ contains an infinite number of elements and the process is a continuous time stochastic process. The processes' ensemble consists of a group of sine waves of various amplitudes some of which are apparently phase shifted 180° (multiplied by negative $A$). Each of these sine waves is a realization (sample function) of the process. For a fixed value of $t = t_1 \geq 0$, $X(t_k)$ is a random variable, one point belonging to each realization. The Fig. 6.7 shows a picture of the ensemble of the process.

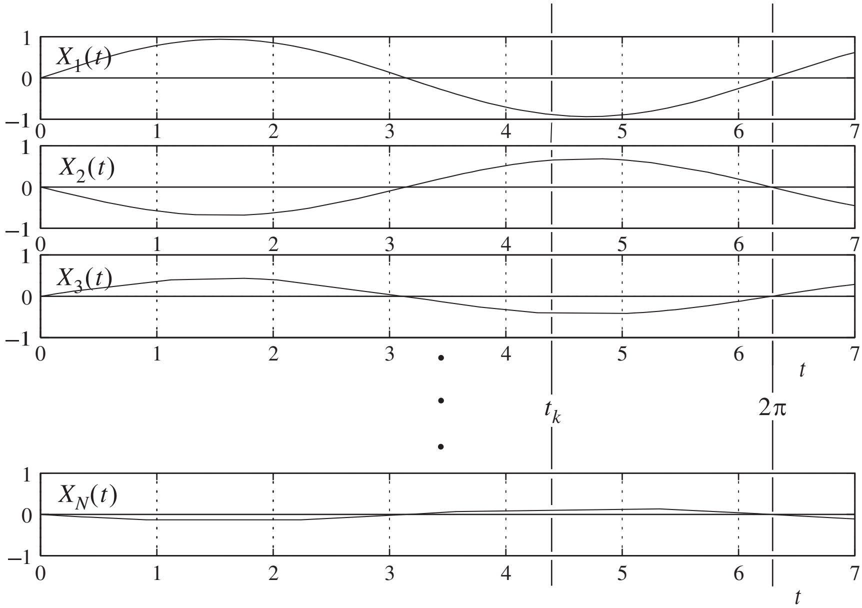


**Fig. 6.7** Ensemble for a (constructed) simple stochastic process composed of sine waves with the same frequency and uniformly distributed amplitudes

A discrete time stochastic process can be constructed by sampling $X(t)$ along the time axis, say in such a way that $\{t_k : t_k = (k/n)\pi, k = 0, 1, \ldots\}$, where $n$ is a positive integer. ❐

***Example 6.12.*** **A Discrete Time Gaussian Random Process**

Now consider a discrete time scalar stochastic process $\{x(k), k = 0, 1, \ldots\}$, where for each $k$, $x(k)$ is a zero mean random variable. The variance of $x(k)$ may depend on $k$ and thus the probability density of $y$ can be written

$$f(x,k) = \frac{1}{\sigma(k)\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2 x(k)}}, \ \sigma^2(k) = E\{x^2(k)\}. \tag{6.90}$$

It is also assumed that the value of $x(k)$ for a given $k$ is statistically independent of its values at all other sample points in the index set.

This process is called an independent Gaussian sequence. For any given value of $k$, the distribution of amplitudes across the ensemble is Gaussian. An example of one realization of such a process was given earlier in Fig. 6.1, p. 355 (top, second figure). In this case one should consider the sample number to be the time index $k$. The process in Fig. 6.1 has a mean value 0 and a standard deviation $\sigma = 1$. The example is adapted from Meditch (1969). ❐

Stochastic processes may be described in a way which is similar to that used for random variables except that the time dimension must be taken into account. In particular probability distribution and density functions are used. For a scalar random process the probability distribution function is

$$F(x_1, t_1) = Pr(X(t_1) \leq x_1) \tag{6.91}$$

and the corresponding probability density function is

$$f(x_1, t_1) = \frac{dF(x_1, t_1)}{dx_1}. \tag{6.92}$$

It is possible that the distribution function changes in time. To see how fast this happens, it is necessary to observe the same variable at two (or more) different times. The probability of the simultaneous occurrence of two different values in a certain range is given by the second order joint probability distribution function,

$$F_2(x_1, t_1; x_2, t_2) = Pr(X(t_1) \leq x_1 \text{ and } X(t_2) \leq x_2), \tag{6.93}$$

corresponding to the joint probability density function

$$f_2(x_1, t_1; x_2, t_2) = \frac{\partial^2 F(x_1, t_1; x_2, t_2)}{\partial x_1 \partial x_2}. \tag{6.94}$$

Higher order joint distribution and density functions may be defined but in general these are rarely used in control applications because of their complexity.

If two random processes are under consideration, the distribution and density functions which can be used to display their joint statistical characteristics are the functions

$$F_2(x, t_1; y, t_2) = Pr(X(t_1) \leq x \text{ and } Y(t_2) \leq \mathrm{y}), \tag{6.95}$$

$$f_2(x, t_1; y, t_2) = \frac{\partial^2 F(x, t_1; y, t_2)}{\partial x \partial \mathrm{y}}. \tag{6.96}$$

The generalization of the methodology above to vector random processes is immediate. For a continuous stochastic vector process $\{\mathbf{x}(t), t \in I\}$, $I = \{t : 0 \leq t \leq T\}$, the joint probability distribution function of $m$ random $n$ vectors can be written,

$$F(x_1, t_1; \ldots; x_n, t_n) = Pr(X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n), \tag{6.97}$$

with the corresponding joint probability density function,

$$f(x_1, t_1; \ldots; x_n, t_n) = \frac{\partial^n F(x_1, t_1; \ldots; x_n, t_n)}{\partial x_1 \ldots \partial x_n}. \tag{6.98}$$

Such complications will not however be of direct interest in what follows. What is more important is how to deal with the moments of such distribution and density functions. It is these moments which have the greatest use in technological and scientific applications.

### 6.3.2 Moments of a Stochastic Process

The reason why it is the moments of a random process which are most useful in real applications is because it is only these which can be realistically characterized in any reasonably simple and compact way. It is in general not possible to write down the equation(s) for any particular realization of a process and certainly not for any particular ensemble. What it is possible to write down compactly is the result of taking an average over an ensemble. This is fortunate because it turns out that such ensemble averages can be related to time averages which can be practically measured.

In terms of the first density function, $f_1(x, t)$, the nth moment of a stochastic process, $X(t)$, at a given $t \in T$ is defined by the integral

$$m_n(t) = E\{X^n(t)\} = \int_{-\infty}^{\infty} \chi^n f_1(\chi, t) d\chi. \tag{6.99}$$

The first moment is the mean or average value of the stochastic process $X(t)$ at $t$,

$$m(t) = m_X(t) = E\{X(t)\} = \int_{-\infty}^{\infty} \chi f_1(\chi, t) d\chi. \tag{6.100}$$

Here the notation for the mean, $m_X(t)$, or commonly, $m(t)$, has been introduced. The mean square of $X(t)$ at $t$ is given by $Q(t) = m_2(t)$. The central moments of the process also play an important part in the theory of stochastic processes. The $n$th central moment of $X(t)$ at $t$ is

$$\mu_n(t) = E\{(X(t) - m(t))^n\} = \int_{-\infty}^{\infty} (\chi(t) - m(t))^n f_1(\chi, t) d\chi. \tag{6.101}$$

Of particular importance is the variance of $X(t)$ at $t$, $\mu_2(t) = Q(t)$, which is the square of the standard deviation of $X$,

$$\mu_2(t) = \sigma_X^2 = \sigma^2 = E\{(X(t) - m(t))^2\} = \int_{-\infty}^{\infty} (\chi - m)^2 f_1(\chi, t) d\chi. \tag{6.102}$$

As for random variables, the joint moments of $X(t)$ are of great utility. The moments of $X(t)$ defined in terms of its second order density function are in effect the joint moments of two random processes. The joint moment of $X(t)$ at $t_1$ and $t_2$ is

$$m_{nm}(t_1, t_2) = E\{X_n(t_1) X_m(t_2)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \chi_1 \chi_2 f_2(\chi_1, t_1; \chi_2, t_2) d\chi_1 d\chi_2. \tag{6.103}$$

The quantity $C_{XX}(t_1, t_2) = m_{XX}(t_1, t_2)$ is called the autocorrelation function for $X(t)$ and is in general a function of both $t_1$ and $t_2$. Sometimes this name is shortened to correlation function when the syntax makes it clear what is meant. In the same framework it is possible to talk about two different random variables and to consider how strongly they are correlated. This leads to the cross correlation function, defined by

$$C_{XY}(t_1, t_2) = E\{X(t_1) Y(t_2)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \chi \gamma f_2(\chi, t_1; \gamma, t_2) d\chi d\gamma. \tag{6.104}$$

The central mixed moments have an important role in what follows. The autocovariance function for $X(t)$ is defined by

$$\mu_{XX}(t_1, t_2) = R_{XX}(t_1, t_2) = E\{[X(t_1) - m(t_1)][X(t_2) - m(t_2)]\} \tag{6.105}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\chi_1 - m_1)(\chi_2 - m_2) f_2(\chi_1, t_1; \chi_2, t_2) d\chi_1 d\chi_2. \tag{6.106}$$

When $t_1 = t_2 = t$, this quantity becomes the variance of $X(t)$, $\sigma_X{}^2$. The normalized version of the autocovariance function is the correlation coefficient function

$$\rho_{XX}(t_1, t_2) = \frac{\mu_{XX}(t_1, t_2)}{\sigma_X(t_1)\sigma_X(t_2)} = \frac{R_{XX}(t_1, t_2)}{\sigma_X(t_1)\sigma_X(t_2)}. \tag{6.107}$$

The interpretation of this function as a measure of linear dependence, introduced in connection with random variables, carries over to random processes as well. In this same vein the cross covariance function is defined as

$$\mu_{XY}(t_1, t_2) = R_{XY}(t_1, t_2) = E\{[X(t_1) - m_X(t_1)][Y(t_2) - m_Y(t_2)]\}. \tag{6.108}$$

***Example 6.13.* Autocovariance Function of a Scalar Stochastic Process**

For the continuous time scalar stochastic process of Example 6.11 the mean value of the process is

$$m(t) = E\{X(t)\} = E\{A\sin(t)\} = 0, \text{ for all } t, \tag{6.109}$$

since $A$ is uniformly distributed symmetrically around 0, between –1 and +1. $A$ is thus a zero mean random variable. The covariance function of $X(t)$ is

$$R_X(t_1, t_2) = E\{A^2 \sin(t_1)\sin(t_2)\} \tag{6.110}$$

$$= E(A^2)\sin(t_1)\sin(t_2) \tag{6.111}$$

$$= \frac{1}{3}\sin(t_1)\sin(t_2). \tag{6.112}$$

$$\Rightarrow Q(t) = \frac{1}{3}(\sin(t))^2, \tag{6.113}$$

where $Q(t)$ is the variance of $X(t)$. The example is from Meditch (1969). ❐

It is a straightforward process to extend the concepts above to the case of vector stochastic processes. Consider a vector stochastic process, $\mathbf{X}(t)$, $\{t \in I\}$, the mean of the process is the mean of the components of the process,

$$\mathbf{m}_X(t) = E\{\mathbf{X}(t)\} = [E\{X_1(t)\}, E\{X_2(t)\}, \ldots, E\{X_n(t)\}]^T. \tag{6.114}$$

The second order joint moment matrix or autocorrelation matrix is

$$C_X(t_1, t_2) = E\{\mathbf{X}(t_1)\mathbf{X}^T(t_2)\} \tag{6.115}$$

$$= \begin{bmatrix} E\{X_1(t_1)\ X_1(t_2)\} & E\{X_1(t_1)\ X_2(t_2)\} & \ldots & E\{X_1(t_1)\ X_n(t_2)\} \\ E\{X_2(t_1)\ X_1(t_2)\} & E\{X_2(t_1)\ X_2(t_2)\} & \ldots & E\{X_2(t_1)\ X_n(t_2)\} \\ \ldots & \ldots & \ldots & \ldots \\ E\{X_n(t_1)\ X_1(t_2)\} & E\{X_n(t_1)\ X_2(t_2)\} & \ldots & E\{X_n(t_1)\ X_n(t_2)\} \end{bmatrix}, \tag{6.116}$$

where the last matrix has been written out in detail for clarity. Each matrix element of $\mathbf{C}_X(t_1, t_2)$ is a scalar joint moment function.

The covariance matrix of the process is

$$\mathbf{R}_X(t_1, t_2) = E\{[\mathbf{X}(t_1) - \mathbf{m}(t_1)][\mathbf{X}(t_2) - \mathbf{m}(t_2)]^T\}. \tag{6.117}$$

It can be shown that the covariance matrix, $\mathbf{R}_X(t_1, t_2)$, and the second order joint moment matrix, $\mathbf{C}_X(t_1, t_2)$, have the following important properties:

$$1.\ \mathbf{R}_X(t_1, t_2) = \mathbf{R}_X^T(t_1, t_2), \text{ for all } t_1, t_2, \tag{6.118}$$

$$2.\ \mathbf{C}_X(t_1, t_2) = \mathbf{C}_X^T(t_1, t_2), \text{ for all } t_1, t_2, \tag{6.119}$$

$$3.\ \mathbf{Q}_X(t) = \mathbf{R}_X(t, t) \geq 0, \text{ for all } t, \tag{6.120}$$

$$4.\ \mathbf{Q}'(t) = \mathbf{C}_X(t, t) \geq 0, \text{ for all } t, \tag{6.121}$$

$$5.\ \mathbf{C}_X(t_1, t_2) = \mathbf{R}_X(t_1, t_2) + \mathbf{m}_X(t_1)\mathbf{m}_X^T(t_2), \text{ for all } t_1, t_2. \tag{6.122}$$

The notation $\mathbf{M} \geq 0$, where $\mathbf{M}$ is a square symmetric matrix, means that $\mathbf{M}$ is positive semi-definite (or non-negative definite), i.e.,

$$\mathbf{x}^T\mathbf{M}\ \mathbf{x} \geq 0, \text{ for } \mathbf{x} \in \Re^n. \tag{6.123}$$

### 6.3.3 Stationary Processes

In order to establish a useful connection between the physical implications of the theory of stochastic processes and technological applications it is necessary to introduce two important concepts: stationarity and ergodic processes. Stationarity makes it possible to attach reliable statistical characteristics to stochastic processes and ergodicity makes it possible to connect realistically achievable measurements to the basically theoretical ensemble concept which has been described earlier. The concept of ergodicity will be described in the next section.

A stochastic process, $\mathbf{X}(t)$, $t \in T$, is said to be strictly stationary if its probability distribution or density functions are invariant for an arbitrary change of the time parameter, i.e.,

$$F_n(x_1, t_1; \ldots; x_n, t_n) = F_n(x_1, t_1 + \tau; \ldots; x_n, t_n + \tau), \tag{6.124}$$

for all $(t_j + \tau) \in T$, $j = 1, 2, \ldots, n$, each $n$ and an arbitrary time increment, $\tau$. The sign of the time increment has not been specified thus the probability densities are only dependent on time differences, not absolute values. More abstractly, the joint probability distribution and density functions which characterize a strictly stationary stochastic process are invariant with respect of a change of time origin. In the definition of a strictly stationary stochastic process above, it is clear that the measurable qualities of the distribution, its moments, are not as yet defined. It is important here that this definition be specialized to make possible a direct application to technical problem formulations.

A stochastic process, $\mathbf{X}(t)$, $t \in T$, is wide sense stationary if its first and second order moments are finite and if its second order moment is only dependent on the time difference, $t_1 - t_2$. That is if

$$1.\ \mathbf{m}(t) = \langle E\{\mathbf{X}(t)\}\rangle = \text{constant (in its components)}, \tag{6.125}$$

and

$$2.\ \mathbf{C}_X(t,t) = E\{\mathbf{X}(t)\ X^T(t)\} < \infty,\ \text{in its elements}, \tag{6.126}$$

and its second order moment matrix has the quality that

$$3.\ \mathbf{C}_X(t_1,t_2) = E\{\mathbf{X}(t_1)\ X^T(t_2)\} = \mathbf{C}_X(t_1 - t_2), \tag{6.127}$$

then the stochastic process, $\mathbf{X}(t)$, is said to be wide sense stationary. In the list above, another condition is often substituted for 3 or added to the list which is

$$4.\ \mathbf{R}_X(t_1,t_2) = E\{[\mathbf{X}(t_1) - \mathbf{m}(t_1)][\mathbf{X}(t_2) - \mathbf{m}(t_2)]^T\} = \mathbf{R}_X(t_1 - t_2). \tag{6.128}$$

This is a useful but in reality unnecessary clarification. 1, 2 and 3 are sufficient in themselves or points 1, 2 and 4.

Clearly any strictly stationary process with a finite second order moment matrix is also wide sense stationary. The converse is not in general true except for an important exception, a Gaussian process, which is completely specified by its mean and covariance functions. Thus a wide sense stationary Gaussian stochastic process is also strictly stationary.

### *Example 6.14.* A Constant Process

A constant process is given by

$$X(t) = Y, \tag{6.129}$$

where $Y$ is a random variable. $X(t)$ is obvious strictly stationary and will also be wide sense stationary if

$$E\{Y^2\} < \infty. \tag{6.130}$$

❐

***Example 6.15.* A Dynamic Process**

Let $\{X(t), t \geq 0\}$ be the scalar stochastic process which is defined by the differential equation,

$$\dot{x} = -\omega_0 x + v, \tag{6.131}$$

where $\omega_0 = 1/\tau$, $\tau$ is the time constant, v is a zero mean random variable which has a constant standard deviation $\sigma$. Assuming that x(0) = 0, any realization of the process has the general form,

$$x(t) = \frac{v}{\omega_0}(1 - e^{-\omega_0 t}), \tag{6.132}$$

for all $t \geq 0$. Moreover

$$C_X(t, t) = \frac{\sigma^2}{\omega_0^2}(1 - e^{-\omega_0 t})^2, \tag{6.133}$$

but as $\mathrm{E}\{x(t)\} = 0$,

$$R_X(t_1, t_2) = E\{X(t_1)X(t_2)\} = \frac{\sigma^2}{\omega_0^2}(1 - e^{-\omega_0 t_1})(1 - e^{-\omega_0 t_2}), \tag{6.134}$$

which is clearly not stationary. However for large times

$$R_X(t_1, t_2) \rightarrow \frac{\sigma^2}{\omega_0^2}, \text{for } t_1, t_2 \rightarrow \infty. \tag{6.135}$$

Thus in the steady state, the stochastic process $X(t)$ becomes wide sense stationary. ❐

### *6.3.4 Ergodic Processes*

A process is said to be ergodic if a certain statistic (or moment) calculated by averaging over the members of an ensemble at a fixed time can also be calculated by time averaging over any particular sample function or realization, given sufficient time.

Consider the stochastic process, $\{X(t), -(T/2) \leq t \leq (T/2)\}$ and see Fig. 6.8. If one pictures a set of $N$ (large) random functions stacked up on top of each other on $N$ time axes (each of length $T$) placed above each other, an ensemble average of $X(t)$ (or some function of it) can be calculated by averaging over the $N$ ensemble members at a given fixed time, $t_1$. This is a scientific average, related to the ensemble concept. On the other hand if the $N$ ensemble members or sample functions are set end to end and an average taken over the time interval $NT$ then this time average should be representative of the

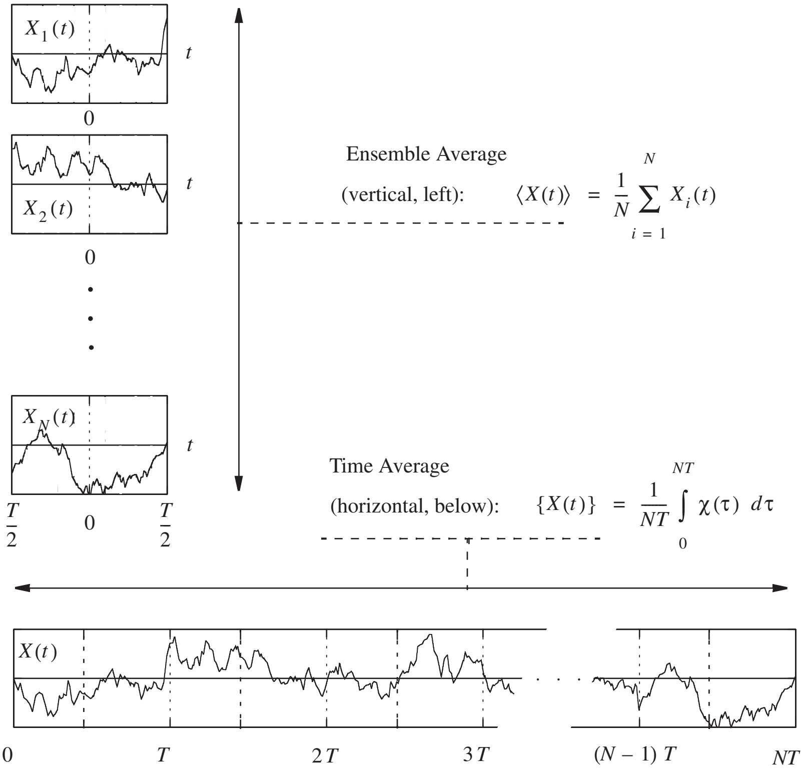

$X_1(t)$
$t$
0
$X_2(t)$
$t$
0
$X_N(t)$
$t$
$\frac{T}{2}$
0
$\frac{T}{2}$
Ensemble Average
(vertical, left): $\langle X(t)\rangle = \frac{1}{N}\sum_{i=1}^{N} X_i(t)$
Time Average
(horizontal, below): $\{X(t)\} = \frac{1}{NT}\int_0^{NT} \chi(\tau)\, d\tau$
$X(t)$
0
$T$
$2T$
$3T$
$(N-1)\,T$
$NT$

**Fig. 6.8** Visualization of the transition from ensemble to time averages

ensemble. This is a technical average, related to a simple measurement. The ergodic assumption is that the ensemble and time averages will give exactly the same result. Another way to say the same thing more exactly is to say that for a function $g[X(t)]$ of the stochastic variable, $X(t)$,

$$\langle g[X(t)]\rangle = E\{g[X(t)]\}, \tag{6.136}$$

where $\langle\cdot\rangle$ denotes a time average and $E\{\cdot\}$ denotes an ensemble average. This means that

$$E\{g[X(t)]\} = \int_{-\infty}^{\infty} g(\chi)f(\chi)d\chi = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} g[x(\tau)]d\tau, \tag{6.137}$$

thus taking a time average is equivalent to taking an ensemble average.

There are many definitions of ergodicity. There is ergodicity in the mean, the mean square and the correlation function (see Soong, 1973, p. 53). In what follows the ergodic assumption will be applied to all of these statistical measures without further comment. In general it is very difficult to determine if a certain system in fact is ergodic experimentally or theoretically: measurements of the necessary quality are very difficult to carry out and even then the statistical interpretation of the experimental results is difficult.

The ergodic assumption obviously has very great technical implications: what is not generally available in practice is a representative collection of sample functions from a very large collection of identically prepared systems. What is available is a certain sample function from one system on which one can make measurements over more or less limited observation periods. In general then ergodic properties must be considered to be a hypothesis. Ergodic conditions are nearly always assumed to obtain because of their great technical utility and because experience has shown that large errors do not seem to be made in this way.

***Example 6.16*. A Constructed Non-ergodic Stationary Process**

In general stationarity does not imply ergodicity. Consider a process whose ensemble is a set of constant functions (for example sine waves with various constant zero offsets). Such a process is stationary since it has a constant distribution function. The average over one member of the ensemble will not be the same as that over one sample function. Thus the stationary process is not at the same time ergodic and ergodic processes are just a subset of stationary processes. ❐

### 6.3.5 Independent Increment Stochastic Processes

A well known example of a process of independent increments is Brownian motion. The observation of Brownian motion was first reported in 1785 by a Dutch physicist, Jan Ingenhausz. The effect was however named for Robert Brown who was a botanist who studied flower pollen suspended in water under a microscope and published his results in 1828. He discovered that the pollen particles which he was studying executed an apparently random motion of a very irregular nature. This he first concluded incorrectly was a primitive form of life but later, correctly, was due to molecular bombardment. A more complete scientific and mathematical approach was taken later first by Einstein and Smoluchowski and then in more detail by Wiener and Lévy (see note 6.6.2 at the end of this chapter for more historical details).

If one imagines a small particle suspended in a liquid bombarded by molecules, from a given starting point, the motion of the particle will be a succession of connected straight line segments. The function described by projecting this motion onto $x$, $y$ and $z$ axes would intuitively be continuous but would have very sharp corners everywhere. Also the particle would wander away from its starting point and would in the mean eventually wander out to an infinite distance from its starting point. It will be useful to keep this picture of a sample function of a random walk process in mind in what follows.

A stochastic process with uncorrelated or independent increments is the basic model which is to be used in this book to model noise in dynamic systems. A process with independent increments, $\mathbf{X}(t)$, $t \geq 0$, can be defined by the following characteristics.

1. Initial value:

$$\mathbf{X}(t_0) = 0. \tag{6.138}$$

2. Given any sequence of time instants, $t_1$, $t_2$, $t_3$, $t_4$, such that $t_0 \leq t_1 \leq t_2 \leq t_3 \leq t_4$, the increments, $\mathbf{X}(t_2) - \mathbf{X}(t_1)$ and $\mathbf{X}(t_4) - \mathbf{X}(t_3)$ have zero means and are independent of each other

$$E\{\mathbf{X}(t_2) - \mathbf{X}(t_1)\} = E\{\mathbf{X}(t_4) - \mathbf{X}(t_3)\} = 0, \tag{6.139}$$

$$E\{[\mathbf{X}(t_2) - \mathbf{X}(t_1)][\mathbf{X}(t_4) - \mathbf{X}(t_3)]\} = 0. \tag{6.140}$$

The mean value of the independent increment process can be computed directly as

$$\mathbf{m}(t) = E\{\mathbf{X}(t)\} = E\{\mathbf{X}(t) - \mathbf{X}(t_0)\} = 0, \ t \geq 0. \tag{6.141}$$

The covariance matrix of the process can be found similarly as

$$\begin{aligned}
\mathbf{R}_X(t_1, t_2) &= E\{\mathbf{X}(t_1)\mathbf{X}^T(t_2)\} && (6.142)\\
&= E\{[\mathbf{X}(t_1) - \mathbf{X}(t_0)][\mathbf{X}(t_2) - \mathbf{X}(t_1) + \mathbf{X}(t_1) - \mathbf{X}(t_0)]^T\} && (6.143)\\
&= E\{[\mathbf{X}(t_1) - \mathbf{X}(t_0)][\mathbf{X}(t_1) - \mathbf{X}(t_0)]^T\} && (6.144)\\
&= E\{\mathbf{X}(t_1)\mathbf{X}^T(t_1)\} && (6.145)\\
&= \mathbf{Q}(t_1), \ t_2 \geq t_1 \geq t_0, && (6.146)
\end{aligned}$$

where $\mathbf{Q}(t) = E\{\mathbf{X}(t)\ \mathbf{X}^T(t)\}$ is the variance of the process. In the same way it can be shown that

$$\mathbf{R}_X(t_1, t_2) = \mathbf{Q}(t_2), \ t_1 \geq t_2 \geq t_0. \tag{6.147}$$

These equations show that an independent increment process cannot be wide sense stationary because the variance of the process is time dependent. It can also be shown that $\mathbf{Q}(t)$ increases monotonically in time, i.e.,

$$\mathbf{Q}(t_2) \geq \mathbf{Q}(t_1) \ \Leftrightarrow \ \mathbf{Q}(t_2) - \mathbf{Q}(t_1) \geq 0, \ \text{for all } t_0 \leq t_1 \leq t_2, \tag{6.148}$$

or in other words, the difference is positive semi-definite.

Assuming that the matrix function $\mathbf{Q}(t)$ is absolutely continuous it is possible to write it as the integral,

$$\mathbf{Q}(t) = \int_0^t \mathbf{V}(\tau) d\tau, \tag{6.149}$$

where $\mathbf{V}(t)$ is a symmetric positive semi-definite matrix function. It follows that

$$E\{[\mathbf{X}(t_2) - \mathbf{X}(t_1)][\mathbf{X}(t_2) - \mathbf{X}(t_1)]^T\} = \mathbf{Q}(t_2) - \mathbf{Q}(t_1) \tag{6.150}$$

$$= \int_{t_1}^{t_2} \mathbf{V}(\tau) d\tau. \tag{6.151}$$

Combining this result with that from Eqs. (6.146) and (6.147), the covariance of the independent increment process can be written as

$$\mathbf{R}_X(t_1, t_2) = \int_{t_0}^{min(t_1, t_2)} \mathbf{V}(\tau) d\tau. \tag{6.152}$$

As mentioned earlier, a random walk or Brownian motion is one of the best examples of a process with independent increments. Such a process is also called a Wiener or Wiener-Lévy process. It is in fact the process in the integrand in Eq. (6.152).

***Example 6.17*. Wiener Process**

Let $\{X(t), t \geq 0\}$ be a scalar stochastic process such that:

1. $X(0) = 0$,
2. the process is an independent increment process,
3. for $t_2 \geq t_1$, the increments have a Gaussian distribution:

$$Pr([X(t_2) - X(t_1)] \leq x) = \frac{1}{\sqrt{2\pi(t_2 - t_1)}} \int_{-\infty}^{x} \exp\left(-\frac{\chi^2}{2(t_2 - t_1)}\right) d\chi, \tag{6.153}$$

then this is a normalized Wiener (or Wiener-Lévy) process. If $W(t) = \sigma X(t)$, where $X(t)$ is a normalized Wiener process then

$$E\{W(t)\} = E\{\sigma X(t)\} = \sigma E\{X(t)\} = 0, \tag{6.154}$$

because $X(t)$ is a zero mean process. Also the increments of the process have zero means,

$$E\{X(t_2) - X(t_1)\} = 0, \tag{6.155}$$

and the variance of the incremental process is

$$E\{[W(t_2) - W(t_1)]^2\} = \sigma^2 E\{[X(t_2) - X(t_1)]\} = \sigma^2|t_2 - t_1|, \tag{6.156}$$

for any $t_1$ and $t_2$, because the increments have a Gaussian distribution function (Eq. (6.153)). This means that any Wiener process can be formed from a nomalized Wiener process and this is often convenient in calculations. From Eq. (6.156), if say a pollen particle in a liquid suspension starts at $t_1 = 0$, then the uncertainty in its position will increase with $\sqrt{\text{time}}$:

$$Q(t) = \mu_2(t) = \sigma^2 t. \tag{6.157}$$

This is a general characteristic of Wiener processes. A pictorial representation of the process is shown on Fig. 6.9. Starting from a very localized variance at zero time, the process spreads out according to Eq. (6.157).

A Wiener process is the simplest possible example of a diffusion process. It can be shown that almost all sample functions of a Brownian motion process are nondifferentiable. Intuitively they are continuous but have corners nearly everywhere. These sample functions are also of unbounded variation: this property precludes the use of ordinary Riemann integrals. Thus it is necessary to develop new mathematical tools in order to do rigorous calculations with random processes as inputs to dynamic systems. ❐

A Wiener process cannot itself be a good choice for a general noise model. It is not stationary and can become infinite for large times. It can however be used to obtain a model which is generally applicable for noise modelling. An obvious method of attack is to differentiate the Wiener process formally in order to obtain a process which has no D.C. or zero frequency component.

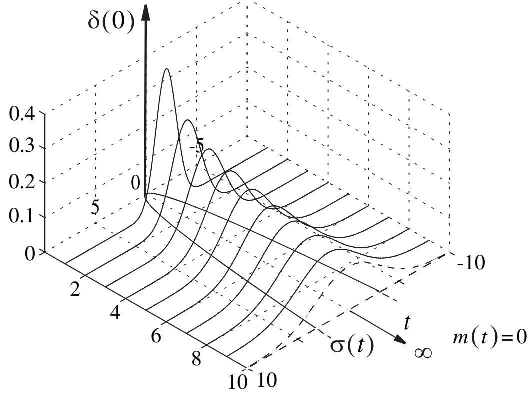


**Fig. 6.9** The spreading out of the variance of a normalized Wiener process in time shown up to $t = 10$ only. At $t = 0$, the variance is $\delta(0)$

***Example 6.18.*** **White Noise: Derivative of an Independent Increment Process**

Proceeding purely formally from the above, it is possible to show that white noise is the derivative of an independent increment process. Consider now the derivative of the vector independent increment process, $\{\mathbf{X}(t),\ t \geq t_0\}$:

$$\dot{\mathbf{X}}(t) = \frac{d\mathbf{X}(t)}{dt}. \tag{6.158}$$

The mean of the derivative process can be found from

$$E\{\dot{\mathbf{X}}(t)\} = \frac{d}{dt}E\{\mathbf{X}(t)\} = 0. \tag{6.159}$$

The covariance matrix of the derivative process can be obtained by formally differentiating the integral in Eq. (6.152) and using Leibnitz' differentiation rule for integrals (see starred note *, p. 387). Proceeding in this way,

$$\mathbf{R}_{\dot{X}}(t_1, t_2) = E\{\dot{\mathbf{X}}(t_1)\dot{\mathbf{X}}^T(t_2)\} \tag{6.160}$$

$$= \frac{\partial^2}{\partial t_1 \partial t_2}E\{\mathbf{X}(t_1)\mathbf{X}^T(t_2)\} \tag{6.161}$$

$$= \frac{\partial^2}{\partial t_1 \partial t_2}\mathbf{R}_X(t_1, t_2),\ \ t_1, t_2 \geq t_0. \tag{6.162}$$

Now differentiating equation (6.152) using Leibnitz's rule yields

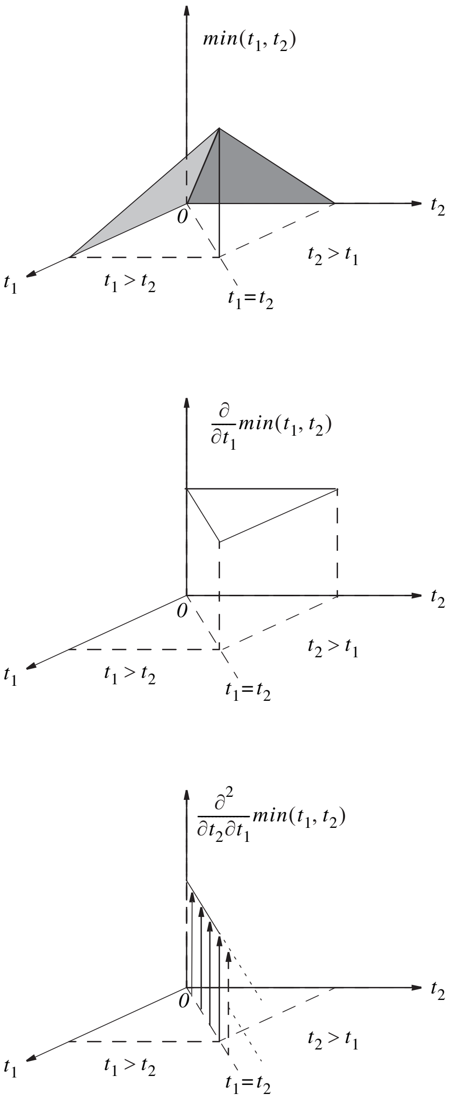
$$\mathbf{R}_{\dot{X}}(t_1, t_2) = \mathbf{V}(min(t_1, t_2))\frac{\partial^2}{\partial t_1 \partial t_2}min(t_1, t_2) \tag{6.163}$$

$$= \mathbf{V}(t_1)\delta(t_1 - t_2),\ \ t_1, t_2 \geq t_0, \tag{6.164}$$

where $\delta(\ldots)$ is the Dirac delta function. This is the covariance function of a white noise process. $\mathbf{V}(\ldots)$ is the intensity or strength of the white noise signal. See Fig. 6.10 for a graphical 'derivation' of Eq. (6.164). Equation (6.164) implies that a white noise process is correlated only when $t_1 = t_2$.

White noise is thus an extremely irregular process. From the above it is generated formally by differentiating a Wiener process which is also very irregular. For practical simulation purposes, white noise is generated by using a sampled random number generator with a constant sample time. This sample time has to be much smaller than the fastest time constant in the system being studied. In principle the sample time should be varied in a random way in the same way as the amplitude of the signal. The random number generator used most often has a Gaussian distribution function with a variance of 1. A Gaussian distribution function is used because it is simple and often encountered in physical systems.

**Fig. 6.10** Graphical 'derivation' of a white noise signal. See example 6.18



*Leibnitz' rule: Suppose that

$$I(u) = \int_{a(u)}^{b(u)} f(x, u)\ dx \ \Rightarrow\ I = I(b, a, u), \tag{6.165}$$

then

$$\frac{dI}{du} = \frac{\partial I}{\partial u}\frac{du}{du} + \frac{\partial I}{\partial a}\frac{da}{du} + \frac{\partial I}{\partial b}\frac{db}{du} \tag{6.166}$$

$$= \int_{a(u)}^{b(u)} f_u(x,u)\ dx - f(a,u)\frac{da}{du} + f(b,u)\frac{db}{du}. \tag{6.167}$$

❐

The example above shows the form which will generally be used for white noise in this book from this point. White noise is the derivative of a Wiener process with a mean value zero, a Gaussian amplitude distribution and which has a covariance or correlation matrix which is

$$\mathbf{R}_X(\tau) = E\{\mathbf{X}(t)\ \mathbf{X}^T(t+\tau)\} = \mathbf{V}\,\delta(\tau) \tag{6.168}$$

where $\mathbf{V}$ is a symmetric, square, constant matrix. The diagonal elements of the matrix are the variances of the components of $\mathbf{X}(t)$ (say like $\sigma_{X_i}^2$, $i = 1, 2, \ldots, n$) while the off diagonal elements express the correlation between the components of $\mathbf{X}(t)$ (say like $\sigma_{X_i}\sigma_{X_j}$, where $i,j = 1, 2, \ldots, n$). White noise is in reality a mathematical abstraction because physically speaking, it does not (and cannot) exist. It turns out however that its use leads to relatively simple expressions for state and output noise in linear time-invariant dynamic systems. This is the reason why it is so widely used in disturbance modelling.

Recall that the power dissipated in a resistor is proportional to $V_{RMS}^2$, the squared effective (or squared RMS) voltage across the resistor, or $I_{RMS}^2$, the squared effective (or squared RMS) current through the resistor, with the proportionality constants $1/R$ and $R$ respectively. The corresponding quantity for a random variable is the squared standard deviation (or variance) of the stochastic variable, $\sigma^2$. The diagonal matrix elements of the covariance matrix are thus the power (apart from a constant) in the components of the stochastic variable $\mathbf{X}(t)$. The nondiagonal elements are the power in the cross correlation components. For independent variables these off diagonal elements are zero. Thus the description of stochastic variables is in terms of their power content not the details of their specific realization (or actual time dependence).

As is clear from the example above, white noise is closely related to a process of independent increments. White noise obtains its name from the fact that its spectrum in the frequency domain is perfectly flat for all frequencies. White light contains all colors or equivalently all frequencies in the visible light range. Hence the name: white noise. More details about white noise and its use will appear later in this chapter: it is in fact the main noise and disturbance model which will be used in this book.

The approach to deriving the white noise process above is seemingly very abstract. It is useful to try to develop a more intuitive physical model or paradigm for this process as it is in such universal use.

### *Example 6.19.* An Undamped Random Walk: White Noise

In order to obtain a more intuitive physical picture of a Wiener process it is useful to consider a simple physical picture of a random walk. Consider a grain of pollen suspended in a liquid. Such a particle is held in suspension by the random motion of the molecules of the fluid around it. At random times the pollen grain is hit by molecules of the liquid in what can be considered to be instantaneous hard sphere collisions. The equation of motion of the pollen particle in one dimension for a single collision event is then

$$ma = m\dot{v} = F(t)\ \delta(t - t_0), \tag{6.169}$$

where $m$ is the mass of the particle, $a$ is the acceleration, $v$ is the velocity, $t_o$ is(are) some arbitrary collision(s) time(s). $F(t)$ is the strength of the forces exerted on the pollen grain in the collisions. Equation (6.169) can be rewritten as

$$\frac{dv}{dt} = \frac{F(t)}{m}\ \delta(t - t_0). \tag{6.170}$$

The quantity $F(t)/m$ determines the acceleration of the particle. It can be seen that the motion of the pollen grain is governed by a succession of delta function accelerations with a random strengths at random times: see Fig. 6.11.

Because the pollen particle is effectively free between collisions, its velocity will be constant between collisions. To find its path, the velocity components must be integrated in time. The motion in each dimension can then be represented by two integrators in series. The input to first integrator is the delta function forces, the resulting velocities are the output of the first integrator (a set of random step functions) and the output of the second integrator is the path length (a set of random ramps). The velocity is a white noise process and the path length (a random walk) is a Wiener process.

It should also be noted that there is no damping of the velocity of the particles built into the physical model. This means that the particle velocities can in principle become infinite. This implies that a collection of particles can contain infinite energy and this is of course not possible. Thus the model above is a mathematical idealization which is not physically reasonable.

It is easy to simulate the overall system above, using a uniform distribution to define the different collision times and a Gaussian amplitude distribution for the delta function collision force intensities. These collision force intensities are used to drive the first integrator which in turn drives a second integrator. The output of the first integrator is white noise. The output of the second integrator is a random walk or Wiener process. The results of such a simulation are shown on Figs. 6.11, 6.12 and 6.13.

Figure 6.13 shows a random walk for a particle in two dimensions. It is composed of two independent Wiener processes, $X_1(t)$ and $X_2(t)$. Each of these processes is generated separately as shown in Figs. 6.11 and 6.12 using uniformly distributed (in time) delta function forces, which have a Gaussian amplitude distribution. It is the vertical and horizontal velocities of the particle which are white noise processes, $V_1(t)$ and $V_2(t)$. ❐

The large frequency content of a white noise process is one of the characteristics which makes it non-physical; it is the quality which gives it an infinite energy

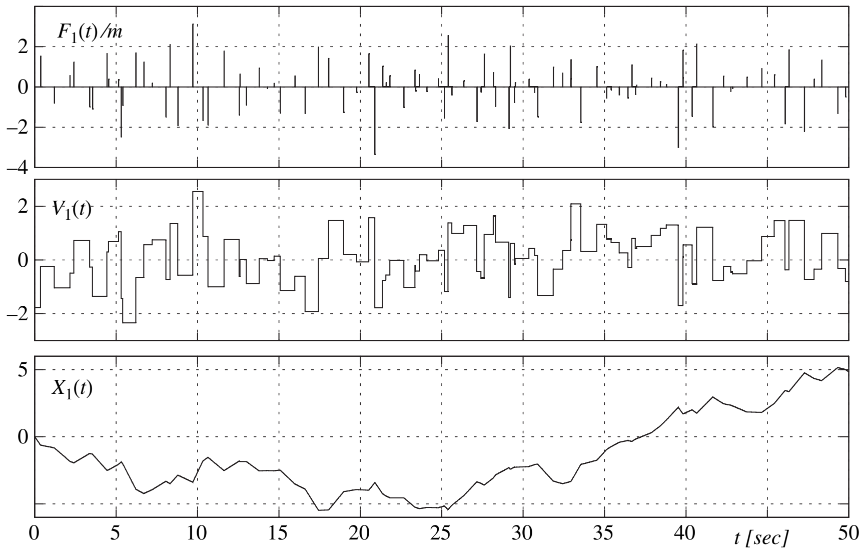


**Fig. 6.11** Construction of an independent Wiener process: $x_1$ dimension

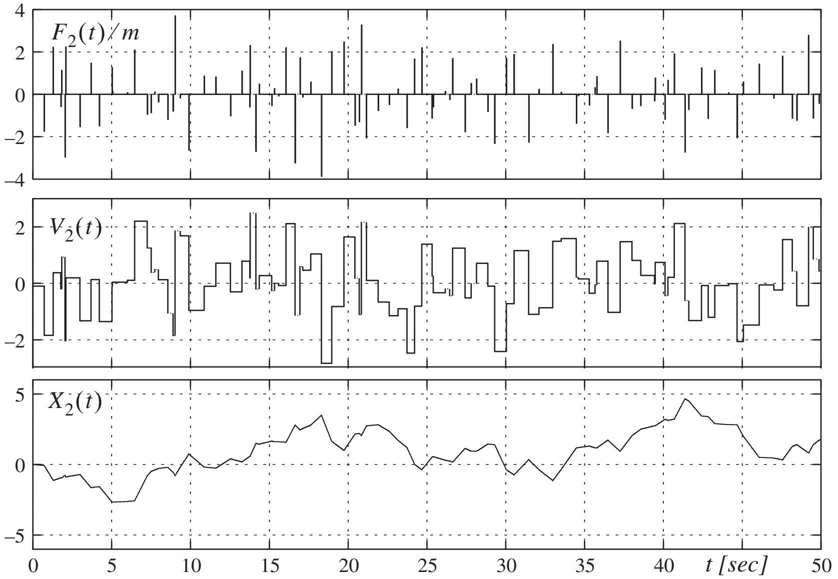


**Fig. 6.12** Construction of an independent Wiener process: $x_2$ dimension

content. It can also be shown that the Wiener process is only an accurate model for a random walk for times which are large compared with the mean time between collisions.

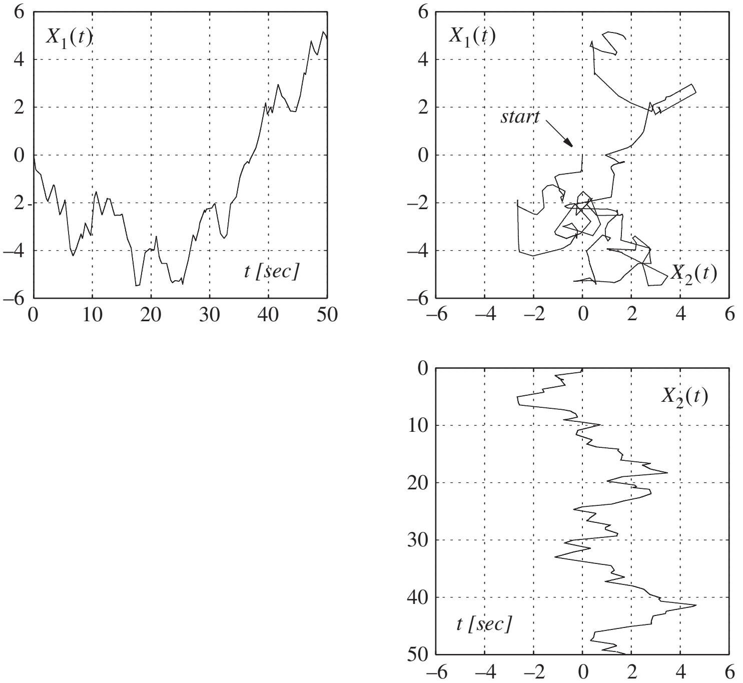


**Fig. 6.13** Construction of a two dimension random walk using the independent one dimensional random walks of Figs. 6.11 and 6.12

It is possible to construct a more physical model of a random walk which includes the necessary requirement that it have a finite bandwidth. This is detailed in Example 6.20 below.

### *Example 6.20.* A Damped Random Walk: Bandlimited White Noise

The physical model in Example 6.19 can easily be extended to include the viscous damping force on the pollen particles in addition to the random collision forces. The differential equation which is the result of this extension is classical and is known as the Langevin equation. It is

$$\frac{dv}{dt} = -\frac{b}{m}v + \frac{F(t)}{m}\delta(t - t_0), \tag{6.171}$$

where $b$ is a damping coefficient and $F(t)$ is the same Gaussian distributed intensity as in Example 6.19 above. It implies the existence of processes in which the kinetic energy associated with the pollen grain is dissipated to other degrees of freedom such as the molecules in the suspension liquid, etc.

The result of solving the Langevin equation is a process which has much less high frequency content than an undamped white noise/Wiener process. This process, $v(t)$, is an Ornstein-Uhlenbeck process. It has a Gaussian distribution but unlike the Wiener process, it does not have independent increments. It can be shown however that the overall system, its displacement as well as its velocity, is an accurate description of the random walk of a pollen particle. This is true both for long times compared to a collision time but also for short times. A typical realization of the process would look like an undamped random walk but it would be smoother, indicating a lower high frequency content. Such a system also has a finite energy content, making it more physical than the ideal process.

In fact it is possible to give a more detailed model for the damping parameter, $b$, by using Stokes's law for a viscous liquid. It can be expressed as

$$b = 6\pi\eta a, \tag{6.172}$$

where $\eta$ is the viscosity of the liquid and $a$ is the particle radius. It can be shown that introducing this damping coefficient results is a low pass filtering of the white noise which would otherwise be the result of solving Eq. (6.170) above (see Example 6.19). The time constant of this low pass filter can be shown to be

$$\tau_r = \frac{m}{b} = \frac{m}{6\pi\eta a}. \tag{6.173}$$

For a typical pollen particle this time constant is on the order of 1 microsecond which implies a high frequency cutoff at about $6.28 \cdot 10^6$ Hz. This is of course quite high and means that the bandlimited white noise process is close to being undamped white noise for most practical purposes.

Another example of physically generated, low pass filtered, white noise is the thermal noise generated in an electrical conductor by the thermal agitation of the electrons. Such noise has a flat spectrum up to $10^{12} = 1000$ GHz at room temperature (20°C). Above this the spectrum falls off with frequency. ❐

## 6.4 Noise Propagation: Frequency and Time Domains

The response of linear systems to random signals can be built up around the theory of deterministic signals in such networks. It is also a requirement that the statistical measures which have been considered above be interpreted in the light of this linear theory. The response of a linear network which has an impulse response, $\mathbf{g}(t)$, to an input signal $\mathbf{u}(t)$ is given by the convolution integral:

$$\mathbf{y}(t) = \int_0^\infty \mathbf{g}(\tau)\,\mathbf{u}(t-\tau)\,d\tau = \int_{-\infty}^t \mathbf{g}(t-\tau)\,\mathbf{u}(\tau)\,d\tau. \tag{6.174}$$

This equation is valid both for scalar and vector systems. A simple example of a scalar system is a low pass filter.

***Example 6.21.*** **Impulse Response of a Low Pass Filter**

The response of a first order low pass filter with a D.C. gain of 1 is defined by the differential equation

$$\frac{dy(t)}{dt} = \frac{1}{T}(-y(t) + u(t)), \tag{6.175}$$

where $T$ is the time constant. The impulse response of the filter is an exponential function

$$g(t) = \begin{cases} 0, & t < 0 \\ \frac{1}{T}e^{-\frac{t}{T}}, & t \geq 0 \end{cases}. \tag{6.176}$$

For an arbitrary input $u(t)$ the response of the filter is given by the integral

$$y(t) = \int_0^\infty \frac{1}{T}e^{-\frac{\tau}{T}}u(t-\tau)d\tau. \tag{6.177}$$

❒

In the frequency domain the relation of the input to a system to its output is given by the transfer function from input to output,

$$\mathbf{Y}(j\omega) = \mathbf{G}(j\omega)\mathbf{U}(j\omega), \tag{6.178}$$

which is in general a matrix equation as shown. This is merely the result of transforming the convolution integral of Eq. (6.174) to the frequency domain. $\mathbf{G}(j\omega)$ is the Fourier transform of the impulse response, $\mathbf{g}(t)$:

$$\mathbf{G}(j\omega) = \int_{-\infty}^{\infty} \mathbf{g}(\tau)e^{-j\omega\tau}d\tau, \tag{6.179}$$

where $\omega = 2\pi f$, and $f$ is the frequency. This is of course only one half of a Fourier transform pair. The other half is

$$\mathbf{g}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{G}(j\omega)e^{j\omega\tau}d\omega. \tag{6.180}$$

The equations above are valid both for scalar and vector systems.

The Fourier transform is a general way of associating a frequency response with a time response. A time response can be seen as a superposition of sine waves having certain amplitude and phase relationships. Any time response can be reproduced from sine waves (possibly infinitely many) in this way. A delta function contains sine waves of equal amplitudes at all frequencies. The impulse response of a filter is the result of propagating the continuous flat spectrum of a delta function pulse through the dynamics of the filter. A Fourier transform of the impulse response of the filter is thus in reality it's transfer function. This relationship is often used in signal analysis. For more details the reader can consult (see Papoulis, 1977).

***Example 6.22.*** **Frequency Response of a Low Pass Filter**

For the differential equation of Eq. (6.175) the frequency response corresponding to the impulse response is simply

$$G(j\omega) = \frac{1}{1 + j\omega T} = \frac{Y(j\omega)}{U(j\omega)}. \tag{6.181}$$

This low pass filter obviously has a gain of 1 at D.C. ($\omega = 0$). ❐

It is possible to apply these same frequency domain concepts to the treatment of random processes as will be clear in the next sections.

### 6.4.1 Continuous Random Processes: Time Domain

At this point it is clear that in order to deal with random variables in dynamic systems it will be necessary to consider averages of the variables or some function of the variables. This is necessary because it is very difficult to deal with the sample functions of a random process: it is not in fact possible to write down an equation for a single realization of the process.

Consider the problem of finding the mean (or expectation value) of Eq. (6.174) when $\mathbf{u}(t)$ is a stochastic process,

$$\mathbf{m}_y(t) = E\{\mathbf{y}(\mathrm{t})\} = \mathrm{E}\left\{\int_0^{\infty} \mathbf{g}(\tau)\mathbf{u}(t-\tau)\mathrm{d}\tau\right\} = \mathrm{E}\left\{\int_{-\infty}^{\mathrm{t}} \mathbf{g}(t-\tau)\mathbf{u}(\tau)\mathrm{d}\tau\right\}. \tag{6.182}$$

It can be shown that this expression has good meaning if the dynamic system is asymptotically stable and $\mathbf{y}(t)$ is a stochastic process which has finite second order moments (see Åström (1970)). This then means that one can write

$$\mathbf{m}_y(t) = E\left\{\int_0^{\infty} \mathbf{g}(\tau)\mathbf{u}(t-\tau)d\tau\right\} = \int_0^{\infty} \mathbf{g}(\tau)E\{\mathbf{u}(t-\tau)\}d\tau \tag{6.183}$$

$$= \int_{\infty}^{0} \mathbf{g}(\tau)\mathbf{m}_u(t-\tau)d\tau, \tag{6.184}$$

where the operations of integration and expectation have been exchanged. Likewise for the covariance of the process one has

$$\mathbf{R}_{uy}(t_1, t_2) = E\{\mathbf{u}(t_1)\mathbf{y}^T(t_2)\} \tag{6.185}$$

$$= \int_0^\infty \mathbf{g}(\tau)\ \mathbf{R}_u(t_1, t_2 - \tau)\ d\tau, \tag{6.186}$$

and

$$\mathbf{R}_y(t_1, t_2) = E\{\mathbf{y}(t_1)\ \mathbf{y}^T(t_2)\} \tag{6.187}$$

$$= \int_0^\infty \int_0^\infty \mathbf{g}(\varsigma)\mathbf{R}_u(t_1 - \varsigma, t_2 - \tau)\mathbf{g}^T(\tau)\ d\varsigma\ d\tau. \tag{6.188}$$

In particular for a random process which is strongly stationary such that

$$\mathbf{m}_u(t) = \mathbf{m}_u = \text{constant}, \tag{6.189}$$

$$\mathbf{R}_u(t_1, t_2) = \mathbf{R}_u(t_1 - t_2), \tag{6.190}$$

one has

$$\mathbf{m}_y = \int_0^\infty \mathbf{g}(\tau)\ d\tau \cdot \mathbf{m}_u, \tag{6.191}$$

$$\mathbf{R}_{uy}(t_1 - t_2) = \int_0^\infty \mathbf{g}(\tau)\ \mathbf{R}_u(t_1 - t_2 + \tau)\ d\tau, \tag{6.192}$$

$$\mathbf{R}_y(t_1 - t_2) = \int_0^\infty \int_0^\infty \mathbf{g}(\varsigma)\mathbf{R}_u(t_1 - t_2 - \varsigma + \tau)\ \mathbf{g}^T\ (\tau)\ d\varsigma\ d\tau. \tag{6.193}$$

In the two last equations it should be noticed that both equations are only functions of the time difference $t_1 - t_2$. Thus it is common to write them as functions of a single variable: for example $\tau = t_1 - t_2$.

***Example 6.23*. Finite Time Averaging**

From a technical view point, in order to make useful averages it is convenient to use the ergodic assumption and assume that the ensemble averages which have been important up to this point will be equivalent to the time averages which can practically be measured. Thus it is necessary here to consider the nature of time averages and how such averages propagate through linear dynamic systems. In order to obtain reproducible results it is important to consider only wide sense stationary stochastic processes here.

A time average which one might imagine has the same accuracy as the ensemble averages considered earlier might be

$$y(t) = \lim_{T\to\infty} \frac{1}{T} \int_{t-T}^{t} x(\tau)\ d\tau, \tag{6.194}$$

for a certain realization of the continuous stochastic process, $X(t)$, $t \geq 0$. This is nothing more or less than a linear filter with an impulse response

$$g(t) = \begin{cases} \frac{1}{T}, & 0 \leq \mathrm{t} \leq \infty \\ 0, & \mathrm{t} < 0 \end{cases}. \tag{6.195}$$

While this should yield good results, it is inconvenient to have to carry out the average over an infinite time period. Thus a practically applicable time average would be

$$y(t) = \frac{1}{T}\int_{t-T}^{t} x(\tau)\ d\tau. \tag{6.196}$$

Such a finite interval time averager can be realized in a number of different ways, for example using an operational amplifier integrator with a time constant of T = RC.

To consider the finite interval time averager from an ensemble point of view it is useful to consider a number of different experiments with it, all collecting sample functions over the time interval $T$. This makes is possible to consider the output $Y(t)$ as a stochastic process generated by another stochastic process $X(t)$. Assuming that the input signal is an ergodic process, the finite interval time average can be written

$$Y(t) = \frac{1}{T}\int_{t-T}^{t} X(\tau)\ d\tau. \tag{6.197}$$

In this way the first moment (average) and covariance functions of $Y(t)$ can be found as

$$m_Y(t) = E\{Y(t)\} = E\left\{\frac{1}{T}\int_{t-T}^{t} X(\tau)\ d\tau\right\} \tag{6.198}$$

$$= \frac{1}{T}\int_{t-T}^{t} E\{X(\tau)\}\ d\tau = \frac{1}{T}\int_{t-T}^{t} m_X(\tau)\ d\tau, \tag{6.199}$$

and

$$R_Y(t_1, t_2) = E\{[Y(t_1) - m_Y(t_1)][Y(t_2) - m_Y(t_2)]^T\} \tag{6.200}$$

$$= E\left\{\frac{1}{T^2}\int_{t_2-T}^{t_2}\int_{t_1-T}^{t_1} [X(\varsigma) - m_X(\varsigma)][X(\tau) - m_X(\tau)]^T\ d\varsigma\ d\tau\right\} \tag{6.201}$$

$$= \frac{1}{T^2}\int_{t_2-T}^{t_2}\int_{t_1-T}^{t_1} R_X(\varsigma, \tau)\ d\varsigma\ d\tau, \tag{6.202}$$

where it has been assumed that it is possible to exchange the operations of expectation and integration. This corresponds to changing the order of the two integrations. It can be shown that this is possible under fairly general conditions. It is not difficult to show that for a wide sense stationary stochastic process that the equations above lead to the results $m_Y = m_X$ and $R_Y(\tau) = R_X(\tau)$.

The results above can be immediately generalized for other impulse responses than those corresponding to finite horizon averaging and for vector stochastic processes.

### 6.4.2 Continuous Random Processes: Frequency Domain

To treat the propagation of a stochastic signal through a linear network it is necessary to obtain a picture of how the covariance matrix looks in the frequency domain. For a wide sense stationary stochastic process this can be done by obtaining the Fourier transform of the covariance matrix. This quantity is called the power spectral density for a scalar variable and the power spectral density matrix for a vector process. Making the variable substitution $\tau = t_1 - t_2$ (which is valid for a wide sense stationary stochastic process) this matrix can be expressed as

$$\mathbf{S}_X(\omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau} \mathbf{R}_X(\tau)\, d\tau. \tag{6.203}$$

From this equation, for symmetry reasons, it is clear that it must also be true that

$$\mathbf{R}_X(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{S}_X(\omega) e^{j\omega\tau}\, d\omega. \tag{6.204}$$

Equations (6.203) and (6.204) are called the Wiener-Khintchine relations (or theorem). What the first matrix equation above expresses is the power density of the stochastic variable $\mathbf{X}(t)$ as a function of frequency. It has the following important and useful properties:

1. $\mathbf{S}_X(-\omega) = \mathbf{S}_X^T(\omega)$, for all ω,
2. $\mathbf{S}_X{}^*(\omega) = \mathbf{S}_X(\omega)$, for all ω,
3. $\mathbf{S}_X(\omega) \geq 0$, for all ω,

where the * denotes the complex conjugate transpose. The inequality in 3 indicates that the spectral power density matrix is positive semi-definite. From the definition of the spectral power density, it should be clear that it expresses the power in a variable per unit frequency.

One of the most important uses of the spectral power density function is exactly to describe the propagation of an input signal to the output of a system.

Consider an asymptotically stable time invariant linear system with a transfer function given by the matrix $\mathbf{G}(j\omega)$. If the input to this system is a wide sense stationary stochastic process, $\{\mathbf{U}(t),\ t \geq 0\}$, and the transfer function of the system is given by

$$\mathbf{G}(s) = \int_0^\infty e^{-s\tau}\mathbf{g}(\tau)\ d\tau, \tag{6.205}$$

then using the fact that $\mathbf{m}_y(t)$ is given by the convolution integral above, Eq. (6.184), the mean value at the output of the system is

$$\mathbf{m}_y = \mathbf{G}(s)|_{s=0}\mathbf{m}_u = \mathbf{G}(0)\ \mathbf{m}_u. \tag{6.206}$$

This expression says that the mean value of the input, when multiplied by the D.C. gain of the network is the mean of the output. With a power spectral density matrix $\mathbf{S}_u(\omega)$ which is constant in time, the output of the system is also a wide sense stationary stochastic process, $\mathbf{Y}(t)$, where

$$\mathbf{S}_{xy}(\omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau}\mathbf{R}_{uy}(\tau)\ d\tau \tag{6.207}$$

$$= \int_{-\infty}^{\infty} e^{-j\omega\tau} \int_0^\infty \mathbf{g}(s)\mathbf{R}_u(\tau - s)\ ds\ d\tau \tag{6.208}$$

$$= \int_{-\infty}^{\infty} \int_0^\infty e^{j\omega s}\mathbf{g}(s)e^{-j\omega(\tau+s)}\mathbf{R}_u(\tau + s)\ ds\ d\tau, \tag{6.209}$$

or

$$\mathbf{S}_{uy}(\omega) = \mathbf{G}(-j\omega)\ \mathbf{S}_u(\omega). \tag{6.210}$$

The power spectral density matrix, $\mathbf{S}_y(\omega)$, is given by

$$\mathbf{S}_y(\omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau}\mathbf{R}_y(\tau)\ d\tau \tag{6.211}$$

$$= \int_{-\infty}^{\infty} e^{-j\omega\tau} \int_0^\infty \int_0^\infty \mathbf{g}(s)\mathbf{R}_u(\tau - s + t)\mathbf{g}^T(t)\ ds\ dt\ d\tau \tag{6.212}$$

$$= \int_{-\infty}^{\infty} \int_0^\infty \int_0^\infty e^{-\mathrm{j}\omega s}\mathbf{g}(s)e^{-\mathrm{j}\omega(\tau-s+t)}\mathbf{R}_u(\tau - s + t)\mathbf{g}^T(t)e^{j\omega t}\ ds\ dt\ d\tau \tag{6.213}$$

or

$$\mathbf{S}_y(\omega) = \mathbf{G}(j\omega)\mathbf{S}_u(\omega)\mathbf{G}^T(-j\omega). \tag{6.214}$$

For a wide sense stationary stochastic variable, $\mathbf{V}(t)$, with zero mean and with covariance matrix $\mathbf{R}_V(t_1 - t_2)$, a quantity proportional to the total A.C. power in the components of the vector is given by

$$tr[\mathbf{R}_V(0)] = E\{[\mathbf{V}(t_1) - \mathbf{m}(t_1)]^T[\mathbf{V}(t_2) - \mathbf{m}(t_2)]\} \tag{6.215}$$

$$= tr\left[\frac{1}{2\pi}\int_{\infty}^{-\infty} \mathbf{S}_V(\omega)d\omega\right] = tr\left[\int_{\infty}^{-\infty} \mathbf{S}_V(f)df\right], \tag{6.216}$$

where $tr[\ldots]$ means the sum of diagonal elements of the matrix, $[\ldots]$, and $f$ is the frequency ($2\pi f = \omega$) (see Example 6.24 below). This corresponds to the total A.C. power in it given that the mean value is zero. The matrix elements of

$$\mathbf{R}_V(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{S}_V(\omega)\ d\omega \tag{6.217}$$

are proportional to the A.C. power in its components.

For a variable which is not zero mean, the matrix elements of $\mathbf{C}_V(0)$ are proportional to the A.C. plus D.C. powers in components of the signal, where

$$C_V(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{S}_V(\omega)d\omega = \int_{-\infty}^{\infty} \mathbf{S}_V(f)df. \tag{6.218}$$

$tr[\mathbf{C}_V(0)]$ is proportional to the total power in the signal $\mathbf{V}(t)$. $\mathbf{R}_V(0)$ is nevertheless still proportional to the A.C. powers in the signal components.

### *Example 6.24.* Stochastic Quadratic Forms

The result of Eq. (6.216) can be obtained in a simple way by direct calculation. Consider the problem of evaluating a quadratic expression of the form,

$$E\{\mathbf{v}^T(t)\mathbf{W}(t)\mathbf{v}(t)\} = E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} v_i(t)W_{ij}(t)v_j(t)\right\}, \tag{6.219}$$

where $\mathbf{v}(t)$ is a stochastic vector (with $n$ elements) and $\mathbf{W}(t)$ is a symmetric weighting matrix. This is clearly a quadratic expression in the elements in the vector $\mathbf{v}(t)$. $\mathbf{W}(t)$ is invariably selected to be a positive semidefinite as what is desired is an expression for the overall power in the system.

The equation above can be evaluated using elementary methods:

$$E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} v_i(t)W_{ij}(t)v_j(t)\right\} = E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(t)v_i(t)v_j(t)\right\} \tag{6.220}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(t) E\{v_i(t) v_j(t)\} \tag{6.221}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(t) C_{v,ij}(t, t) \tag{6.222}$$

$$= tr[W(t)\mathbf{C}_v(t, t)]. \tag{6.223}$$

Thus

$$E\{\mathbf{v}^T(t)\mathbf{W}(t)\mathbf{v}(t)\} = tr[\mathbf{W}(t)\mathbf{C}_v(t, t)]. \tag{6.224}$$

If $\mathbf{v}(t)$ is a wide sense stationary stochastic process with zero mean such that Eq. (6.128) is satisfied and $\mathbf{W}(t)$ is constant, then

$$E\{\mathbf{v}^T(t)\mathbf{W}(t)\mathbf{v}(t)\} = tr[\mathbf{W}(t)\mathbf{R}_v(0)]. \tag{6.225}$$

Moreover if $\mathbf{v}(t)$ is zero mean and a power spectral density matrix which is $\mathbf{S}_v(\omega)$,

$$E\{\mathbf{v}^T(t)\mathbf{W}\mathbf{v}(t)\} = tr\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{W}(t)\mathbf{S}_v(\omega)d\omega\right]. \tag{6.226}$$

These results have already been used without detailed derivation in Eq. (6.217). It is important to remember this technique as it makes it possible to calculate the mean square value of any set of stochastic variables. ❐

***Example 6.25*. White Noise**

As mentioned above, white noise has a constant amplitude spectrum. Its spectral power density matrix reflects this characteristic but is an expression of the 'power' in the noise. Thus for a wide sense stationary white noise process having a constant intensity, $V = \sigma^2$, expressed as

$$S_X(\omega) = V = \sigma^2, \tag{6.227}$$

(a constant). One has then that

$$R_X(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Ve^{j\omega\tau}d\omega = V\delta(\tau), \tag{6.228}$$

as $V$ is constant and the integral expression is a delta function. Thus

$$R_X(\tau) = V\delta(\tau) = \sigma^2\delta(\tau). \tag{6.229}$$

From the expression for the spectrum, it is clear that the white noise signal must contain infinite energy and thus is physically unrealistic and can only be an approximation. Nevertheless white noise is a good model for random disturbances in dynamic systems because in general practical dynamic systems contain their own high frequency filtering. See Fig. 6.14, below. ❐

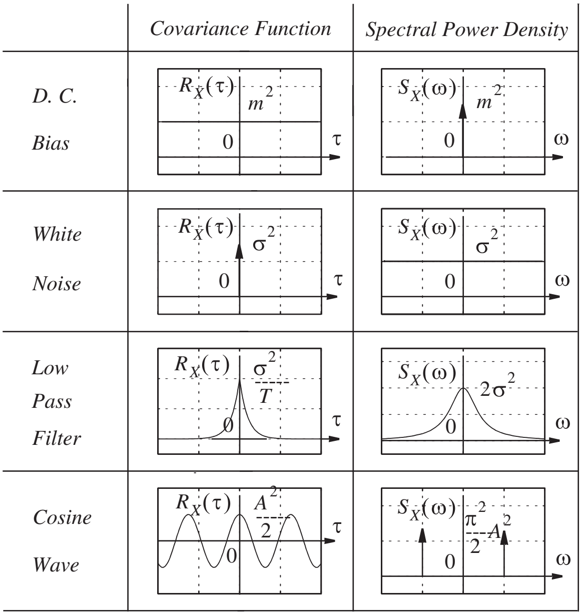


**Fig. 6.14** Covariance functions and power spectral densities of common deterministic and random processes

### *Example 6.26.* Exponentially Correlated Noise

In Examples 6.21 and 6.22 above a unity gain (at $\omega = 0$) low pass filter is treated. White noise filtered with this filter has the covariance function

$$R_X(\tau) = R_X(t_1 - t_2) = \frac{\sigma^2}{T} e^{-\frac{|t_1 - t_2|}{T}} = \frac{\sigma^2}{T} e^{-\frac{|\tau|}{T}}. \tag{6.230}$$

It is easy to Fourier transform this quantity to find the corresponding spectral density function,

$$S_X(\omega) = \frac{2\sigma^2}{1 + \omega^2 T^2}, \tag{6.231}$$

for $T > 0$, remembering the result of Eq. (6.175). Low pass filtered white noise is a very common model for many types of disturbances or noise in dynamic systems. See Fig. 6.14, p. 401. ❐

***Example 6.27*. A Sine Wave and its Autocorrelation Function**

To give an idea of the meaning of Eqs. (6.203) and (6.204) it is a instructive to look at a deterministic signal like a sine wave. It is not difficult to show that the autocorrelation function of a cosine wave, $x(t) = A\cos(\omega_0 t)$, is given by the expression

$$R_X(\tau) = A^2\cos(\omega_0\tau). \tag{6.232}$$

When transformed using Eq. (6.203) this yields the expression

$$S_X(\omega) = \frac{\pi}{2}A^2[\delta(\omega-\omega_0) + \delta(\omega+\omega_0)]. \tag{6.233}$$

This is clearly a spectrum of a signal which consists of two 'teeth', one at $\omega_0$ and another at $-\omega_0$ as one would expect. See Fig. 6.14. ❐

## *6.4.3 Continuous Random Processes: Time Domain*

The description of stochastic processes in the time domain in Sect. 6.4.1 is somewhat unsatisfactory because it involves ensemble averages which can be technically difficult to use. In the previous Sect. 6.4.2, a frequency domain version of the same concepts is given but this also involves measurements which must be carried out over long time periods. This is also technically inconvenient. In this section a method of describing the running time development of stochastic averages will be developed. This will make it more convenient to deal with stochastic processes in real time observer and control systems. This differential equation method is also part of the foundation of the concept of Kalman filtering and also makes it possible to deal with noise in nonlinear systems under certain conditions.

Consider now a linear dynamic system driven by white noise. The system being considered is

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}_v(t)\mathbf{v}(t), \tag{6.234}$$

where $\mathbf{A}(t)$ and $\mathbf{B}_v(t)$ are possibly time varying dynamic and input matrices. $\mathbf{v}(t)$ is here wide sense stationary, white noise with a zero mean and $\mathbf{R}_v(\tau) = \mathbf{V}(\tau)\delta(\tau)$. The initial conditions (I.C.) for the integration are

$$\text{I.C.}: E\{\mathbf{x}(t_0)\} = \mathbf{m}_0, \tag{6.235}$$

$$E\{\mathbf{x}_0(t_0)\mathbf{x}_0^T(t_0)\} = \mathbf{Q}(t_0). \tag{6.236}$$

The solution to the state equation (6.234) can be written formally as

$$\mathbf{x}(t) = \Phi(t, t_0)\ \mathbf{x}(t_0) + \int_{t_0}^{t} \Phi(t, \tau)\ \mathbf{B}_v(\tau)\ \mathbf{v}(\tau)\ d\tau, \tag{6.237}$$

where $\Phi(t, t_0)$ is the transition matrix of the system. This makes it possible to form the product

$$\begin{aligned}\mathbf{x}(t_1)\mathbf{x}^T(t_2) = {} & \Phi(t_1, t_0)\ \mathbf{x}(t_0)\mathbf{x}^T(t_0)\Phi^T(t_2, t_0) \\ & + \Phi(t_1, t_0)\mathbf{x}(t_0)\left[\int_{t_0}^{t_2} \Phi(t_2, \tau)\ \mathbf{B}_v(\tau)\mathbf{v}(\tau)d\tau\right]^T \qquad (6.238) \\ & + \left[\int_{t_0}^{t_1} \Phi(t_1, \tau)\ \mathbf{B}_v(\tau)\mathbf{v}(\tau)d\tau\right]\mathbf{x}^T(t_0)\Phi^T(t_1, t_0) \\ & + \left[\int_{t_0}^{t_1} \Phi(t_1, \tau)\ \mathbf{B}_v(\tau)\mathbf{v}(\tau)d\tau\right]\left[\int_{t_0}^{t_2} \Phi(t_2, \tau)\ \mathbf{B}_v(\tau)\mathbf{v}(\tau)d\tau\right]^T. \end{aligned} \tag{6.239}$$

Taking the expectation value of both sides of this equation, remembering that the white noise source is given by a delta function and letting $t = t_1 = t_2$, an equation for $\mathbf{Q}(t)$ can be found:

$$\begin{aligned}\mathbf{Q}(t) &= \mathbf{R}_v(t, t) \\ &= \Phi(t, t_0)\mathbf{Q}(t_0)\Phi^T(t, t_0) + \int_{t_0}^{t} \Phi(t, \tau)\mathbf{B}_v(\tau)\mathbf{V}(\tau){\mathbf{B}_v}^T(\tau)\Phi^T(t, \tau)d\tau.\end{aligned} \tag{6.240}$$

To convert this integral expression into a differential equation it is only necessary to differentiate it with respect to time. This results in the expression

$$\begin{aligned}\dot{\mathbf{Q}}(t) = \frac{d\mathbf{Q}(t)}{dt} = {} & \frac{d\Phi(t, t_0)}{dt}\mathbf{Q}(t_0)\Phi^T(t, t_0) + \Phi^T(t, t_0)\mathbf{Q}(t_0)\frac{d\Phi^T(t, t_0)}{dt} \\ & + \frac{d}{dt}\left[\int_{t_0}^{t} \Phi(t, \tau)\mathbf{B}_v(\tau)\mathbf{V}(\tau){\mathbf{B}_v}^T(\tau)\Phi^T(t, \tau)d\tau\right].\end{aligned} \tag{6.241}$$

Using now Leibnitz' rule in the form,

$$\frac{d}{dt}\int_0^t f(t, \tau)d\tau = f(t, t) + \int_0^t \frac{\partial f(t, \tau)}{\partial t}d\tau, \tag{6.242}$$

and, remembering that,

$$\frac{d\Phi(t, \tau)}{dt} = \mathbf{A}(t)\Phi(t, \tau), \tag{6.243}$$

for every $t$ and $\tau$, one obtains from Eq. (6.241) the expression

$$\dot{\mathbf{Q}}(t) = \mathbf{A}(t)\mathbf{Q}(t) + \mathbf{Q}(t)\mathbf{A}^T(t) + \mathbf{B}_v(t)\mathbf{V}(t)\mathbf{B}_v^T(t). \tag{6.244}$$

This equation is sometimes called the degenerate Riccati equation (the second order nonlinearities are not present) but the name which will be used here is the time dependent Lyapunov equation. This equation describes the propagation of white noise through a dynamic system both in time and in the stationary and transient cases. It is thus of a completely different nature than the noise descriptions in Sects. 6.4.1 and 6.4.2 above as it is in a form which makes it directly useful for estimation and control purposes on a digital computer. It is the basis of optimal stochastic estimators or Kalman filters as will be seen in the next chapter.

Another important quality of Eq. (6.244) which should be observed is that it only describes the main statistical property of zero mean Gaussian stochastic signals which propagate through linear systems or networks, namely the variance (or squared standard deviation) of the signal. No statement is made about the details of any particular realization of $\mathbf{v}(t)$ or of higher moments of more complex stochastic signals. The reason for this is pure necessity: there is no particular time function which can be written down for $\mathbf{v}(t)$ and more complexity is difficult to treat simply and compactly (and most often is not required).

Equation (6.244) is particularly useful in the stationary state for time invariant stable systems where it reduces to

$$\mathbf{A}\mathbf{Q}_\infty + \mathbf{Q}_\infty\mathbf{A}^T + \mathbf{B}_v\mathbf{V}\ \mathbf{B}_v^T = 0, \tag{6.245}$$

where the infinity subscript on $\mathbf{Q}$ indicates that this is the steady state value of $\mathbf{Q}(t)$. This form of the equation is in general called the Lyapunov (matrix) equation which explains the choice of name above for Eq. (6.244). From Eq. (6.244) it is clear that the solution of Eq. (6.245) is the value of variance matrix at large times:

$$\lim_{t\to\infty}\mathbf{Q}(t) = \mathbf{Q}_\infty = \int_0^\infty e^{\mathbf{A}\tau}\mathbf{B}_v\mathbf{V}\mathbf{B}_v{}^T e^{\mathbf{A}\tau}d\tau. \tag{6.246}$$

Using this equation is often much easier than using Eqs. (6.193) and (6.214): only algebraic equations need be solved. Integrating is often more difficult and time consuming if it is at all possible.

***Example 6.28.*** **A Low Pass Filter with White Noise Input**

Consider a low pass filter described by the state and output equations

$$\dot{x}(t) = \frac{1}{\tau}[-x(t) + u(t)] = \omega_0[-x(t) + u(t)], \quad y(t) = x(t), \tag{6.247}$$

where $\tau$ is the filter time constant and $\omega_0$ is the filter cutoff frequency. Note that if the derivative on the left is zero (stationary state) then $x = u$ and the gain of the filter is 1 for $\omega \to 0$.

The transfer function for the low pass filter can be obtained by Laplace transforming the state and output equations:

$$\begin{aligned} sX(s) &= \frac{1}{\tau}[X(s) + U(s)] \;\Rightarrow\; \left(s + \frac{1}{\tau}\right)X(s) = \frac{1}{\tau}U(s), \\ Y(s) &= X(s), \end{aligned} \tag{6.248}$$

$$\Rightarrow\; H(s) = \frac{X(s)}{U(s)} = \frac{Y(s)}{U(s)} = \frac{1}{1 + s\tau} = \frac{1}{1 + \frac{s}{\omega_0}}. \tag{6.249}$$

It is easy to check that this is the correct answer. Using Mason's formula it is found that

$$H(s) = \frac{\frac{1}{\tau}\frac{1}{s}}{1 - \left(-\frac{1}{\tau}\frac{1}{s}\right)} = \frac{1}{1 + s\tau}. \tag{6.250}$$

It is now assumed that the filter is excited by white noise at its input with an intensity $V = \sigma_u^2$. The noise variance at the output of the filter can be found by using Eqs. (6.193) or (6.214),

$$\sigma_y^2 = \sigma_x^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_y(\omega)d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(j\omega)S_u(\omega)H(-j\omega)d\omega. \tag{6.251}$$

This quantity is an expression of the power in the output signal, $y(t)$. The problem above involves a scalar system, moreover, $S_u(\omega) = \sigma_u^2$, a constant, for all $\omega$ and so (see footnote below):

$$\begin{aligned} \sigma_y^2 &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(j\omega)|^2\sigma_u^2 d\omega = \frac{\sigma_u^2}{2\pi}\int_{-\infty}^{\infty} \frac{1}{1 + \left(\frac{\omega}{\omega_0}\right)^2}\omega_0\, d\left(\frac{\omega}{\omega_0}\right) && (6.252) \\ &= \frac{\omega_0}{2}\sigma_u^2. && (6.253) \end{aligned}$$

Notice that the output noise power increases linearly with the cutoff frequency of the low pass filter. See also Fig. 6.15, p. 406.

It is possible to calculate the result of Eq. (6.253) in another way using the time independent Lyapunov equation. The stationary noise variance (or covariance) for the low pass filter can be found by solving the equation:

The integral in Eq. (6.252) can be found in standard definite integral tables see equation (6.256).

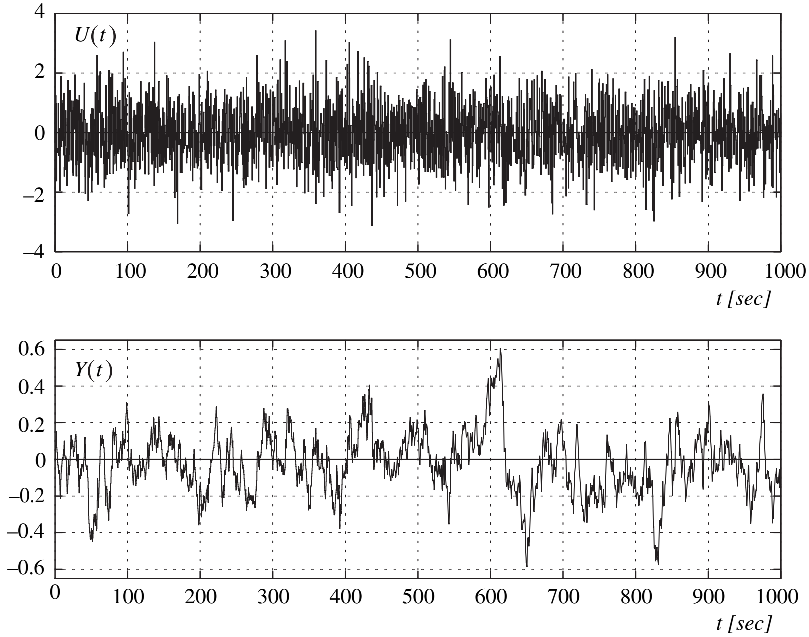


**Fig. 6.15** Simulation of a low pass filtered white noise signal. *Top*: filter input, *bottom*: filter output

$$AQ_\infty + Q_\infty A^T + BVB^T = \left(-\frac{1}{\tau}\right)Q_\infty + Q_\infty\left(-\frac{1}{\tau}\right) + \left(\frac{1}{\tau}\right)V\left(\frac{1}{\tau}\right) = 0. \quad (6.254)$$

$$\Rightarrow \; \sigma_y^2 = \frac{1}{2\tau}\sigma_u^2 = \frac{\omega_0}{2}\sigma_u^2. \quad (6.255)$$

This result has been verified (roughly) by simulation and the graphic results are shown in see Fig. 6.15, above.
Note:

$$\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx = \pi. \quad (6.256)$$

❐

Figure 6.15 is a simulation which shows the result of filtering a white noise signal (with a standard deviation of 1) with a first order low pass filter with a time constant of 10 s. According to Eq. (6.253) or (6.255) this should result in a white noise signal with a standard deviation of $\sqrt{\omega_0/2} = \sqrt{1/20} = 0.224$. The value actually observed is about 0.205 with the realization shown, using the $\pm 3\sigma$ rule of thumb in Example 6.5. However it should be remembered that 0.224 is an average statistical result over a large number of possible realizations, not a single one as in this example.

In this example white noise is integrated using a standard integration routine for a deterministic differential equation. In fact the addition of a noise as a driving source converts the deterministic differential equation the low pass filter (Eq. (6.247)) into a stochastic differential equation. Special numerical integration methods must be used for such differential equations so that the results shown above in Fig. 6.15 are somewhat fortuitously accurate. Further information on this important subject can be found below in the following Sect. 6.4.4, p. 408.

***Example 6.29.* The Langevin Equation**

It is interesting now to return to the Langevin equation treated earlier in Example 6.20. If it is desired to describe the overall movement of the pollen particle, position as well as velocity, the state equations which must be solved are

$$\frac{dx}{dt} = v, \tag{6.257}$$

$$\frac{dv}{dt} = -\frac{b}{m}v + \frac{1}{m}v_1(t), \tag{6.258}$$

where the system is now to be driven by a white noise process, $v_1(t) = F(t)\delta(t - t_0)$, as in Eq. (6.171), with covariance $r_v$. The **A** and $\mathbf{B}_v$ matrices for the system are thus

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{b}{m} \end{bmatrix} \text{ and } \mathbf{B}_v = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}. \tag{6.259}$$

Using these matrices in the time dependent Lyapunov equation (6.244) and writing down the three differential equations which result for the matrix elements of **Q**, one has

$$\dot{q_{11}} = 2q_{12}, \tag{6.260}$$

$$\dot{q_{12}} = -\frac{b}{m}q_{12} + q_{22}, \tag{6.261}$$

$$\dot{q_{22}} = -\frac{2b}{m}q_{22} + \frac{r_v}{m^2}. \tag{6.262}$$

Assuming that the pollen particle starts from a known zero position, the initial conditions for the integration of these equation are $q_{11}(0) = 0$, $q_{12}(0) = 0$ and $q_{22}(0) = r_v/(2bm)$. Use of these initial conditions makes it possible to integrate equations (6.260, 6.261, 6.262). The solutions obtained are

$$q_{11}(t) = \frac{r_v}{b^2}\left[t - \frac{m}{b}\left(1 - e^{-\frac{b}{m}t}\right)\right], \tag{6.263}$$

$$q_{12}(t) = \frac{r_v}{2b^2}\left(1 - e^{-\frac{b}{m}t}\right), \tag{6.264}$$

$$q_{22}(t) = \frac{r_v}{2bm}. \tag{6.265}$$

For small times such that $t \ll \tau_r = \frac{m}{b}$, using the series expansion

$$e^{-\frac{t}{\tau_r}} = 1 - \frac{t}{\tau_r} + \frac{1}{2}\left(\frac{t}{\tau_r}\right)^2 - \ldots, \tag{6.266}$$

it is clear that

$$\langle X^2 \rangle = \frac{r_v}{b^2} t^2, \tag{6.267}$$

and the particle moves as though it is a free particle with a constant speed

$$v = \sqrt{r_v/b^2}. \tag{6.268}$$

For large times such that $t \gg \tau_r = \dfrac{m}{b}$, Eq. (6.266) becomes

$$\langle X^2 \rangle = \frac{r_v}{b^2} t, \tag{6.269}$$

and the particle moves as though it is diffusing and executing a random walk.

It can be shown that covariance is related to the thermally induced kinetic energy of the particle and that

$$r_v = 2bkT, \tag{6.270}$$

where $k$ is Boltzmann's constant and $T$ is the absolute temperature. Thus the mean value of the particle position at large times is

$$\langle x^2 \rangle = \frac{kT}{3\pi\mu a} t = \frac{D}{3\pi\mu} t, \tag{6.271}$$

where $a$ is the radius of the particle, $\mu$ is the coefficient of viscosity and $D = (kT)/a$ is the diffusion constant. ❒

### 6.4.4 Inserting Noise into Simulation Systems

In many cases it is necessary to insert noise sources into systems to simulate the effect of state noise, disturbances, signal inputs or measurement noise. Most often this is done by assuming that the noise is small compared to the input or state signals levels. ‘Small’ in this case means a small amplitude in relation to the size of the deterministic signals and/or sizes of the nonlinearities in the system.

Assume that the noise is to be inserted into the state equation,

$$\dot{\mathbf{x}(t)} = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), \mathbf{v}_1(t)), \tag{6.272}$$

where $\mathbf{v}_1(t)$ is a zero mean noise source which also may be a vector process. In a general nonlinear system, $\mathbf{v}_1(t)$ may be added to the states or the states may be a function of the noise or the noise may be added to only some of the states. Linearization of Eq. (6.272) may be carried out in the usual way to obtain

$$\dot{\Delta\mathbf{x}} = \mathbf{A}\Delta\mathbf{x} + \mathbf{B}\Delta\mathbf{u} + \mathbf{B}_v\Delta\mathbf{v}_1, \tag{6.273}$$

where

$$\mathbf{A} = \left.\frac{\partial\mathbf{f}}{\partial\mathbf{x}}\right|_{\substack{\mathbf{x}=\mathbf{x}_0\\ \mathbf{u}=\mathbf{u}_0\\ \mathbf{v}_1=\mathbf{v}_0}}, \quad \mathbf{B} = \left.\frac{\partial\mathbf{f}}{\partial\mathbf{u}}\right|_{\substack{\mathbf{x}=\mathbf{x}_0\\ \mathbf{u}=\mathbf{u}_0\\ \mathbf{v}_1=\mathbf{v}_0}} \text{ and } \mathbf{B}_v = \left.\frac{\partial\mathbf{f}}{\partial\mathbf{v}_1}\right|_{\substack{\mathbf{x}=\mathbf{x}_0\\ \mathbf{u}=\mathbf{u}_0\\ \mathbf{v}_1=\mathbf{v}_0}}$$

and the nominal linearization point is $\mathbf{x} = \mathbf{x}_0$, $\mathbf{u} = \mathbf{u}_0$ and $\mathbf{v}_1 = \mathbf{v}_0 = 0$, by definition. The noise amplitude is then $\Delta\mathbf{v}_1$. Similar expressions to that in Eq. (6.273) may be associated with ‘small’ noise sources associated with the inputs or with other sources built into the control object or its associated systems. In some cases it may be necessary to use the chain rule for differentiation in order to find the matrix $\mathbf{B}_v$.

To simulate the noise with the proper amplitude it is necessary to determine its standard deviation (amplitude) or variance (power). In Matlab/Simulink this can be done using either of the internal generators: Random Number or Bandlimited Noise. Before using either generators the documentation for it should be consulted.

### Random Number Generator

Noise which is effectively white can be generated by choosing a sample time which is 10–20 times smaller than the smallest time constant in the system being studied. This is equivalent to saying that the sampling frequency is 10–20 times the highest cutoff frequency in the subject system. As the random numbers generated have a variance of 1, the white noise generators must be followed by a gain element which adjusts the amplitude of the noise to the level required in the simulation. If the approximate maximum amplitude of the noise signal is $\pm A$ then the noise amplitude should be selected to be $2A/6 = A/3$ as the noise will then remain in this window 99.75% of the time:

this is the 6σ limit, 6 times the standard deviation. Outliers will of course occur occasionally outside this window.

Many other simulation packages have noise generators based on the use of random number generators.

### Band Limited White Noise

In Simulink the Band Limited White Noise is generated using a random sequence with a correlation time much smaller than the shortest time constant in the system. This correlation time is the sampling time used in the block. In the documentation of the block it is suggested that the correlation time be selected as

$$t_c = \frac{1}{c}\frac{2\pi}{\omega_c}, \tag{6.274}$$

where $c = 100$ and $\omega_c$ is the cut-off frequency of the system. $c$ can be selected smaller, say 10–20 for typical systems. Choosing an extremely small correlation time can increase the sampling time significantly due to the large number of rapid transients introduced into the integration algorithms.

The algorithm used in the block automatically scales the white noise so that the generator outputs a noise covariance which is the noise power, $\sigma^2$, divided by the correlation time, $\sigma^2/t_c$. The purpose of this scaling is to ensure that the approximate white noise has the same covariance as the system would have if responding to real white noise (with an infinite frequency spectrum).

The ‘white’ noise generated using either of the methods above is equivalent. If true white noise is to be inserted into a system, then it is a good idea to low pass filter it before inserting into the system. This will make it easier for the integration routines used to solve differential equations to integrate the system equations without having to deal with the singularities which the steps in the white noise sequences represent. The variance of the noise from low pass filtered white noise is given in Eq. (6.253).

### Reproducing White Noise

In principle a basic characteristic of white noise is that it can never be reproduced because if it were reproduceable, it would have to be correlated with itself. This could be inconvenient in simulation studies because one would always have to look at a system with different noise inputs. Fortunately it is possible to avoid this problem using standard random noise generators. This is done by choosing different seeds for the random number generators. By using the same seed (or seeds) for different simulations, the identically same sequence(s) can be generated reproducably: the generators are deterministic with respect to the sequences which they generate with a given seed. They are however apparently stochastic for the shorter periods of time in which they are used for simulations. If different independent noise sequences are required in different blocks, the seeds used in the different independent generators must be selected to be different.

### Other ‘Noise’ Sources

In many cases other kinds of disturbance signals than white noise are encountered. In such cases it is the Root Mean Square (RMS) value of the signal which is that with should be compared with the standard deviation used for white noise amplitudes. The effective power of for example sine wave is its rms value squared. The effective power of any deterministic signal can be easily found as required. Note that the frequency spectra of noise signals other than Gaussian white noise are very different from white noise: they are in general not white (have flat frequency spectra). This must be taken into account for in simulating them.

### Integrating Stochastic Differential Equations

While it straight forward to insert Gaussian white noise signal into linear or nonlinear systems accurately with the correct mean value, variance and density (or distribution) function, it is not easy to obtain the correct states and outputs using the standard integration routines available in Matlab or Simulink. The reason for this is that inserting noise into differential equations automatically makes them stochastic differential equations which have different mathematical characteristics than deterministic differential equations.

Whereas the integration of deterministic differential equations is based on a first order approximation to their solution, stochastic differential equations must be based on a second order approximation. The reason for this is Eq. (6.157) which implies that at each time step, *h*, an integral of a noise driving term in a differential equation must be updated as

$$\Delta w = \sigma\sqrt{h}, \tag{6.275}$$

where *w* is a Wiener process. This obviously is not possible using a normal integration routine. Equation (6.275) results in the integral of *w* being,

$$\int_0^h w\,dw = \frac{\sigma^2}{2}\left[w(h)^2 - h\right], \tag{6.276}$$

which is not the result which one would expect. This result is obtained using one interpretation of the stochastic integral (the Ito interpretation, see Sect. 6.6.3). Clearly then, an important error will be introduced by using ordinary integration rules and numerical integration algorithms. See Gran (2007), Chapter 5, for more detail.

In general in control applications, the errors introduced using ordinary integration algorithms on systems with noise sources are ignored. It is assumed that the errors introduced (given reasonable signal to noise ratios) are small enough that they may be overcome by the feedback and/or estimation loop(s) in the system. This fortunately also often proves to be the case. One can expect that the results using ordinary integration routines are accurate to an approximation which is no better than $0(h^{3/2})$ under the condition that a fixed step size is used and the noise is assumed constant and equal to $w(h)$ over the step interval. Significant errors may occur if these conditions are not met, depending on the nonlinearities in the control object and the signal to noise ratio. The covariance errors which might be expected will generally be on the order of 5–10% for reasonable signal to noise ratios.

### Evaluating Noise Covariances

Once a system has been simulated using random noise sources for the state and measurement noise, one often wishes to evaluate the results and compare them to the theoretical predictions of say Eqs. (6.244) or (6.245) which are correct for linear systems. This of course cannot be accomplished in a completely satisfactory manner because 1. standard integration routines cannot integrate stochastic differential equation correctly and 2. one can only obtain at one noise realization at a time.

Thus one usually resorts to using standard integration routines and adjusting the sample time of the noise generators so that the results approximate those obtained with Eqs. (6.244) or (6.245). In this process one must insure that the sample time is still on the order of 10 times less than the smallest time constant in the system. To reduce the statistical uncertainty, many runs are usually conducted using a Monte Carlo selection of the seeds in the noise generators and averaging states and outputs over say 10–100 runs. The noise covariances can be calculated using the MATLAB algorithm `cov(X)` or `cov(X,Y)`.

In some cases, for example for financial applications, accurate integration results must be obtained. Such a case could be predicting commodity prices for option speculation. Under these conditions one can find specialized integration routines for stochastic differential equations of varying accuracy and complexity. A useful introduction to the theory of stochastic differential equations may be found in Soong (1973) and specialized integration routines for stochastic differential equations in Gran (2007).

## *6.4.5 Discrete Time Stochastic Processes*

Discrete time random processes have characteristics which are analogous to those of continuous processes and are described in very much the same way. For this reason it is not necessary to go into detail to reformulate the definitions and rules which have been given for continuous time processes. It suffices to sketch roughly a few concepts and to let the reader deal with discrete time systems as analogies to the material which has already been presented above.

Discrete time random variables have already been defined in Sect. 6.3.1. Discrete time vector stochastic processes can be defined by specifying all of its joint probability distributions. If for all real $\mathbf{x}_1$, $\mathbf{x}_2$, ... , $\mathbf{x}_\mathrm{n}$ and for all integers $i_1$, $i_2$, ... , $i_\mathrm{n}$ and all integers $n$ and $k$,

$$\begin{aligned} &Pr(\mathbf{X}(i_1) \leq \mathbf{x}_1, \mathbf{X}(i_2) \leq \mathbf{x}_2, \ldots, \mathbf{X}(i_n) \leq \mathbf{x}_n) = \\ &\qquad Pr(\mathbf{X}(i_1 + k) \leq \mathbf{x}_1, \mathbf{X}(i_2 + k) \leq \mathbf{x}_2, \ldots, \mathbf{X}(i_n + k) \leq \mathbf{x}_n), \end{aligned} \tag{6.277}$$

then the process is stationary, where the vector inequalities are to be understood as earlier, component for component. It is also Gaussian if all of the component distributions are Gaussian.

The mean or expectation value of the stochastic process $\mathbf{X}(k)$ is defined by the equation

$$\mathbf{m}(k) = E\{\mathbf{X}(k)\}. \tag{6.278}$$

Its second order joint moment matrix or autocorrelation matrix is

$$\mathbf{C}_X(i,j) = E\{\mathbf{X}(i)\ \mathbf{X}^T(j)\}, \tag{6.279}$$

while the covariance matrix of the process is

$$R_X(i,j) = E\{[\mathbf{X}(i) - \mathbf{m}(i)][\mathbf{X}(j) - \mathbf{m}(j)]^T\}. \tag{6.280}$$

The variance of the process is

$$\mathbf{Q}_X(k) = E\{[\mathbf{X}(k) - \mathbf{m}(k)][\mathbf{X}(k) - \mathbf{m}(k)]^T\} \tag{6.281}$$

and $\mathbf{C}_X(k, k)$ is the second order moment matrix.

A strictly stationary process is one which has a constant mean value(s) and, at the same time, its joint moment and covariance matrices depend only on the time difference $i$–$j$. A wide sense stationary discrete time process is one which has the characteristics that

$$1.\ \mathbf{m}(t) = \langle E\{\mathbf{x}(k)\}\rangle = \text{constant (in its components)}, \tag{6.282}$$

$$2.\ \mathbf{C}_X(k,k) = E\{\mathbf{X}(k)\ \mathbf{X}^T(k)\} < \infty, \text{in its elements}, \tag{6.283}$$

$$3.\ \mathbf{C}_X(i,j) = E\{X(i)\ \mathbf{X}^T(j)\} = \mathbf{C}_X(i-j), \tag{6.284}$$

or

$$4.\ \mathbf{R}_X(i,j) = E\{[\mathbf{X}(i) - \mathbf{m}(i)][\mathbf{X}(j) - \mathbf{m}(j)]^T\} = \mathbf{R}_X(i-j). \tag{6.285}$$

### *6.4.6 Translating Continuous Noise into Discrete Time Systems*

The model of Eqs. (6.1) and (6.2) commonly arises from physical considerations in continuous time. This is true not only of the deterministic portion of the model but also the stochastic part of it. For this reason it is often necessary to convert a continuous time model to a discrete time one for design purposes. This means that discrete time equivalents are required of the state noise and the measurement noise process models.

A discrete time state noise intensity, $\mathbf{V}_{1d}$, can be derived from a continuous time one, $\mathbf{V}_1$, using the following considerations. Let the sampling time be defined as $T$ and consider the model of Eq. (6.1). If the state noise is ideal low pass filtered white noise, $\mathbf{v}(t)$, (with bandwidth $1/(2T)$) then sampling this source results in the equation

$$\mathbf{v}(k) = \int_{kT}^{(k+1)T} e^{\mathbf{A}[(k+1)T-\tau]}\mathbf{B}_v\mathbf{v}(\tau)\ d\tau. \tag{6.286}$$

Taking the expectation value of both sides of this equation shows that the mean value of the stochastic variable, $\mathbf{v}(k)$, is zero because the mean value of $\mathbf{v}(t)$ is zero.

The covariance of the process, $\mathbf{v}(t)$, has to be calculated from the equation

$$\begin{aligned}\mathbf{V}_{1d} &= E\{\mathbf{v}(k)\mathbf{v}^T(k)\}\\ &= \int_{kT}^{(k+1)T}\int_{kT}^{(k+1)T} e^{\mathbf{A}[(k+1)T-\tau]}\mathbf{B}_v E\{\mathbf{v}(\tau)\mathbf{v}^T(\sigma)\}\mathbf{B}_v{}^T e^{\mathbf{A}^T[(k+1)T-\sigma]}d\tau\ d\sigma\\ &= \int_{kT}^{(k+1)T} e^{\mathbf{A}[(k+1)T-\tau]}\mathbf{B}_v\mathbf{V}_1\ \mathbf{B}_v^T e^{\mathbf{A}^T[(k+1)T-\tau]}d\tau, \qquad (6.287)\end{aligned}$$

where the last equality comes from the fact that the noise source $\mathbf{v}(t)$ is white and thus $E\{\mathbf{v}(\tau)\mathbf{v}^T(\sigma)\} = \mathbf{V}_1\delta(\tau-\sigma)$. Making the variable changes $u = (k+1)T - \tau$ and $-\tau = u$, the following equation relating $\mathbf{V}_{1d}$ and $\mathbf{V}_1$ results:

$$\mathbf{V}_{1d} = \int_0^T e^{\mathbf{A}\tau}\mathbf{B}_v\mathbf{V}_1\ \mathbf{B}_v^T e^{\mathbf{A}\tau}d\tau. \tag{6.288}$$

This integral can be difficult to evaluate but it is often possible to make an approximation which is valid when the sampling time is small compared to the smallest effective time constant in the system. The exponential under the integral sign can under this assumption be replaced by its series expansion,

$$e^{\mathbf{A}\tau} \cong 1 - \mathbf{A}\tau + \frac{1}{2!}\mathbf{A}^2\tau^2 - \ldots. \tag{6.289}$$

This results in an approximation to an order in $T^n$ which can be selected by choosing the relevant number of terms, $n$, in the series expansion. In general only one term in the series expansion is sufficient for most practical cases. This gives the expression

$$\mathbf{V}_{1d} = \mathbf{B}_v \mathbf{V}_1 \ \mathbf{B}_v^T T. \tag{6.290}$$

This expression shows approximately how to convert a continuous time process noise to its discrete time equivalent. This expression is only valid if $T$ is much smaller that the smallest system time constant. For further details see Franklin, et al., (1990), p. 454.

To discretize the measurement noise covariance it should be noticed that the covariance of a continuous white noise process is

$$\mathbf{R}_w(\tau) = \mathbf{V}_2 \delta(\tau). \tag{6.291}$$

This process is to be approximated with a discrete white noise process, $\mathbf{w}(k)$, which can be written

$$\mathbf{R}_w(k) = \mathbf{V}_{2d} \delta(k), \tag{6.292}$$

where $\delta(k)$ is the Kronecker delta and

$$\delta(k) = \begin{cases} 1, & \text{if } \ \mathrm{k} \ = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{6.293}$$

The transition will obviously involve a limiting process as the delta function is a singularity which only has meaning under an integral sign. Define the rectangle approximation to the continuous delta function as,

$$\delta(t) = \lim_{T \to 0} \perp \left(\frac{t}{T}\right), \tag{6.294}$$

where

$$\perp \left(\frac{t}{T}\right) = \begin{cases} \dfrac{1}{T}, \text{ if} - \dfrac{T}{2} \le t \le \dfrac{T}{2} \\ 0, \text{ otherwise} \end{cases}. \tag{6.295}$$

This suggests the identification,

$$\mathbf{V}_2 \delta(t) = \lim_{T \to 0} (\mathbf{V}_{2d} T) \perp \left(\frac{t}{T}\right), \tag{6.296}$$

and hence that

$$\mathbf{V}_{2d} = \frac{\mathbf{V}_2}{T}. \tag{6.297}$$

This is of course because the smallest time which can be resolved in a discrete time system is a sampling time. The effect of sampling on the measurement noise intensity is thus to increase the measurement noise intensity (if the sampling time is small). For further details see Franklin, et al., (1990), p. 456.

### *6.4.7 Discrete Random Processes: Frequency Domain*

The power spectral density matrix, $\mathbf{S}(\omega)$, of a wide sense, discrete time stationary stochastic process, $\mathbf{X}(k)$, can be defined as

$$\mathbf{S}_X(\omega) = \sum_{k=-\infty}^{\infty} \mathbf{R}_X(k)e^{-jk\omega} \tag{6.298}$$

which is valid for $-\pi \leq \omega \leq \pi$ or equivalently $-f_s/2 \leq f \leq f_s/2$ where

$$\omega = 2\pi fT = 2\pi\frac{f}{f_s}, \tag{6.299}$$

and $T$ is the sampling period, $f$ is the frequency and $f_s = 1/T$ is the sampling frequency.

An important property of the spectral density matrix for a zero mean, wide sense stationary, discrete time stochastic process is that

$$E\{\mathbf{X}(k)\ \mathbf{X}^T(k)\} = \mathbf{R}_X(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{S}_X(\omega)d\omega. \tag{6.300}$$

This is true because by definition,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{S}_X(\omega)d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_{k=-\infty}^{\infty} \mathbf{R}_X(k)e^{-jk\omega}\right]d\omega \tag{6.301}$$

$$= \sum_{k=-\infty}^{\infty} \mathbf{R}_X(k)\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-jk\omega}d\omega\right] = \mathbf{R}_X(0), \tag{6.302}$$

and a discrete time delta function is given by

$$\delta(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-jk\theta}d\theta = \begin{cases} 1, & \text{for k = 0} \\ 0, & \text{otherwise} \end{cases}. \tag{6.303}$$

The transfer function picture of the propagation of a stochastic process through a dynamic system is also valid for a discrete time system. An asymptotically stable linear time invariant system which is described by the transfer function $\mathbf{H}(z)$ is assumed to have as an input, $\mathbf{U}(k)$, a wide sense stationary discrete time stochastic process with the spectral density matrix, $\mathbf{S}_u(\omega)$. The output, $\mathbf{Y}(k)$, of such a system is a realization of a discrete time stochastic process which has the spectral density matrix, $\mathbf{S}_y(\omega)$, given by

$$\mathbf{S}_y(\omega) = \mathbf{H}(e^{j\omega})\mathbf{S}_u(\omega)\mathbf{H}^T(e^{-j\omega}). \tag{6.304}$$

The description of white noise in discrete time is very close to that which is used for white noise in continuous time. Consider a stochastic process which is a sequence of mutually uncorrelated, zero mean, stochastic variables with constant variance matrix, $\mathbf{Q}$. The covariance of this wide sense stationary process is then

$$\mathbf{R}_u(i-j) = \mathbf{Q}\delta(i-j)\begin{cases} \mathbf{Q}, & \text{for } \mathrm{i=j} \\ 0, & \text{for } \mathrm{i \neq j} \end{cases}, \tag{6.305}$$

where the Kronecker delta function can be written

$$\delta(i-j) = \begin{cases} 1\,, & \text{for } \mathrm{i=j} \\ 0\,, & \text{for } \mathrm{i \neq j} \end{cases}. \tag{6.306}$$

The spectral density matrix of the process is then

$$\mathbf{S}_u(\omega) = \mathbf{Q}, \tag{6.307}$$

which is the discrete time analog of white noise. It is not an equivalent to continuous time white noise because it has a somewhat different frequency spectrum due to the effect of sampling. This is because the Fourier transform of equation ( 6.298) is that of a pulse of finite width, the sampling time, as this is the minimum time which can be resolved in a discrete time system. This is not the same as a continuous time delta function.

***Example 6.30.*** **Sampled White Noise**

In reality discrete time white noise should be obtained by sampling band limited continuous time white noise. Practically it is never done this way. Discrete time noise is nearly always generated by drawing random numbers at constant sample times like the responses in Figs. 6.12 and 6.12. White noise is by definition Gaussian so it is the lower figure on Fig. 6.1 which is most often relevant. The sample number $n$ then becomes the time index, $k$.

The noise spectrum which results from using a random number generator is not white. This is because the spectrum which is generated can be seen as the Fourier transform of a flat, band limited, noise spectrum (with a bandwidth $\omega_c = 1/[2T]$), convoluted with a periodic delta functions at whole multiples of the sampling frequency. In detail the amplitude spectrum of band limited noise white is given by an expression of the form

$$V_d(j\omega) = \frac{\tau}{T}\frac{\sin\left(\frac{\omega\tau}{2}\right)}{\frac{\omega\tau}{2}} * \sum_{i=-\infty}^{\infty} \delta\left(\omega - i\frac{2\pi}{T}\right), \tag{6.308}$$

where the $*$ denotes convolution, $\tau$ is the width of the sampling pulse and $T$ is the time between samples. In the special case where $\tau = T$, the sampling pulses completely fill the time between samples. This is exactly the result which is usually obtained by using a zeroth order hold network which is the conventional way of sampling. In this case the equation above becomes

$$V_d(j\omega) = \frac{\sin\left(\frac{\omega T}{2}\right)}{\frac{\omega T}{2}} * \sum_{i=-\infty}^{\infty} \delta\left(\omega - i\frac{2\pi}{T}\right) \tag{6.309}$$

or equivalently in terms of ordinary frequencies,

$$V_d(jf) = \frac{\sin(\pi f T)}{\pi f T} * \sum_{i=-\infty}^{\infty} \delta\left(f - i\frac{1}{T}\right). \tag{6.310}$$

This is a collection of *sin* $x/x$ spectra, each centered at the sampling frequency and its harmonics: $f_{sn} = n/T$ or equivalently $\omega_{sn} = (2\pi n)/T$, where $n = \ldots - 2, -1, 0, 1, 2, \ldots$. It is clear that the sample time should be small enough that the main sampling spectrum and its harmonics should be separated from each other by at least $5/T$ for proper system noise performance.

Figure 6.16 below shows the main sampled white noise amplitude spectrum on a linear plot of the expressions in Eqs. (6.309) and (6.310). The main spectrum is that centered around $f_{s0} = 0$. Only the main spectrum is shown on the plots: its sampled copies (centered at multiples of the sample frequency) are suppressed for the sake of simplicity. Other details have also been suppressed for simplicity, see Papoulis (1977). In the top figure the sample frequency is 100 Hz and in the bottom it is 200 Hz. The dashed curves on both plots show the responses of first and second order systems with –3dB frequencies of 10 Hz. The second order system is critically damped and falls off more rapidly than the first order system.

It is clear that if the spectrum of Fig. 6.16 is to be used to represent white noise in a control system then the system bandwidth should not be above about $f_{s1}/10$ or so as shown. Otherwise the noise bandwidth will not be sufficiently large to give a good approximation of continuous time white noise. In fact it should be somewhat above this in some cases. This is easy to see from the log-log plots in Fig. 6.17 corresponding to those in Fig. 6.16. Here it is obvious that it is possible to use the 100 Hz sampled white noise for a second order system while the 200 Hz sampled white noise will be more reasonable for the first order system.

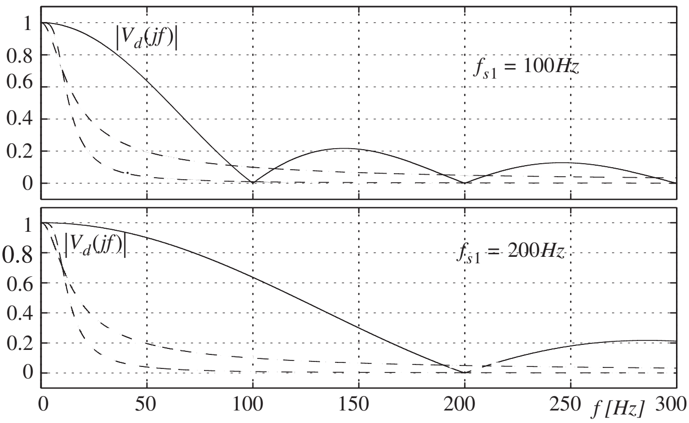


**Fig. 6.16** 'White' noise spectrum derived using a random number generator and a constant sample time T, plotted on linear axes. The sampling frequency is 100 Hz (*above*) and 200 Hz (*below*). A second order filter cuts off more rapidly than a first order one

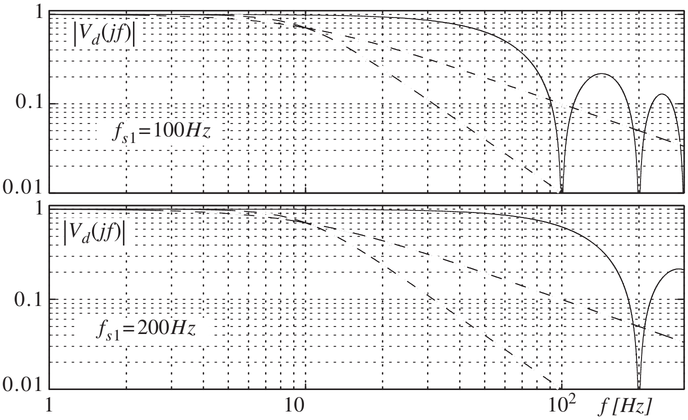


**Fig. 6.17** White noise spectrum of Fig. 6.16 plotted on logarithmic axes. Again a second order filter cuts off more rapidly than a first order one

### *Example 6.31.* Sampled Low Pass Filtered White Noise

A scalar low pass filtered white noise process, with intensity $\sigma^2$, has a autocovariance function which is

$$R_X(i-j) = \sigma^2 e^{-\left|\frac{(i-j)T}{\tau}\right|}, \tag{6.311}$$

where $T$ is the sampling time and $\tau$ is the time constant of the low pass filter. The power spectral density function of the process is then

$$S_X(\omega) = \frac{\sigma^2(1 - e^{-(2T)/\tau})}{(e^{j\omega} - e^{-T/\tau})(e^{-j\omega} - e^{-T/\tau})}. \tag{6.312}$$

It should be noted here that in order for this expression to be valid, it is necessary to sample the low pass filtered noise spectrum with a sampling frequency which is 4–10 times larger than the bandwidth of the spectrum. Otherwise aliasing will occur and the expression above will be invalid. ❐

### 6.4.8 Discrete Random Processes: Running in Time

What has to be accomplished is to find a discrete time analog to the time dependent Lyapunov equation. The dynamic linear difference equation which describes the propagation of a zero mean stochastic process, **v**(k), is

$$\mathbf{x}(k+1) = \mathbf{F}(k)\ \mathbf{x}(k) + \mathbf{G}_v(k)\ \mathbf{v}(k), \tag{6.313}$$

where the dynamic and input matrices are constant. The covariance of $\mathbf{v}(k)$ is $\mathbf{R}_v(k)$. Initial conditions for this process are $\mathbf{m}(0) = \mathbf{m}_0$ and $\mathbf{R}_v(0) = \mathbf{R}_0$.

Taking the expectation of each side of Eq. (6.313) yields

$$\mathbf{m}(k+1) = \mathbf{F}(k)\ \mathbf{m}(k), \tag{6.314}$$

because $\mathbf{v}(k)$ is a zero mean process. The mean value will thus propagate normally through the system.

To find out how the noise covariance propagates through the system it is useful to introduce the notation,

$$\mathbf{Q}(k) = E\{\tilde{\mathbf{x}}(k)\ \tilde{\mathbf{x}}^T(k)\}, \tag{6.315}$$

where $\tilde{\mathbf{x}}(k) = \mathbf{x}(k) - \mathbf{m}(k)$.

Clearly because of Eqs. (6.313) and (6.314), $\tilde{\mathbf{x}}(k)$ satisfies Eq. (6.313). Thus using this fact it is possible to form the equation,

$$\tilde{\mathbf{x}}(k+1)\ \tilde{\mathbf{x}}^T(k+1) = [\mathbf{F}(k)\ \mathbf{x}(k) + \mathbf{G}_v(k)\mathbf{v}(k)][\mathbf{F}(k)\mathbf{x}(k) + \mathbf{G}_v(k)\mathbf{v}(k)]^T \tag{6.316}$$

$$= \mathbf{F}(k)\tilde{\mathbf{x}}(k)\ \tilde{\mathbf{x}}^T(k)\ \mathbf{F}(k) + \mathbf{F}(k)\ \tilde{\mathbf{x}}(k)\mathbf{v}^T(k)\ \mathbf{G}_v^T(k) + \mathbf{G}_v(k)\tilde{\mathbf{x}}(k)\mathbf{v}^T(k)\ \mathbf{F}^T(k) + \mathbf{G}_v(k)\ \mathbf{v}(k)\mathbf{v}^T(k)\ \mathbf{G}_v^T(k). \tag{6.317}$$

Taking the expectation value of each side of this equation and remembering that there is no correlation between the noise source, $\mathbf{v}(k)$, and $\tilde{\mathbf{x}}(k)$, it is possible to obtain the equation

$$\mathbf{Q}(k+1) = \mathbf{F}(k)\mathbf{Q}(k)\mathbf{F}^T(k) + \mathbf{G}_v(k)\mathbf{R}_v(k)\mathbf{G}_v^T(k). \tag{6.318}$$

This equation shows how the noise covariance propagates in time and might be called the time dependent Lyapunov difference equation in analogy with the equivalent equation found for continuous time systems. It suggests that if the noise source $\mathbf{v}(k)$ is zero mean and Gaussian then its behavior is completely specified by Eqs. (6.314) and (6.318). Again, it is only exactly valid for linear systems which have a form like that in Eq. (6.313).

***Example 6.32*. A First Order Scalar Stochastic System**

A linear first order, discrete time system driven by a Gaussian, zero mean, white noise source $v(k)$ can be expressed as

$$x(k+1) = f\,x(k) + g\;v(k), \tag{6.319}$$

where $f$ and $g$ are constants. The noise covariance of $v(k)$ is assumed to be $r_v$. At the time $k_0$ the mean value of the process is $m_0$ while the initial covariance is $r_0$.

The mean value is given by the equation,

$$m(k+1) = fm(k), \tag{6.320}$$

and hence the mean value of $x(k)$ at any time is

$$m(k) = f^{k\,-\,k_0}m_0. \tag{6.321}$$

Equation (6.318) for the system above yields the expression

$$q(k+1) = f^2q(k) + g^2r_v. \tag{6.322}$$

Using iterative methods, the solution of Eq. (6.322) can be easily obtained, it is

$$q(k) = f^{2(k-k_0)}r_0 + \frac{1-f^{2(k-k_0)}}{1-f^2}g^2r_v. \tag{6.323}$$

In the stationary state, corresponding to $k_0 \to -\infty$, and assuming that $|f| < 1$, the following solutions are obtained:

$$\begin{aligned} m(k) &\to 0, \\ q(k) &\to \frac{g^2r_v}{1-f^2}. \end{aligned} \tag{6.324}$$

The covariance of the state noise can be calculated from

$$r_v(k+\tau,k) = r_v(\tau) = f^{\tau}q(k) = \frac{g^2f^{|\tau|}r_v}{1-f^2}. \tag{6.325}$$

This corresponds to a spectral density which is

$$\begin{aligned} S_x(\omega) &= \frac{1}{2\pi}\frac{g^2 r_v}{(e^{j\omega}-f)(e^{-j\omega}-f)} \\ &= \frac{1}{2\pi}\frac{g^2 r_v}{1+f^2-2f\cos(\omega)}. \end{aligned} \tag{6.326}$$

□

## 6.5 Summary

The main purpose of this chapter is to provide simple mathematical and physical tools to deal with the problem of analyzing the effect of stochastic signals in linear systems. Because of the nature of random signals (or more specifically noise), this has involved a review of methods of applying statistical and signal analysis methods to linear systems. Thus it has been necessary to take a very broad view of linear systems theory and it is this broad background which is often the main difficulty in learning about noise in dynamic systems.

The review of statistical methods has concentrated on first random variables and vectors and then on random scalar and vector processes. This has made it possible to introduce the concept of independent increment, random walk or Wiener processes in a natural way. As a bi-product of this treatment, white noise has emerged as a precondition for constructing a Wiener process. A Wiener process is just an integrated white noise process, subject to certain initial conditions. While this treatment is heavily dependent on a dynamic system picture formulated in the time domain, the name white noise itself comes from frequency domain considerations. Thus it has been a requirement that some basic results of signal analysis be reviewed for linear systems. Some of these results have made it clear that white noise is only a mathematical idealization which has to be used with some care in modelling noise sources in real systems. Experience shows however that reasonably accurate results can be obtained when this is done properly.

The main results of the chapter with respect to later chapters are two equations which can predict the time development of two of the main statistical measures which can be used on stochastic signals: the mean values and the covariance matrix. These measures are sufficient to completely characterize the propagation of Gaussian distributed noise in linear systems but give no information about any particular realization of the noise process. The propagation of the covariance matrix has been shown to be predicted by an equation which has been called the time dependent Lyapunov equation. This usage in not common but will be useful in what follows as it makes it possible to separate noise propagation from uncertainty reduction due to the use of measurements (and the associated measurement noise) in a straight forward way. This will become apparent in the next chapter.

The reader should be cautioned that the results obtained in this chapter, while being formally correct, have been obtained by using methods which are not rigorous. Fortunately, used in the proper framework on linear systems, the results presented are also rigorously correct. In order to treat the subject of random signals in general systems correctly, somewhat more advanced methods are necessary. The interested reader is referred in the first case to 'Notes for Chapter 6' below and in the second to the references at the end of this book.

## 6.6 Notes

### *6.6.1 The Normal Distribution*

The normal distribution was first discovered by Abraham De Moivre in 1733 as a bi-product of his investigation of the limiting form of the binomial distribution. This work was not generally known and it was not until this distribution was rediscovered by Karl Fredrich Gauss in 1809 and Pierre Simon de Laplace in 1812 that it became generally known. In fact Gauss did breach the subject briefly much earlier in work published in 1780 but the he went into depth with it in work published first in 1812. Both Gauss and Laplace were led to the normal distribution in connection with their study of the general question of the interpretation of observational errors. They are also credited with a number of applications of the normal distribution to various matters in the general theory of probability.

Laplace is responsible for the first statement of the Central Limit Theorem, though it was at first incomplete. Due to the great influence of Gauss and Laplace it was for a long period of time believed that stochastic distributions of nearly all kinds would approach a normal distribution in the limit if only a sufficiently large number of observations could be assembled. This went so far that it was supposed that any deviation of a stochastic variable from its supposed mean was regarded as an 'error'. Even though this is now known to be in error, many types of observational data (demographic, biological, astronomical and physical) yield, as far as can be determined, approximate distributions which are very close to being normal. According to Cramér, '...mathematical proof tells us that, under certain qualifying conditions, we are justified in expecting a normal distribution, while statistical experience shows us that, in fact, distributions are often approximately normal' (Cramér, 1946, p. 232).

### *6.6.2 The Wiener Process*

The study of Brownian motion was the beginning of the investigation of stochastic differential equations and the mathematical theory thus evolved is one of the most important developments in the theory of stochastic processes. The first satisfactory theories for Brownian motion were developed independently by Albert Einstein and M. von Smoluchowski at the beginning of the 1900s. Norbert Wiener and P. Lévy gave the first rigorous treatments of the process and hence the names Wiener and Wiener-Lévy are attached firmly to it. The most significant theoretical contribution is due to Wiener and dates from about 1923.

### 6.6.3 Stochastic Differential Equations

In general the problem which has to be solved when considering a system of the type considered in this text with a stochastic input is

$$\frac{dx(t)}{dt} = f(x(t), t) + g(x(t), t)\ v(t) \tag{6.327}$$

where $f$ and $g$ are nonlinear real valued functions and $v(t)$ is a random input (usually assumed to be white noise). This problem with additive white noise is a generalization of the Langevin problem. The process $v(t)$ is delta function correlated and is not integrable in the ordinary way. Thus Eq. (6.327) has no real mathematical meaning and has to be reformulated for use on the problem at hand.

Recalling that $v(t)$ is formally the derivative of a Wiener process, it can be written as

$$v(t) = \frac{dw(t)}{dt}, \tag{6.328}$$

where $w(t)$ is now a Wiener process. This means that Eq. (6.327) can be reformulated as the equation

$$dx(t) = f(x(t), t)\ dt + g(x(t), t)\ dw. \tag{6.329}$$

.

This equation only has meaning if its integral is defined,

$$x(t) - x(t_0) = \int_{t_0}^{t} f(x(\tau), \tau) d\tau + \int_{t_0}^{t} g(x(\tau), \tau)\ dw(\tau), \tag{6.330}$$

where it is assumed that proper initial conditions have be specified for the integration. The first integral can be defined as a Riemann integral but the second cannot. In the theory of stochastic processes the second Stieltjes integral is commonly defined in two ways: as a Stratonovich or as a Ito integral. Because the increment of a Wiener process, $dw$, has a magnitude which is proportional to $\sqrt{dt}$, neither of these integrals can be of an ordinary type and new integration rules have to be derived. To go further here is beyond the scope of the current treatment. The interested reader is referred to the references for a more complete development, in particular the list of references should read: Åstrøm (1970), Doob (1953), McGarthy (1974) and Soong (1973).

## 6.7 Problems

### *Problem 6.1*

Consider the problem of expressing the results of an experiment where a coin is flipped four times. To associate a stochastic variable with the experiment, a 1 will be scored for a heads and a 0 for tails. For four flips of the coin the maximum score is 4 while the minimum is 0. For a given series of experiments, the score is a stochastic variable, $X$.

a. By enumerating the possible results of the experiment, find the probability density function for the experiment and sketch it on a graph.
b. Given the probability density function above, find the corresponding probability distribution function and sketch it on a graph.
c. What is the most probable result for the score of the experiment?
d. What is the most probable value of $X^2$?
e. What is the average result for a large number of trials and what is the standard deviation?

### *Problem 6.2*

A random variable, $X$, has the probability density function,

$$f(x) = \frac{a}{x^2 + 1},$$

and is defined for all real $x$.

a. Find the value of the constant $a$.
b. Find the distribution function which corresponds to the p.d.f. above.
c. Find the probability that the stochastic variable, $Y = X$, is between 1/4 and 1.
d. Find the probability that the stochastic variable, $Y = X^2$, lies between 1/4 and 1.

### *Problem 6.3*

The probability density function for a stochastic variable, $X$, is

$$F(x) = \begin{cases} 1 - e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}.$$

a. Find the corresponding probability density function.
b. Find the probability that $X \geq 5$.
c. Find the probability that $-5 \leq X \leq 5$.

### *Problem 6.4*

In a game of darts a round target is used. By independent experiment it has been found that the probability of a dart hitting between $r$ and $r + dr$ is

$$Pr(r \geq R \geq r + dr) = a\left[1 - \left(\frac{r}{b}\right)^2\right]dr,$$

where $R$ is the distance of the hit from the center of the target, $a$ is a constant and $b$ is the radius of the target. Assume that during a game that there are no complete target misses.

a. What is probability of hitting the bull's eye (center of the target) if it has a radius $c$?
b. Is the answer completely correct? Explain.

***Problem 6.5***

Find a set of differential equations which can be used to find the running mean value and root mean square (RMS) value of an arbitrary continuous input signal. By definition, for any given initial time, $t_0$, and for $t \rightarrow \infty$,

$$v_{avg}(t) = \frac{1}{t}\int_{t_0}^{t} v(\tau)d\tau,$$

$$v_{rms}(t) = \left(\frac{1}{t}\int_{t_0}^{t} [v(\tau)]^2 d\tau\right)^{\frac{1}{2}}.$$

a. What is the meaning of these differential equations?
b. Can these equations be used for the practical determination of the mean and RMS values of a signal? Under what circumstances?
c. If the answer to b. is yes then what is (are) the advantage(s) and disadvantage(s) of doing this?
d. Can the equations above be used on a random (noise) signal?

Find the linearized equations corresponding to the differential equations above.

e. What do these equations tell about the dynamics of the 'measuring instruments' which can be constructed using the equations above?
f. On what time scales can one expect an exact answer?

***Problem 6.6***

The average of a discrete signal $u(k)$ can be calculated as the expression,

$$\langle x(k)\rangle = \frac{1}{k}\sum_{j=1}^{k} u(j),$$

while its average squared value can be calculated from

$$\langle x(k)^2\rangle = \frac{1}{k}\sum_{j=1}^{k} u(j)^2.$$

Given that the sampling time is $\Delta T$, derive recursive expression for the average and squared averaged values of the input which can be used for on line calculation of these quantities.

a. Compare the expressions found with those of Problem 6.5.
b. Will the averages calculated using the discrete recursion equations above agree with those found using the continuous differential equations of Problem 6.5?
c. What information is lost by using the discrete averaging equations rather than the continuous ones?

### *Problem 6.7*

A system is described by the differential equation,

$$\dot{x}(t) = -ax(t) + 2u(t),$$

where $u(t) \in N(0, \mu^2)$ is white noise.

a. What is the stationary variance of $x(t)$?

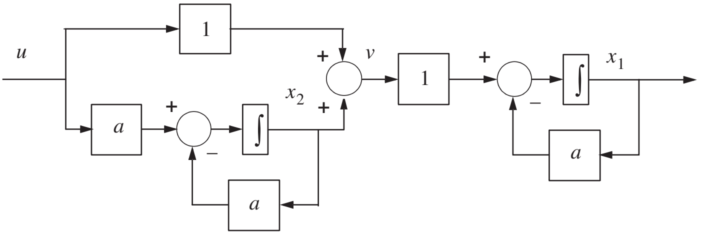
Assume now that $u(t)$ is low pass filtered white noise which has been pre-processed by the system

$$\dot{y}(t) = -8y(t) + v(t),$$
$$u(t) = y(t).$$

b. What are the state equations of the system? Note that $v(t) \in N(0, \mu^2)$
c. What is now the covariance of $x(t)$ calculated in the time domain?
d. What is the covariance of $x(t)$ calculated in the frequency domain?

### *Problem 6.8*

A block diagram of a second order system is given below.



a. What is the transfer function of the system from $u$ to $x_1$?
b. Sketch a Bode plot of the asymptotic frequency response of the system.
c. What is the D.C. gain of the network?
d. What is such a filter called?
e. Write down the state equations of the network.

Now assume that the input is Gaussian white noise such that $u(t) \in N(0, \sigma_u^2)$. Given that for a low pass filter with cutoff frequency $\omega_0$, a D.C. gain of one, driven with white noise $u(t) \in N(0, \sigma_u^2)$, the output is

$$\sigma_y^2 = \frac{\omega_0}{2}\sigma_u^2.$$

f. Calculate the size of the signal at $v$ in the diagram above using the information given.
g. Now calculate the noise at the output of the filter a the point $x_1$, using the result of the point f. above and the information given.
h. Is the same result obtained by using the Lyapunov equation for the overall system? Why is the result found using the Lyapunov equation different than that obtained using the given information?

### *Problem 6.9*

A second order system is described by a transfer function which is,

$$H(s) = \frac{X(s)}{U(s)} = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0 s + \omega_0^2},$$

where $s$ is the Laplace operator, $\zeta$ and $\omega_0$ are positive constants. This system can be represented simply in at least two different differential equation formulations.

a. What are two of these formulations? Draw block diagrams for each of these realizations.

The input, $u(t)$, to either of the system realizations in question a. is assumed to be a white noise source with the characteristics, $u(t) \in N(0, \sigma_u^2)$

b. What is the steady state covariance of the output of the two possible realizations of the system, $x(t)$? Should it be the same or different?
c. What happens when $\zeta \to 0$? Can it ever be zero?

### *Problem 6.10*

Find a system which can realize the transfer function

$$H(s) = \frac{X(s)}{U(s)} = \frac{\omega_0^2 s}{s^2 + 2\zeta\omega_0 s + \omega_0^2}.$$

a. What is such a filter called?
b. Does it have a useful application? Which?
c. Draw a Bode plot of the system's response.

Assume that the input is driven by a white noise signal which is $u(t) \in N(0, \sigma_u^2)$.

d. What is the noise signal at the output of the filter?
e. What happens when the damping parameter goes to zero? Is this in practice possible? Under what conditions is such a system used?

### *Problem 6.11*

Assume that it is given that a certain test system has as its output a stationary stochastic process, $Y(t)$, with the following covariance function:

$$R_Y(\tau) = \begin{cases} a\left(1 - \frac{|\tau|}{b}\right), & |\tau| \leq \mathrm{b} \\ 0, & \text{otherwise} \end{cases}.$$

a. Find the power spectral density matrix for the system and sketch it.
b. Can this spectral density be realized as the output of a system with finite order which is driven by a white noise source? Explain you answer.

Select the coefficients of the first order transfer function,

$$H(s) = \frac{b}{s + a},$$

such that the output power of this system for low frequencies is the same as $Y(t)$ at low frequencies. It is to be assumed that the noise input of the system is $u(t) \in N(0, 1)$.

c. What are the constants of the transfer function $H(s)$?
d. Plot the two spectra for $a = 1$ and $b = 1$ up to a frequency $\omega = 2\pi$.
e. What is the 3 dB power bandwidth for the $Y(t)$ signal (the frequency above which lies half to the total signal power)?

***Problem 6.12***

Consider the following dynamic system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -3 & 2 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t),$$

$$y = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

It is assumed that this system is driven with a zero mean white noise source with intensity, $V = 1$.

a. Derive and expression for the power spectral density matrix for this system and for its output in the stationary state.
b. Calculate the stationary variance matrix for the states of the system and find the noise power on the output of it.

Consider the quadratic expression

$$J = E\left\{ \int_{t_0}^{t_1} \mathbf{x}^T \mathbf{R}\, \mathbf{x}\; d\tau \right\}.$$

c. How can R be selected, given that it has to be symmetrical and so that

$$J = E\left\{ \int_{t_0}^{t_1} y^2(\tau)\; d\tau \right\}.$$

d. How does this integral develop for large times?

# Chapter 7
# Optimal Observers: Kalman Filters

**Abstract** This chapter has the purpose of reviewing the most important design aspects of Kalman filters as well as some of their most important properties. Heuristic derivations are given of the Kalman filter 'equations for both continuous time and discrete time dynamic systems. It is shown that the state mean values propagate according to the same observer equations as given in Chap. 4. Moreover it is shown that the state noise propagates according to the time dependent Lyapunov equation derived in Chap. 6. When measurements are made on the system this equation has to be modified with a term which expresses the decrease of uncertainty which the measurements make possible. The combination of these two results yields the main stochastic design equation for Kalman filters: the Riccati equation. Solving this equation immediately gives the optimal observer gain for a Kalman filter. Combining a Kalman filter with optimal or LQR feedback results in a very robust controller design: the LQG or Linear Quadratic Gaussian regulator.

## 7.1 Introduction

In many practical situations very few of the states or functions of the states of a dynamical system may be measured directly without error. In general this occurs either because the states or the measurements of the states (or most commonly, both) are corrupted by noise. In such cases it is reasonable to consider the problem of finding an optimal estimate of all the states of the system given noisy measurements of some or all of the other states. Here 'optimal' may mean optimal in the sense of least squares, minimum variance, or some other optimality criterion.

Briefly this can be accomplished by forcing a mathematical model of the system dynamics to follow the states of the plant or control object itself. The effects of the state noise are accounted for by effectively propagating the state noise through the same mathematical model and filtering it from the state estimates with a weight depending on the measurement noise. Recognizing that no analytical solution is actually required, the state and noise differential equations are solved recursively to find the state estimates. One of the most common such mathematical models

E. Hendricks et al., *Linear Systems Control*, DOI: 10.1007/978-3-540-78486-9_7,
© Springer-Verlag Berlin Heidelberg 2008

and its associated noise suppression algorithm is called a Kalman filter. Currently the large computational burden which this entails is often placed on a digital computer but in some cases may also be carried using analog methods. It is also possible to apply the basic Kalman filter to parameter estimation and/or nonlinear processes though the sense in which it is optimal is changed.

Kalman filters are at present used in many of the control systems which are familiar. Navigation systems for airplanes, ships and spacecraft based on such filters are very common and applications to automotive navigation systems are currently in production in several companies. Attitude control for these transport systems, including optimal estimators are also becoming common. Positioning systems for oil well drilling platforms, which require great accuracy and where cost is not a object are also a natural application area. Automotive engine controllers using nonlinear Kalman filters have recently been put into production and will become more widely used as emission restrictions for automobiles become more stringent. Thus Kalman filters are becoming more prevelent as the desired accuracy and functionality of every day products increases and the price of microprocessors and microcontrollers continues to fall.

In order to use the Kalman-Bucy algorithm effectively, its method of operation should be understood. This is particularly true when the algorithm is to be applied to discretized systems. In order to show how the original linear algorithm works, the next three sections of this chapter will analyze and describe the linear continuous and the discrete Kalman filters (CKF and DKF). The use of optimal stochastic observers together with optimal (LQR) controllers will be discussed in the final section of this chapter. Such systems are called LQG or (optimal) Linear Quadratic Gaussian control systems.

## 7.2 Continuous Kalman Filter

The system for which a continuous Kalman filter (CKF) is to be realized has the usual form except that noise sources are added to model the system disturbances. The state equation is

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{B}_v(t)\mathbf{v}_1(t) \tag{7.1}$$

and the output equation is

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{v}_2(t). \tag{7.2}$$

For a Kalman filter to exist for the system above, it is necessary that the system be completely observable (or reconstructable). Alternatively the unobservable portion of the system has to be stable.

The term $\mathbf{v}_1(t)$ models the process or state or noise: this disturbance is usually built into the control object itself. $\mathbf{v}_2(t)$ is the observation or measurement noise which is most often external to the control object. Most often the components of the state and measurement noise sources are assumed to be uncorrelated. It is in other words assumed that the joint noise process can be described by white noise with a matrix intensity,

$$\mathbf{R}_V(t_1, t_2) = E\left\{\begin{bmatrix}\mathbf{v}_1(t_1)\\ \mathbf{v}_2(t_1)\end{bmatrix}\begin{bmatrix}\mathbf{v_1}^T(t_2) & \mathbf{v}_2^T(t_2)\end{bmatrix}\right\} = \mathbf{V}(t_1)\delta(t_1 - t_2), \tag{7.3}$$

where clearly,

$$\mathbf{V}(t_1) = \begin{bmatrix}E\{\mathbf{v}_1(t_1)\mathbf{v}_1^T(t_2)\} & E\{\mathbf{v}_1(t_1)\mathbf{v}_2^T(t_2)\}\\ E\{\mathbf{v}_2(t_1)\mathbf{v}_1^T(t_2)\} & E\{\mathbf{v}_2(t_1)\mathbf{v}_2^T(t_2)\}\end{bmatrix} = \begin{bmatrix}\mathbf{V}_{11}(t_1, t_2) & \mathbf{V}_{12}(t_1, t_2)\\ \mathbf{V}_{21}(t_1, t_2) & \mathbf{V}_{22}(t_1, t_2)\end{bmatrix}. \tag{7.4}$$

If $\mathbf{V}_{12}(\ldots) = \mathbf{V}_{21}(\ldots) = 0$ then the state and measurement noise sources are uncorrelated and this will be assumed initially for the sake of simplicity. Also it will be assumed that $\mathbf{V}_{22}(\ldots) > 0$. In other words there is noise on all the measurements.

The initial conditions for the control object are

$$E\{\mathbf{x}(t_0)\} = \mathbf{x}_0, \tag{7.5}$$

$$E\{[\mathbf{x}(t_0) - \mathbf{x}_0][\mathbf{x}(t_0) - \mathbf{x}_{\mathbf{0}}]^T\} = \mathbf{Q}_0. \tag{7.6}$$

A linear, full order, optimal, stochastic observer is to be constructed which has the general form,

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{M}(t)\hat{\mathbf{x}}(t) + \mathbf{N}(t)\mathbf{u}(t) + \mathbf{L}(t)\mathbf{y}(t), \tag{7.7}$$

where $\mathbf{M}(t)$, $\mathbf{N}(t)$ and $\mathbf{L}(t)$ are to be determined. Clearly, one should choose $\mathbf{B}(t) \equiv \mathbf{N}(t)$ immediately on the basis of the observer derived in Chap. 4. The basic requirements of this observer are that 1. the state estimates are unbiased and 2. the covariance of the state estimate error is a minimum.

A method of dealing with this problem is to focus on the error equation as in Chap. 4, where the error is defined by the difference between the states and the state estimates,

$$\mathbf{e}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t). \tag{7.8}$$

If the observer equation is subtracted from the state equation including noise sources, assuming that $\mathbf{B}(t) = \mathbf{N}(t)$, one obtains

$$\begin{aligned}\dot{\mathbf{e}}(t) &= \dot{\mathbf{x}}(t) - \dot{\hat{\mathbf{x}}}(t)\\ &= \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}_v(t)\mathbf{v}_1(t) - \mathbf{M}(t)\hat{\mathbf{x}}(t) - \mathbf{L}(t)\mathbf{C}(t)\mathbf{x}(t) - \mathbf{L}(t)\mathbf{v}_2(t)\\ &= [\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]\mathbf{x}(t) - \mathbf{M}(t)\hat{\mathbf{x}}(t) + \mathbf{B}_v(t)\mathbf{v}_1(t) - \mathbf{L}(t)\mathbf{v}_2(t).\end{aligned} \tag{7.9}$$

Taking the mean value of both sides of this equation, since the order of expectation and differentiation can be reversed (they are both linear operations) and the mean values of the noise sources $\mathbf{v}_1(t)$ and $\mathbf{v}_2(t)$ are zero, it is clear that for an unbiased error that it has to be true that

$$\mathbf{M}(t) = \mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t). \tag{7.10}$$

Thus the observer of equation (7.7) has to have the form,

$$\dot{\hat{\mathbf{x}}}(t) = [\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]\hat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{L}(t)\mathbf{y}(t),$$

or

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{A}(t)\hat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{L}(t)[\mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t)]. \tag{7.11}$$

This is the equation for the optimal observer and it is the same as for a deterministic observer. It suggests that the state estimate is the mean value of the state,

$$\hat{\mathbf{x}}(t) = E\{\mathbf{x}(t)\}. \tag{7.12}$$

The initial condition for the estimation error is $\mathbf{e}(t_0) = \mathbf{x}(t_0) - \hat{\mathbf{x}}(t_0) = \mathbf{e}_0$.

Using the value for $\mathbf{M}(t)$ in Eq. (7.10), the error differential equation becomes

$$\dot{\mathbf{e}}(t) = [\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]\mathbf{e}(t) + [\mathbf{B}_v(t) - \mathbf{L}(t)]\begin{bmatrix}\mathbf{v}_1(t)\\ \mathbf{v}_2(t)\end{bmatrix}. \tag{7.13}$$

Equation (7.13) has the same form as the system analyzed with the time dependent Lyapunov equation in Chap. 6, if the overall noise input matrix is redefined to be $\mathbf{B}'_v(t) = [\mathbf{B}_v(t) - \mathbf{L}(t)]$. Thus it is possible to write down immediately the equation which describes the time development of the variance of the reconstruction error:

$$\begin{aligned}\dot{\mathbf{Q}}(t) =&[\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]\mathbf{Q}(t) + \mathbf{Q}(t)[\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]^T\\ &+ \mathbf{B}_v(t)\mathbf{V}_1(t)\mathbf{B}_v^T(t) + \mathbf{L}(t)\mathbf{V}_2(t)\mathbf{L}^T(t).\end{aligned} \tag{7.14}$$

The initial condition for this differential equation is $\mathbf{Q}(t_0) = \mathbf{Q}_0$.

Using the solution to the LQR problem of Chap. 5, it is possible to find the optimal value of $\mathbf{L}(t)$ which minimizes the estimation error covariance matrix from Eq. (7.14). In Appendix C it is shown that this gain is given by

$$\mathbf{L}(t) = \mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t). \tag{7.15}$$

This is the Kalman gain matrix. As $\mathbf{L}(t)$ is time dependent, the filter is in fact optimal in time at every instant.

Using the result above for the Kalman gain, a more compact and convenient form of equation (7.14) can be obtained which is the optimal observer Riccati equation:

$$\begin{aligned}\dot{\mathbf{Q}}(t) = {} & \mathbf{A}(t)\mathbf{Q}(t) + \mathbf{Q}(t)\mathbf{A}^T(t) \\ & + \mathbf{B}_v(t)\mathbf{V}_1(t)\mathbf{B}_v^T(t) - \mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t)\mathbf{C}(t)\mathbf{Q}(t).\end{aligned} \tag{7.16}$$

Equations (7.11) and (7.16) are the solution to the optimal stochastic observer or Kalman filtering problem.

The initial conditions for the state estimate time update and the covariance time/measurement update are the expected start values of the state vector, $\mathbf{x}(t_0)$,

$$E\{\mathbf{x}(t_0)\} = \hat{\mathbf{x}}(t_0), \tag{7.17}$$

and the expected start value of the error covariance matrix, $\mathbf{Q}(t_0)$,

$$E\Big\{[\mathbf{x}(t_0) - \hat{\mathbf{x}}(t_0)][\mathbf{x}(t_0) - \hat{\mathbf{x}}(t_0)]^T\Big\} = \mathbf{Q}(t_0). \tag{7.18}$$

This is effectively the information available to the filter about the system before it is started up.

The term $\mathbf{A}(t)\mathbf{Q}(t) + \mathbf{Q}(t)\mathbf{A}^T(t)$ above in Eq. (7.16) represents the internal response of the system due to the state noise input $\mathbf{B}_v(t)\mathbf{V}_1(t)\mathbf{B}_v^T(t)$ (which increases the statistical uncertainty because of the state noise). These three terms together with the time derivative are the time dependent Lyapunov equation for the system. The negative term, $-\mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^-1(t)\mathbf{C}(t)\mathbf{Q}(t)$, represents the decrease the overall uncertainty as a result of the measurements. One can think of the solving of the Riccati equation as a method of using the apriori information available about the state and measurement noise to find the error covariance of the system at a given time $t$.

Thus a Kalman filter is constructed in such a way that it keeps track in time of the two main variables which describe the Gaussian distributed stochastic state vector: its mean value, $\hat{\mathbf{x}}(t)$, and its error covariance matrix, $\mathbf{Q}(t)$. Notice also that the measurement noise in Eq. (7.13) goes directly into the reconstruction error and is multiplied by the observer gain, $\mathbf{L}(t)$. This implies that a large gain will lead to a large measurement noise sensitivity.

The Kalman filter as derived above can be shown to be the minimum mean square linear estimator and a better linear estimator cannot be made using any techniques. It can also be shown that under certain additional assumptions (Gaussian initial conditions, state and measurement noise) the Kalman filter is the best minimum mean square estimator of all estimators, linear or nonlinear.

The gain in the Kalman filter is obviously mainly dependent on the aprori noise assumptions embodied in $\mathbf{V}_1(t)$ and $\mathbf{V}_2(t)$. This means that in principle the necessary gain can be computed off line and stored for later use. This makes it unnecessary to calculate the Riccati equation on line though in fact this is quite possible. This is technically possible because in contrast to the optimal regulator or LQR Riccati equation, the observer equation is solved forward in time: it has initial conditions rather than terminal constraints. Moreover $\mathbf{Q}(t)$ is symmetric (containing $n(n+1)/2$ independent elements) and this is some aid in saving computation. Not having to solve the Riccati equation on line however reduces the computational burden significantly.

Control objects which are characterized by constant (or approximately constant) process and measurement noise intensities lend themselves to the use of a constant gain matrix $\mathbf{L}(t)$. This is because for such systems, $\mathbf{Q}(t)$ goes to a constant value for large times which is independent of the initial conditions. In the time invariant case with constant system matrices and constant noise intensities, a constant solution can be obtained by setting the time derivative in Eq. (7.16) equal to zero. This yields the algebraic or time independent Riccati equation. The solution to this equation gives the error variance of a Kalman filter which is optimal for steady state operation,

$$\mathbf{0} = \mathbf{AQ} + \mathbf{QA}^T + \mathbf{B}_v\mathbf{V}_1\mathbf{B}_v^T - \mathbf{QC}^T\mathbf{V}_2^{-1}\mathbf{CQ}, \tag{7.19}$$

and from which a constant Kalman gain is obtained which is

$$\mathbf{L} = \mathbf{QC}^T\mathbf{V}_2^{-1}. \tag{7.20}$$

Under midely restrictive conditions equation (7.19) provides a gain which makes the Kalman filter asymptotically stable with $\mathbf{A}_\mathbf{L} = \mathbf{A} - \mathbf{LC}$. Such an observer is an optimal steady state observer. This observer is often used instead of a time dependent one even though its performance is less than optimal – it is only optimal in the sense that its error covariance is minimum only for large times (or in the steady state) with respect to all other time invariant observers at the selected operating point.

### 7.2.1 Block Diagram of a CKF

A block diagram of a continuous Kalman filter can be found on Fig. 7.1 while a summary of the equations which characterize it and the process to which it is to be applied are given in Table 7.1. The process or system to be filtered must be

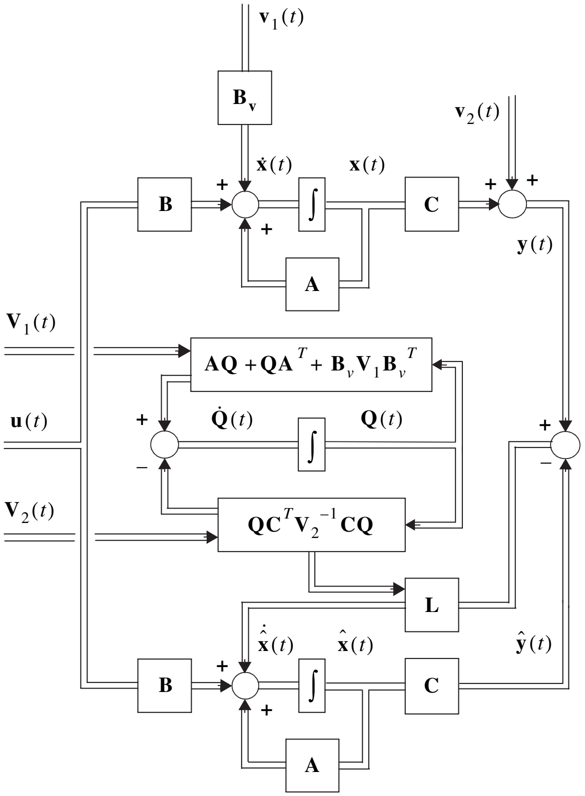


**Fig. 7.1** Block diagram of a conventional continuous Kalman filter. The double lines indicate that vector quantities are being connected

linear or linearized. It is characterized by the dynamic, control, and output matrices: $\mathbf{A}(t)$, $\mathbf{B}(t)$ and $\mathbf{C}(t)$ respectively. In addition to the control input $\mathbf{u}(t)$, the system is driven by two vector, zero mean, normal, and independent white noise sources. These are the state noise, $\mathbf{v}_1(t)$, and the measurement noise, $\mathbf{v}_2(t)$, with covariance matrices $\mathbf{V}_1(t)$ and $\mathbf{V}_2(t)$ respectively. These matrices are assumed to be known for the process to be filtered. The equations given assume that the different state noise sources may be 'mixed' (to create correlated state noise) and possibly scaled by the input gain matrix $\mathbf{B}_v(t)$. The actual process and measurement descriptions are given in the first row of Table 7.1.

The Kalman filter itself is shown in the middle and bottom of Fig. 7.1 and has the same form as a Luenbuerger observer with a gain $\mathbf{L}(t)$. It has two main elements: the state estimate time/ measurement update and the covariance time/ measurement update.

The state estimate time update is simply an exact copy of the dynamics of the process (as well as they are known) fed with the same control input as the system itself. Such an observer is called an identity ovserver. If one imagines that the process model has no inputs but $\mathbf{u}(t)$ then its states will be the zeroth order estimates of the states of the process itself.

**Table 7.1** Summary of the Continuous and Discrete Kalman Filter Equations (Linear Systems Only)

| | Continuous Kalman Filter | Discrete Kalman Filter (open form) |
|---|---|---|
| Control Object | $\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{B}_v(t)\mathbf{v}_1(t)$ | $\mathbf{x}(k+1) = \mathbf{F}(k)\mathbf{x}(k) + \mathbf{G}(k)\mathbf{u}(k) + \mathbf{G}_v(k)\mathbf{v}_1(k)$ |
| Measurement Model | $\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{v}_2(t)$ | $\mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k) + \mathbf{v}_2(k)$ |
| State Estimate Time Update | $\dot{\hat{\mathbf{x}}}(t) = \mathbf{A}(t)\hat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{L}(t)[\mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t)]$ | $\hat{\mathbf{x}}(k)^- = \mathbf{F}(k-1)\mathbf{x}(k-1)^+ + \mathbf{G}(k-1)\mathbf{u}(k-1)$ |
| Covariance Time Update | $\dot{\mathbf{Q}}(t) = \mathbf{A}(t)\mathbf{Q}(t) + \mathbf{Q}(t)\mathbf{A}^T(t) + \mathbf{B}_v(t)\mathbf{V}_1(t)\mathbf{B}_v^T(t)$<br>$- \mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t)\mathbf{C}(t)\mathbf{Q}(t)$ | $\mathbf{Q}(k)^- = \mathbf{F}(k-1)\mathbf{Q}(k-1)^+\mathbf{F}^T(k-1)$<br>$+ \mathbf{G}(k-1)\mathbf{V}_1(k-1)\mathbf{G}^T(k-1)$ |
| Initial Conditions for Time Update above | $E\{\mathbf{x}(0)\} = \hat{\mathbf{x}}(0)$<br>$E\{(\mathbf{x}(0) - \hat{\mathbf{x}}(0))(\mathbf{x}(0) - \hat{\mathbf{x}}(0))^T\} = \mathbf{Q}(0)$ | $E\{\mathbf{x}(0)\} = \hat{\mathbf{x}}(0)$<br>$E\{(\mathbf{x}(0) - \hat{\mathbf{x}}(0))(\mathbf{x}(0) - \hat{\mathbf{x}}(0))^T\} = \mathbf{Q}(0)$ |
| Kalman Gain Matrix | $\mathbf{L}(t) = \mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t)$ | $\mathbf{L}(k) = \mathbf{Q}(k)^-\mathbf{C}^T(k)[\mathbf{C}(k)\mathbf{Q}(k)^-\mathbf{C}^T(k) + \mathbf{V}_2]^{-1}$ |
| State Estimate Measurement Update | Included in time update above | $\hat{\mathbf{x}}(k)^+ = \hat{\mathbf{x}}(k)^- + \mathbf{L}(k)[\mathbf{y}(k) - \mathbf{C}(k)\hat{\mathbf{x}}(k)^-]$ |
| Covariance Measurement Update | Included in time update above | $\mathbf{Q}(k)^+ = [\mathbf{I} - \mathbf{L}(t)\mathbf{C}(t))]\mathbf{Q}(k)^-$ |
| Other Assumptions | $E\{\mathbf{v}_1(t)\mathbf{v}_2^T(\tau)\} = 0,$ for all t, τ<br>$\mathbf{v}_1(t) \in N(0, \mathbf{V}_1),\ \mathbf{v}_2(\mathbf{t}) \in N(0, \mathbf{V}_2)$ | $E\{\mathbf{v}_1(j)\mathbf{v}_2^T(k)\} = 0,$ for all j, k<br>$\mathbf{v}_1(k) \in N(0, \mathbf{V}_1),\ \mathbf{v}_2(k) \in N(0, \mathbf{V}_2)$ |

The state estimate measurement update involves first finding the output innovations or residuals $\mathbf{e}(t) = \mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t) = \mathbf{y}(t) - \hat{\mathbf{y}}(t)$ which are the differences between the outputs of the process, $\mathbf{y}(t)$, and the estimates of the outputs, $\hat{\mathbf{y}}(t)$. The second step is to multiply the innovations with the weighting matrix, $\mathbf{L}(t)$, and to add the results to the summing junction of the process model.

The matrix $\mathbf{L}(t)$ is again the Kalman gain matrix and its elements are roughly the ratios between the statistical measures of the uncertainty in the state estimates and the uncertainties in the measurements. If the elements of $\mathbf{V}_1(t)$ are small (corresponding to little state noise) while $\mathbf{V}_2(t)$'s are large (great measurement noise) then this means that the measurements cannot be trusted and $\mathbf{L}(t)$ will be small. Thus the residuals will only give small corrections in the state estimates and the filter's output will be dependent mostly on the process model. Conversely if the elements of $\mathbf{V}_1(t)$ are large while $\mathbf{V}_2(t)$'s matrix elements are small then there is a great deal of state noise but the measurements can be trusted and $\mathbf{L}(t)$ 's elements will be large. The residuals will then have a great influence on the state estimates.

In order to find $\mathbf{L}(t)$, the time development of the assumed state noise must be projected and the effects of the measurement noise taken into account using the Riccati equation, Eq. (7.16). This is the function of the block in the center of Fig. 7.1 and it is of a completely different nature than the state estimate time/ measurement update component of the filter. It has no inputs from the system itself but only 'internal' inputs $\mathbf{V}_1(t)$ and $\mathbf{V}_2(t)$, which again are assumed to be known. These are inserted into the matrix Riccati equation as shown.

One of Kalman's contribution was to show that the gain which yields the minimum variance and least square error estimates of the states is given by $\mathbf{L}(t) = \mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t)$. Thus another way to look at the filter of Fig. 7.1 is a Luenberger observer with a stochastically optimized gain, the Kalman gain, $\mathbf{L}(t)$.

***Example 7.1.* Time Dependent Kalman Filter**

This example concerns a Kalman filter designed around a first order low pass filter. The system equations are

$$\dot{x}(t) = -ax(t) + bu(t) + v_1(t), \tag{7.21}$$

$$y(t) = x(t) + v_2(t), \tag{7.22}$$

where $a$ and $b$ are constants and $v_1(t) \in N(0, V_1)$ and $v_2(t) \in N(0, V_2)$ are white noise sources such that

$$E\{v_1(t)\} = 0, R_{v_1}(t_1 - t_2) = \sigma_{v_1}^2 \delta(t_1 - t_2), \tag{7.23}$$

$$E\{v_2(t)\} = 0, R_{v_2}(t_1 - t_2) = \sigma_{v_2}^2 \delta(t_1 - t_2). \tag{7.24}$$

The initial conditions for the system are $E\{x(0)\} = 0$ and $E\{x^2(0)\} = q_0$. The Riccati equation for this system can be written as

$$\dot{q} = -2aq + \sigma_{v_1}^2 - \left(\frac{1}{\sigma_{v_2}^2}\right)q^2. \tag{7.25}$$

In this simple case the Riccati equation can be directly integrated to find $q(t)$:

$$\begin{gathered} q(t) = q_1 + \frac{q_1 + q_2}{[(q_0 + q_2)/(q_0 + q_1)]e^{2\alpha t - 1}}, \\ \Rightarrow q(t) \to q_1, \quad \text{for} \quad t \to \infty, \end{gathered} \tag{7.26}$$

where $\alpha = \sqrt{a^2 + (\sigma_{v_1}^2/\sigma_{v_2}^2)}$, $q_1 = \sigma_{v_2}^2(\alpha - a)$, $q_2 = \sigma_{v_2}^2(\alpha - a)$. The optimal time dependent observer for this system is then

$$\dot{\hat{x}} = -a\hat{x} + bu + (q(t)/\sigma_{v_2}^2)(y - \hat{x}), \hat{x}(0) = 0, \tag{7.27}$$

$$\to \dot{\hat{x}} = -a\hat{x} + bu + (q_1/\sigma_{v_2}^2)(y - \hat{x}), \quad \text{for} \quad t \to \infty, \tag{7.28}$$

where the Kalman gain is obviously

$$l(t) = L(t) = Q(t)C^T(t)V_2^{-1}(t) = q(t)/\sigma_{v_2}^2. \tag{7.29}$$

Notice that for large $t$ the Kalman gain is:

$$l(t) = \frac{q_1}{\sigma_{v_2}^2} = \sqrt{a^2 + (\sigma_{v_1}^2/\sigma_{v_2}^2)} - a \approx \frac{\sigma_{v_1}}{\sigma_{v_2}}, \quad \text{for} \quad t \to \infty. \tag{7.30}$$

In other words the Kalman gain is approximately proportional to the square root of the ratio of the state noise intensity to the measurement noise intensity. Note that this result is not dependent on the deterministic input signal.

The response of the Kalman filter above has been simulated for the system parameters $a = b = 1, \sigma_{v_1} = 0.5, \sigma_{v_2} = 0.1$. The results are shown on Fig. 7.2. The response of the system to the square pulse input of the first box in the figure is shown on the second figure. It is the typical response of a first order system corrupted with white state and measurement noise. The output of the system with measurement noise is superimposed on the actual state of the system. Also shown for reference is the uncertainty in the state estimate, $x(t) \pm \sqrt{q(t)}$. The third figure shows the state and the state estimate (which is smoother than the state). Even in the presence of a large state noise component, the state estimate is close to the actual state of the system.

It should be noted that the remarks made in the last chapter with respect to the accuracy of simulations using noise and deterministic differential equation integration algorithms also apply to systems which include Kalman filters here. Only results which are approximately correct can be expected. This is true here and Fig. 7.2 exemplifies this in the error covariances calculated: they are not exactly what may be predicted theoretically.

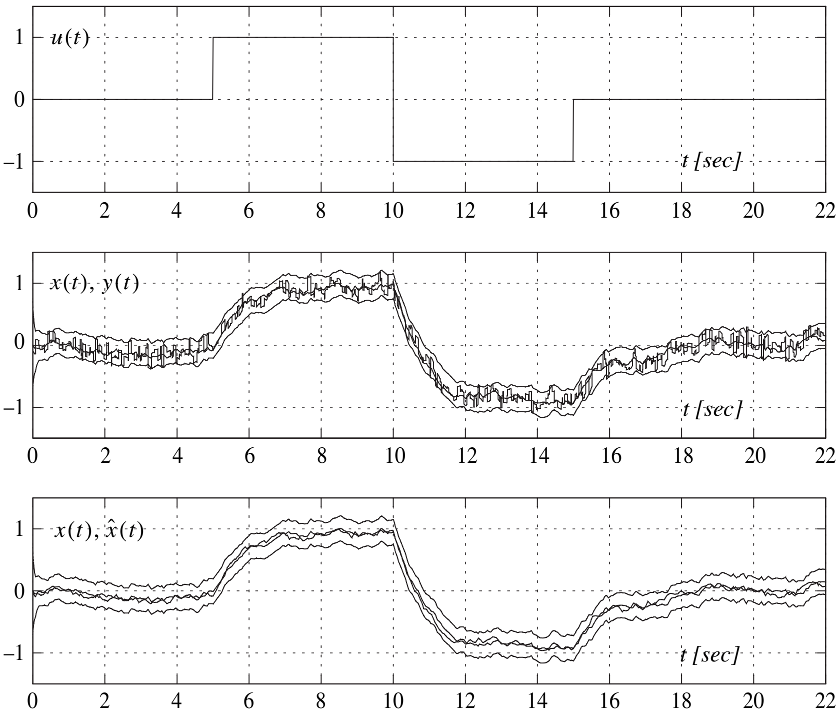


**Fig. 7.2** Simulation of the response of a Kalman filter for the first order system of Example 7.1. The three figures show the input signal, the state and output of the system and the state and state estimate of the system. The smoother trace in the last two figures is the state estimate

Because of the time dependence of the error convariance, the Kalman gain is time dependent as show in Eq. (7.29). In fact this gain is shown in Fig. 7.3 for the same simulation shown in Fig. 7.2. It should be noticed that the Kalman gain rapidly reaches a final value which is close to a constant which is for this example,

$$l(t \to \infty) = \sqrt{1 + (0.5/0.1)^2} - 1 = 4.099, \tag{7.31}$$

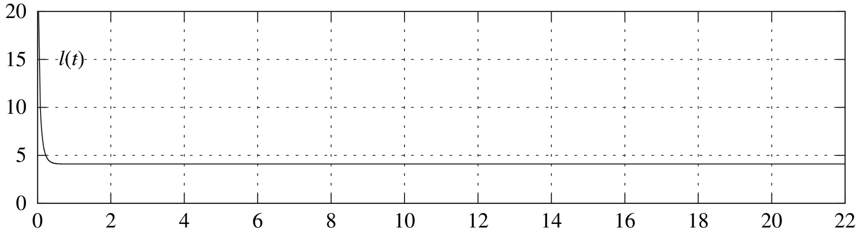


**Fig. 7.3** Time dependent Kalman gain for the Kalman filter of Example 7.2

as predicted by Eq. (7.30). That this commonly is the case is often used to justify the use of a constant Kalman gain even though it is in principle time varying. This implies that often the Kalman gain for a given set of noise sources can be calculated off line, thus simplifying the use of the filter. The simulation was made with SIMNON, Elmquist (1975). □

***Example 7.2.*** **Estimation of the States of a Hydraulic Servo**

This example is based on the hydraulic servo cylinder in Example 3.26. What is desired is to construct a Kalman filter which can be used to estimate the state variables of the cylinder given noisy measurements of the velocity of the cylinder in the stationary state.

In Example 3.26 it was shown that the state equations for the cylinder can be written

$$\begin{aligned}\dot{v}(t) &= \frac{1}{M}(-C_f v + A_c(p_1 - p_2) + f),\\ \dot{p}_1(t) &= \frac{\beta}{V}(-A_c v - C_l(p_1 - p_2) + q),\\ \dot{p}_2(t) &= \frac{\beta}{V}(A_c v + C_l(p_1 - p_2) + q).\end{aligned} \tag{7.32}$$

Here the first state equation in the original model has been ignored and the orginal state and input variables have been used.

In real estimation and control problems it is always advantageous to reduce the order of the system if possible. Here the order of the system can be reduced by subtracting the last equation from the second in Eq. (7.32). Moreover the leakage terms are small and can be ignored. Thus the system reduces to

$$\begin{aligned}\dot{v}(t) &= -\frac{C_f}{M}v + \frac{A_c}{M}\Delta p,\\ \dot{\Delta p}(t) &= -\frac{2\beta A_c}{V}v + \frac{2\beta}{V}q.\end{aligned} \tag{7.33}$$

Often hydraulic cylinders are very poorly damped so that the damping term, $C_f/M$, is small. The state equation (7.33) can thus be reformulated as

$$\begin{aligned}\dot{x}_1 &= x_2,\\ \dot{x}_2 &= -\omega_0^2 x_1 + \bar{\omega}_0^2 u,\end{aligned} \tag{7.34}$$

where $u = q$, $x_1 = v$, $x_2 = (A_c/M)\Delta p$, $\omega_0^2 = 2\beta A_c^2/(VM)$ and $\bar{\omega}_0^2 = 2\beta A_c/(VM)$. These equations have the same form as those for an undamped, forced harmonic oscillator, see Example 4.4. The measurement which is to be made is

$$y(t) = [1 \quad 0]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{7.35}$$

The state noise sources will be assumed to be zero mean white noise such that $v_{111}(t) \in N(0, V_{111})$ on $x_1$ and $v_{122}(t) \in N(0, V_{122})$ on $x_2$. The measurement noise is assumed to be of the same type and is given by $v_2(t) \in N(0, V_2)$. The Riccati equation which must be solved is

$$\mathbf{0} = \mathbf{V}_1 - \mathbf{Q}\mathbf{C}^T\mathbf{V}_2^{-1}\mathbf{C}\mathbf{Q} + \mathbf{A}\mathbf{Q} + \mathbf{Q}\mathbf{A}^T \qquad \text{or}$$

$$\begin{aligned}\mathbf{0} = {} & \begin{bmatrix} V_{111} & 0 \\ 0 & V_{122} \end{bmatrix} - \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\frac{1}{V_2}[1 \quad 0]\begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix} \\ & + \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{bmatrix}\begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix} + \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix}\begin{bmatrix} 0 & -\omega_0^2 \\ 1 & 0 \end{bmatrix}.\end{aligned} \tag{7.36}$$

This matrix equation gives three quadratic equations which must be solved simultaneously to find the matrix elements of $\mathbf{Q}$ and hence the Kalman gain matrix:

$$\begin{aligned} 0 &= V_{111} - \frac{q_{11}^2}{V_2} + 2q_{12}, \\ 0 &= 0 - \frac{q_{11}q_{12}}{V_{12}} - \omega_0^2 q_{11} + q_{22}, \\ 0 &= V_{122} - \frac{q_{12}^2}{V_2} - 2\omega_0^2 q_{12}. \end{aligned} \tag{7.37}$$

The solution of these equations is

$$\begin{aligned} q_{12} &= -\omega_0^2 V_2 \pm \sqrt{\omega_0^4 V_2^2 + V_{122}V_2}, \\ q_{11} &= \sqrt{V_2(V_{111} + 2q_{12})}, \\ q_{22} &= \omega_0^2 q_{11} + \frac{q_{11}q_{12}}{V_2}, \end{aligned} \tag{7.38}$$

where the positive sign has to be selected in the first equation in order for $\mathbf{Q}$ to be positive definite. The Kalman gain can then be calculated to be

$$\mathbf{L} = \mathbf{Q}\mathbf{C}^T\mathbf{V}_2^{-1} = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\frac{1}{V_2} = \frac{1}{V_2}[q_{11} \quad q_{12}] \tag{7.39}$$

❐

***Example 7.3*. Estimation of a Constant**

In many cases Kalman filters are not only used to estimate states but also constants or parameters in the model of the plant which is of interest. Such applications are becoming increasing common because of the desire to use less time adjusting control systems for different examples of the same product. Such adjustments are costly and cannot be maintained during product ageing and wear. Moreover the more complex algorithms necessary are easily within the performance envelope of inexpensive modern microprocessors or microcontrollers.

Let the state $x(t)$ represent the constant. A constant can be represented as the initial condition of an integrator with no input. This constant is then given by the state equation,

$$\dot{x}(t) = 0, \tag{7.40}$$

with the initial condition $x(0) = \langle x_0 \rangle$, which has a variance $q(0) = q_0$. Assuming a measurement of the variable is made with observation noise, $v_2(t)$, the measurement model is

$$y(t) = x(t) + v_2(t), \tag{7.41}$$

where $v_2(t)$ is white noise with an intensity, $V_2$, which is constant.

An observer for the system can be constructed by appending an innovation based on the measurement to the state equation (7.40),

$$\dot{\hat{v}}(t) = l(t)[y(t) - \hat{x}(t)], \tag{7.42}$$

where the initial condition is $\hat{x}(0) = \langle x_0 \rangle$ and $l(t)$ is the Kalman gain. From the system model this gain is determined as

$$l(t) = \frac{q(t)}{V_2}. \tag{7.43}$$

$q(t)$has to be found by solving the Riccati equation,

$$\dot{q}(t) = -\frac{q^2(t)}{V_2}, \tag{7.44}$$

with the initial condition $q(0) = q_0$.

This simple equation can be integrated directly and the solution is

$$q(t) = \frac{q_0 V_2}{V_2 + q_0 t}, \tag{7.45}$$

for $t \geq 0$. From this the Kalman gain is

$$l(t) = \frac{q_0}{V_2 + q_0 t}, \tag{7.46}$$

for $t \geq 0$. It is interesting to note that as $t$ becomes large, the error variance, $q(t)$ goes to zero as does the Kalman gain. Thus for large times the estimated constant or parameter becomes constant and nothing in the algorithm can change it, not even new and possibly significant measurements. ❐

***Example 7.4*. Estimation of a Slowly Varying Constant**

It is possible to rework the observer above so that it can estimate a constant effectively over a longer time horizon. This is done by using a slightly more complex model for the dynamics of the constant. To include the possibly of long term changes in the value of the constant, it can be modelled as an integrated white noise process, i. e., a Wiener process. The model then becomes

$$\dot{x}(t) = v_1(t), \tag{7.47}$$

$$y(t) = x(t) + v_2(t), \tag{7.48}$$

with $v_1(t)$ and $v_2(t)$ white noise sources with zero mean and intensities $V_1$ and $V_2$ respectively. The Riccati equation for this system is

$$\dot{q}(t) = V_1 - \frac{q^2}{V_2}, \tag{7.49}$$

with the initial condition $q(0) = q_0$. Equation (7.48) can be integrated directly and yields for $q(t)$ the expression

$$q(t) = \sqrt{V_1 V_2} \tanh\left(\sqrt{\frac{V_1}{V_2}} t + \tanh\left(\frac{q_0}{\sqrt{V_1 V_2}}\right)\right). \tag{7.50}$$

The time dependent Kalman gain is of course given by

$$l(t) = \frac{q(t)}{V_2} \rightarrow \sqrt{\frac{V_1}{V_2}}, \quad \text{for} \quad t \rightarrow \infty. \tag{7.51}$$

The observer for this system can be written down immediately as

$$\dot{\hat{x}}(t) = l(t)[y(t) - \hat{x}(t)]. \tag{7.52}$$

It is interesting to look at the frequency response of the observer when it attains steady state operating conditions. The transfer function from the filter 'input', $y(t)$, to the filter output, the state estimate, can be found by Laplace transforming equation (7.52) and solving for the output divided by the 'input'. This yields

$$H(s) = \frac{\sqrt{V_1/V_2}}{s + \sqrt{V_1/V_2}}. \tag{7.53}$$

This is a first order low pass filter with unity gain and a cut-off frequency which is $\sqrt{V_1/V_2}$. The filter is apparently attempting to separate the signal (information about the constant) from the background white noise with a first order filter which is the best it can manage, given the system order. ❒

In the examples in this section analytic solutions for the Kalman gain have been obtained as far as is possible. In general, if this is possible, it gives the most useful form for the Kalman gain as one can see immediately the effect of changing the nature of the noise source and of changing the system parameters. It is also a valuable way see the sensitivity of the filter to possible modelling errors and to check the correctness of the solution which has been found. Even when it is not feasible to find exact analytic solutions, perturbation or other forms of approximate analytic solutions can often be obtained. The alternative is to obtain numerical solutions and to plot these as a function of the parameter changes or other effects which one is investigating. This gives an idea of how small changes affect the system under study close to a given operating point. Unfortunately this gives a very poor feeling for the overall behavior of the filter under very different operating conditions, for example during initialization, significant operating point changes or in the presence of outliers (large, but statistically insignificant noise bursts or pulses).

## 7.3 Innovation Process

One of the most interesting and important characteristics of a Kalman filter is the nature of its innovations or residuals: the difference between the measured output(s) and the estimate(s) of the output(s). This quantity is defined by the equation,

$$\mathbf{i}_n(t) = \mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t), \tag{7.54}$$

where the hat indicates the estimate of the state $\mathbf{x}(t)$. This may be thought of as the bearer of the extra detailed information on the state of the observer (apart from that in the observer model itself) carried back to the observer model input. The purpose of this section is to show that under ideal circumstances, the innovations process is white noise which is the same size as the measurement noise source. In ideal conditions (accurate model), the state noise source does not contribute to the innovations, a somewhat surprizing conclusion.

In this section only the case of uncorrelated, zero mean, white state noise, $\mathbf{v}_1(t)$ (intensity $\mathbf{V}_1(t)$) and measurement noise, $\mathbf{v}_2(t)$ (intensity $\mathbf{V}_2(t)$) will be considered. The method of attack will be to calculate the covariance matrix of the noise source associated with the innovations in Eq. (7.106). The estimation error will as usual be defined by

$$\mathbf{e}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t). \tag{7.55}$$

If the integral of the innovations is denoted by $\mathbf{n}(t)$ then,

$$\dot{\mathbf{n}}(t) = \mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t) = \mathbf{C}(t)\mathbf{e}(t), \tag{7.56}$$

with the initial condition $\mathbf{n}(t_0) = 0$. Using the same considerations as used to derive equation (7.13), the following matrix equation can be written down for the errors and innovations:

$$\begin{bmatrix} \dot{\mathbf{e}}(t) \\ \dot{\mathbf{n}}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t) & 0 \\ \mathbf{C}(t) & 0 \end{bmatrix} \begin{bmatrix} \mathbf{e}(t) \\ \mathbf{n}(t) \end{bmatrix} + \begin{bmatrix} \mathbf{I} & -\mathbf{L}(t) \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1(t) \\ \mathbf{v}_2(t) \end{bmatrix}, \tag{7.57}$$

where $\mathbf{L}(t)$ is the Kalman gain matrix. This is a set of linear matrix differential equations driven by a noise source. Such a system has a noise output which can be calculated using the time dependent Lyapunov equation, of Sect. 6.4.3. The result is

$$\begin{aligned} \dot{\mathbf{Q}}(t) =& \begin{bmatrix} \mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t) & 0 \\ \mathbf{C}(t) & 0 \end{bmatrix} \mathbf{Q}(t) + \mathbf{Q}(t) \begin{bmatrix} \mathbf{A}^T(t) - \mathbf{C}^T(t)\mathbf{L}^T(t) & \mathbf{C}^T(t) \\ 0 & 0 \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{I} & -\mathbf{L}(t) \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1(t) & 0 \\ 0 & \mathbf{V}_2(t) \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ -\mathbf{L}^T(t) & \mathbf{I} \end{bmatrix}. \end{aligned} \tag{7.58}$$

The initial condition for the integration of this equation is $\mathbf{Q}(t_0) = \text{diag}[\mathbf{Q}_0\ \mathbf{0}]$. Choosing a block structure for the system corresponding to the four obvious blocks in Eq. (7.58) the following structure for $\mathbf{Q}(t)$ is suggested

$$\mathbf{Q}(t) = \begin{bmatrix} \mathbf{Q}_{11}(t) & \mathbf{Q}_{12}(t) \\ \mathbf{Q}_{12}^T(t) & \mathbf{Q}_{22}(t) \end{bmatrix}. \tag{7.59}$$

Using this block structure and carrying through the multiplication indicated in Eq. (7.59), the following subsiduary matrix differential equations can be derived:

$$\begin{aligned} \dot{\mathbf{Q}}_{11}(t) =& [\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]\mathbf{Q}_{11}(t) + \mathbf{Q}_{11}(t)[\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]^T \\ &+ \mathbf{V}_1(t) + \mathbf{L}(t)\mathbf{V}_2(t)\mathbf{L}^T(t), \end{aligned} \tag{7.60}$$

$$\dot{\mathbf{Q}}_{12}(t) = \mathbf{C}(t)\mathbf{Q}_{11}(t) + \mathbf{Q}_{12}(t)[\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]^T - \mathbf{V}_2(t)\mathbf{L}^T(t), \tag{7.61}$$

$$\dot{\mathbf{Q}}_{22}(t) = \mathbf{C}(t)\mathbf{Q}_{12}^T(t) + Q_{12}(t)\mathbf{C}^T(t) + \mathbf{V}_2(t). \tag{7.62}$$

The initial conditions for the integration of these equations are $\mathbf{Q}_{11}(t_0) = \mathbf{Q}_0$, $\mathbf{Q}_{12}(t_0) = \mathbf{0}$ and $\mathbf{Q}_{22}(t_0) = \mathbf{0}$ respectively.

The first equation is for the Kalman filter itself. Substituting for the Kalman gain in Eq. (7.60) from $\mathbf{L}(t)$ in equation (7.15), the standard Riccati equation (7.16) for a Kalman filter is obtained. Thus the first equation is simply the error covariance of the estimation error.

If the Kalman gain in Eq. (7.15) is used in Eq. (7.61) then it is reduced to a homogeneous equation:

$$\dot{\mathbf{Q}}_{12}(t) = \mathbf{C}(t)\mathbf{Q}_{22}(t) - \mathbf{V}_2(t)\mathbf{V}_2^{-1}(t)\mathbf{C}(t)\mathbf{Q}_{22}(t) + \mathbf{Q}_{12}(t)[\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]^T$$

or

$$\dot{\mathbf{Q}}_{12}(t) = \mathbf{Q}_{12}(t)[\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]^T. \tag{7.63}$$

Given the initial condition for this equation, $\mathbf{Q}_{12}(t) = \mathbf{0}$.

Using this solution for $\mathbf{Q}_{12}(t)$ in the last equation, (7.62), it is found that the variance of the innovation error is the solution of the equation,

$$\dot{\mathbf{Q}}_{22}(t) = \mathbf{V}_2(t). \tag{7.64}$$

The integral form of this equation is

$$\mathbf{Q}_{22}(t) = \int_{t_0}^{t} \mathbf{V}_2(\tau)d\tau. \tag{7.65}$$

The covariance matrix of a linear system can be written in terms of the variance matrix as

$$\mathbf{R}(t_1, t_2) = \begin{cases} \mathbf{Q}(t_1)\Phi^T(t_2, t_1), & \text{for } t_2 \geq t_1 \\ \Phi(t_1, t_2)\mathbf{Q}(t_2), & \text{for } t_1 > t_2 \end{cases}, \tag{7.66}$$

where $\Phi(t, \tau)$ is the transition matrix of the system. The transition matrix for the portion of the system in Eq. (7.57) which produces the innovations is $\mathbf{I}$, the identity matrix. Thus the covariance matrix which corresponds to equation (7.65) is

$$\mathbf{R}_n(t_1, t_2) = \int_{t_0}^{min(t_1,t_2)} \mathbf{V}_2(\tau)d\tau. \tag{7.67}$$

This matrix shows that the innovations process is a white noise process with the same intensity as the measurement noise. Thus if a perfect model of the system is used in a Kalman filter, the errors between its estimates and the state of the system will be white noise. In fact this is one of the most important ways of judging the quality of a given Kalman filter design: the whiteness of the noise on the innovations. In general failure to use a perfect model of the plant will introduce deterministic excitations into the white noise of the innovations signals.

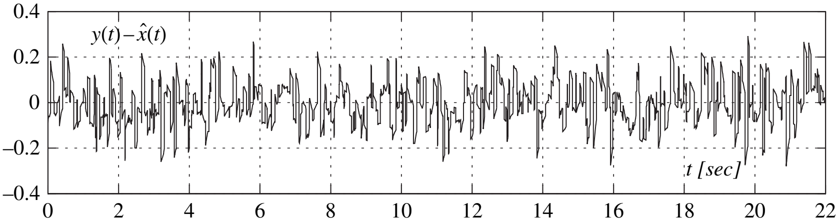


**Fig. 7.4** Innovation in the Kalman filter for the first order system of Fig. 7.2

***Example 7.5.* Innovation for a Simple Kalman Filter**

As an example of the dependence of the dependence of the innovation of a Kalman filter only on the state noise, the system of Example 7.2 is used. For the simulation of this example, the innovation was also calculated. This quantity is shown below in Fig. 7.4. In this example, $\sigma_{v_2} = 0.1$ , which agrees approximately with the $\pm 3\sigma$ amplitude seen in the figure. This simulation was made using SIMNON, Elmquist (1975). ❒

## 7.4 Discrete Kalman Filter

Discrete Kalman filters (DKF) are based on the same considerations as the continuous Kalman filters but their form can vary more. This is because the method of administering the time sampling and equation updating may be different from one application to another and thus from one realization to another. Two methods of constructing a discrete Kalman filter will be treated here: one which shows clearly how the filter works but has an open form and a second which will yield a closed form which is on the surface apparently more like a continuous filter. In all cases it must be remembered that the dynamics of differential and difference equations are different because the active element in a continuous block diagram is an integrator, while in a discrete diagram it is a time delay. This also helps to give the discrete algorithm a different appearance and function than in the continuous case.

### *7.4.1 A Real Time Discrete Kalman Filter (Open Form)*

The system which is to be considered has a form which is

$$\mathbf{x}(k+1) = \mathbf{F}(k)\mathbf{x}(k) + \mathbf{G}(k)\mathbf{u}(k) + \mathbf{G}_v(k)\mathbf{v}_1(k) \tag{7.68}$$

and the measurement equation is

$$\mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k) + \mathbf{v}_2(k). \tag{7.69}$$

In Eqs. (7.68) and (7.69), $\mathbf{v}_1(k)$ and $\mathbf{v}_2(k)$ are zero mean, uncorrelated white noise sequences with covariances $\mathbf{V}_1(k)$ and $\mathbf{V}_2(k)$. Here it will also be assumed that both noise sources have a Gaussian distribution for the sake of simplicity.

What is required is a recursive estimator for the state of a system at the time $t_k$ (or $k$ here) just before and just after a measurement $\mathbf{y}(k)$. If the state before the measurement is denoted by $\hat{\mathbf{x}}(k)^-$ and the state after a measurement by $\hat{\mathbf{x}}(k)^+$, then an estimate is to be made of the form,

$$\hat{\mathbf{x}}(k)^+ = \mathbf{M}(k)\hat{\mathbf{x}}(k)^- + \mathbf{L}(k)\mathbf{y}(k), \tag{7.70}$$

where the **M** and **L** are possibly time varying gains which are to be determined in such a way to obtain optimum filtering characteristics.

Concentrating once again on the filter error, the errors before and after measurements are given by

$$\hat{\mathbf{x}}(k)^+ = \mathbf{x}(k) + \mathbf{e}(k)^+, \tag{7.71}$$

$$\hat{\mathbf{x}}(k)^- = \mathbf{x}(k) + \mathbf{e}(k)^-. \tag{7.72}$$

Starting with the first equation above, solving for $\mathbf{e}(k)^+$,

$$\mathbf{e}(k)^+ = \hat{\mathbf{x}}(k)^+ - \mathbf{x}(k). \tag{7.73}$$

Then substituting for $\hat{\mathbf{x}}(k)^+$ from equation (7.71) and for $\hat{\mathbf{x}}(k)^-$ from the Eq. (7.72) one obtains, remembering that $\mathbf{y}(k) = \mathbf{C}(k)\mathbf{x}(k) + \mathbf{v}_2(k)$ and after collecting like terms,

$$\mathbf{e}(k)^+ = [\mathbf{M}(k) + \mathbf{L}(k)\mathbf{C}(k) - \mathbf{I}]\mathbf{x}(k) + \mathbf{M}(k)\mathbf{e}(k)^- + \mathbf{L}(k)\mathbf{v}_2(k). \tag{7.74}$$

The expectation value of the noise source is zero and the expectation value of the errors is required to be zero. This means that taking the expectation value of both sides of equation (7.74), the requirement is that

$$\mathbf{M}(k) = \mathbf{I} - \mathbf{L}(k)\mathbf{C}(k). \tag{7.75}$$

Using this expression in Eq. (7.70), the following estimator equation can be found,

$$\begin{aligned}\hat{\mathbf{x}}(k)^+ &= [\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]\hat{\mathbf{x}}(k)^- + \mathbf{L}(k)\mathbf{y}(k)\\ &= \hat{\mathbf{x}}(k)^- + \mathbf{L}(k)[\mathbf{y}(k) - \mathbf{C}(k)\hat{\mathbf{x}}(k)^-].\end{aligned} \tag{7.76}$$

This equation shows how the measurement of the output, $\mathbf{y}(k)$, is to be used to progress from the time before sampling to the time after sampling. This equation is called the state measurement time update equation.

To obtain an equation for the error covariance matrix measurement update, equivalent to that above for the state measurement update equation, it is convenient to start with the estimation error which can be shown from Eqs. (7.69), (7.71), (7.72) and (7.76) to be

$$\mathbf{e}(k)^+ = [\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]\mathbf{e}(k)^- + \mathbf{L}(k)\mathbf{v}_2(k). \tag{7.77}$$

The error covariance matrix is by definition,

$$\begin{aligned}\mathbf{Q}(k)^+ &= E\{\mathbf{e}(k)^+\mathbf{e}^T(k)^+\}\\ &= E\{[\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]\mathbf{e}(k)^-[\mathbf{e}^T(k)^-(\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k))^T + \mathbf{v}_2^T(k)\mathbf{L}^T(k)]\\ &\qquad + \mathbf{L}(k)\mathbf{v}_2(k)[\mathbf{e}^T(k)^-(\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k))^T + \mathbf{v}_2^T(k)\mathbf{L}^T(k)]\},\end{aligned} \tag{7.78}$$

where Eq. (7.77) has been used. By definition

$$E\{\mathbf{e}(k)^-\mathbf{e}^T(k)^-\} = \mathbf{Q}(k)^-, \quad E\{\mathbf{v}_2(k)\mathbf{v}_2^T(k)\} = \mathbf{V}_2(k). \tag{7.79}$$

The measurement errors and noise source sequences are uncorrelated, thus

$$E\{\mathbf{e}(k)^-\mathbf{v}_2^T(k)\} = E\{\mathbf{v}_2(k)\mathbf{e}^T(k)^-\} = 0, \tag{7.80}$$

which can be used to simplify Eq. (7.78)

$$\mathbf{Q}(k)^+ = [\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]\mathbf{Q}(k)^-[\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]^T + \mathbf{L}(k)\mathbf{V}_2(k)\mathbf{L}^T(k). \tag{7.81}$$

This is the error covariance measurement update equation.

In order to fix the optimal value of the Kalman gain, $\mathbf{L}(k)$, some criterion for optimality must be given. In this case the trace of the diagonal elements of the error covariance matrix should be minimized. This corresponds to minimizing a quantity proportional to the total A.C. power in the error,

$$J(k) = tr\ E\{\mathbf{e}(k)^+\mathbf{e}^T(k)^+\} = tr\ \mathrm{Q}(k)^+. \tag{7.82}$$

To carry this through with the system in matrix form, the following derivative identity for matrices is required,

$$\frac{\partial}{\partial \mathbf{A}}[tr(\mathbf{ABA}^T)] = 2\mathbf{AB}, \tag{7.83}$$

where $\mathbf{B}$ has to be symmetric. The differentiation is to be carried out with respect to $\mathbf{L}(k)$ because the index or criterion function is to be minimized with respect to the Kalman gain.

Using Eqs. (7.71) and (7.72) in Eq. (7.82) and (7.83) and setting the result equal to zero gives

$$-2[\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]\mathbf{Q}(k)^{-}\mathbf{C}^{T}(k) + 2\mathbf{L}(k)\mathbf{V}_{2}(k) = 0. \tag{7.84}$$

Solving this equation for $\mathbf{L}(k)$ yields the optimum (Kalman) gain,

$$\mathbf{L}(k) = \mathbf{Q}(k)^{-}\mathbf{C}^{T}(k)[\mathbf{C}(k)\mathbf{Q}(k)^{-}\mathbf{C}^{T}(k) + \mathbf{V}_{2}(k)]^{-1}. \tag{7.85}$$

It can be shown by calculating the Hessian matrix of $J(k)$ that this value of the gain actually does give the minimum value of the index. Substitution of this value of the gain into Eq. (7.81) yields a simple expression for the covariance measurement update,

$$\mathbf{Q}(k)^{+} = [\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]\mathbf{Q}(k)^{-}, \tag{7.86}$$

after some manipulation.

Thus the equations that describe the measurement updating of the estimate of $\mathbf{x}(k)$ and the error covariance matrix are Equations (7.87) (repeated from (7.76)) and (7.88) with the optimal gain given in equation (7.89) are

$$\hat{\mathbf{x}}(k)^{+} = \hat{\mathbf{x}}(k)^{-} + \mathbf{L}(k)[\mathbf{y}(k) - \mathbf{C}(k)\hat{\mathbf{x}}(k)], \tag{7.87}$$

$$\mathbf{Q}(k)^{+} = [\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]\mathbf{Q}(k)^{-}, \tag{7.88}$$

$$\mathbf{L}(k) = \mathbf{Q}(k)^{-}\mathbf{C}^{T}[\mathbf{C}(k)\mathbf{Q}(k)^{-}\mathbf{C}^{\mathbf{T}}(k) + \mathbf{V}_{2}(k)]^{-1}. \tag{7.89}$$

What must now be accomplished is to describe how the state and covariance estimates propagate between sampling instants. This can be done in more or less the same way as for a continuous time system. Consider the system estimator

$$\hat{\mathbf{x}}(k+1) = \mathbf{F}(k)\hat{\mathbf{x}}(k) + \mathbf{G}(k)\mathbf{u}(k). \tag{7.90}$$

Subtracting Eq. (7.68) from this equation, the following result is found,

$$\mathbf{e}(k+1) = \mathbf{F}(k)\mathbf{e}(k) - \mathbf{G}_{v}(k)\mathbf{v}_{1}(k), \tag{7.91}$$

where $\mathbf{e}(k) = \hat{\mathbf{x}}(k) - \mathbf{x}(k)$. Taking the expectation value of both sides of equation (7.40) shows that the estimator above will be unbiased (will have mean value zero) since the mean value of the error and the noise is zero.

The error covariance is given by

$$\begin{aligned}\mathbf{Q}(k+1) &= E\{\mathbf{e}(k+1)\mathbf{e}^T(k+1)\}\\ &= E\{(\mathbf{F}(k)\mathbf{e}(k) - \mathbf{G}_v(k)\mathbf{v}_1(k))(\mathbf{F}(k)\mathbf{e}(k) - \mathbf{G}_v(k)\mathbf{v}_1(k))^T\}\\ &= E\{\mathbf{F}(k)\mathbf{e}(k)\mathbf{e}^T(k)\mathbf{F}^T(k) - \mathbf{F}(k)\mathbf{e}(k)\mathbf{v}_1{}^T(k)\mathbf{G}_v{}^T(k)\\ &\qquad - \mathbf{G}_v(k)\mathbf{v}_1(k)\mathbf{e}^T(k)\mathbf{F}^T(k) + \mathbf{G}_v(k)\mathbf{v}_1(k)\mathbf{v}_1{}^T(k)\mathbf{G}_v^T(k)\}\end{aligned}$$

or

$$\mathbf{Q}(k+1) = \mathbf{F}(k)\mathbf{Q}(k)\mathbf{F}^T(k) + \mathbf{G}_v(k)\mathbf{V}_1(k)\mathbf{G}_v^T(k). \tag{7.92}$$

This is the discrete time version of the time dependent Lyapunov equation.

In order to make the equations above compatible with the measurement update equations their arguments have to be changed to agree with those used above:

$$\hat{\mathbf{x}}(k)^- = \mathbf{F}(k-1)\hat{\mathbf{x}}(k-1)^+ + \mathbf{G}(k-1)\mathbf{u}(k-1), \tag{7.93}$$

$$\mathbf{Q}(k)^- = \mathbf{F}(k-1)\mathbf{Q}(k-1)^+\mathbf{F}^T(k-1) + \mathbf{G}_v(k-1)\mathbf{V}_1(k-1)\mathbf{G}_v^T(k-1). \tag{7.94}$$

A summary of the equations of the open form of the discrete Kalman filter (DKF) is given in Table 7.1 on p. 438.

### 7.4.2 Block Diagram of an Open Form DKF

The discrete open form Kalman filter has approximately the same block diagram as the continuous version except for the structure of the internal block and the switching networks required by the sampling process. This may be seen in Fig. 7.5.

The substitutions above imply that the continuous white noise signals must be replaced by discrete white noise sequences. In the same way the input and output vectors are now discrete sampled signals and the operations denoted by the summing points are to be performed at the sampling times $t_k$, $k = 0, 1, 2, \ldots$. This does not imply however that the conditions for the validity of the filter equations have been changed or that the filter is of a principally different nature than in the continuous case. The requirements of the system to be filtered (translated to discrete time) and the functions of the covariance time/measurement update are the same as before. The equations characterizing the discrete process and the Kalman filter are again to be found in Table 7.1. Here they may be compared to the equations for the CKF.

The main differences between the continuous and the discrete Kalman filters are the differences that always exist between continuous and discrete systems

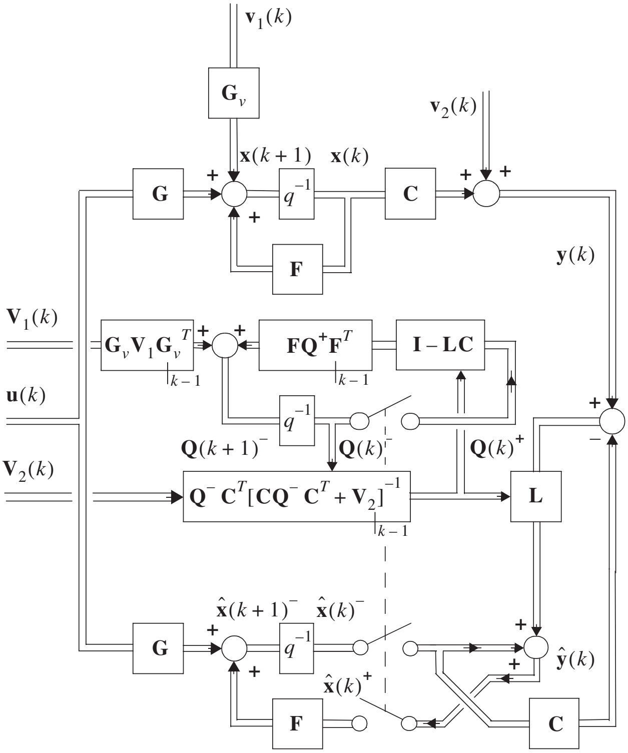


**Fig. 7.5** Block diagram of an open form discrete Kalman filter

and the explicit separation that becomes necessary in the time and measurement updates in time. This is shown in Fig. 7.6. After the sampling time $t_{k-1}$ the state estimates are $\hat{\mathbf{x}}(k-1)^+$ and the error covariances are $\mathbf{Q}(k-1)^+$. Time updating these estimates and covariances consists of propagating them through the system dynamics to the time $t_k^-$ just before a measurement is taken. This is accomplished by using the discrete time transition matrix of the system $\mathbf{F}(k)$ and the covariance time update equation. At the time update the state estimates and covariance matrix are denoted by $\boldsymbol{x}(k)^-$and $\mathbf{Q}(k)^-$. Taking a measurement at a time $t_k$ makes it possible to perform a measurement update bringing these two quantities into the new time interval as $\hat{\mathbf{x}}(k)^+$ and $\mathbf{Q}(k)^+$. A new time update can now be performed and so on. Thus the separation of the time and measurement updates in the discrete Kalman filter is due to the fact that measurements on and inputs to the system may only occur at sampling times. In the continuous case these activities occur concurrently and at all times.

The main disadvantage that the discrete filter has in relation to the continuous one is exactly the same as that discrete regulators have in relation to continuous regulators: for finite sampling intervals there is less information available to the filter about the state of the system because it only works at

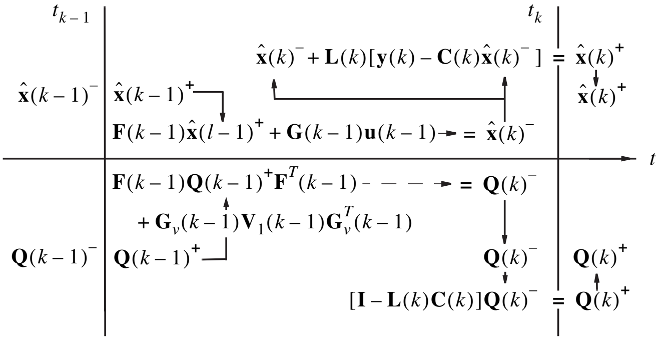

$t_{k-1}$
$t_k$
$\hat{\mathbf{x}}(k)^- + \mathbf{L}(k)[\mathbf{y}(k) - \mathbf{C}(k)\hat{\mathbf{x}}(k)^-] = \hat{\mathbf{x}}(k)^+$
$\hat{\mathbf{x}}(k-1)^-$
$\hat{\mathbf{x}}(k-1)^+$
$\hat{\mathbf{x}}(k)^+$
$\mathbf{F}(k-1)\hat{\mathbf{x}}(l-1)^+ + \mathbf{G}(k-1)\mathbf{u}(k-1) = \hat{\mathbf{x}}(k)^-$
$t$
$\mathbf{F}(k-1)\mathbf{Q}(k-1)^+\mathbf{F}^T(k-1) = \mathbf{Q}(k)^-$
$+ \mathbf{G}_v(k-1)\mathbf{V}_1(k-1)\mathbf{G}_v^T(k-1)$
$\mathbf{Q}(k-1)^-$
$\mathbf{Q}(k-1)^+$
$\mathbf{Q}(k)^-$
$\mathbf{Q}(k)^+$
$[\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]\mathbf{Q}(k)^- = \mathbf{Q}(k)^+$

**Fig. 7.6** Time axis showing the explicit time separation of the time and measurement updates in an open form discrete Kalman filter. Sampling takes place at the time events, $t_{k-1}$ and $t_k$. The detailed method of operation of the filter is indicated on the diagram as the actual measurement and time update equations shown leading up to the time instant, $t_k$

sampling times. Only when the sampling period approaches zero can the discrete filter's response become as fast and as noise free as that of a continuous filter. The main advantage of the discrete time filter is that it can be easily realized immediately in a lower level computer language (for example Assembler) in an event driven microprocessor.

***Example 7.6.*** **Discrete Estimation of a Constant**

A constant can be estimated from the system dynamics and measurement models:

$$x(k+1) = x(k), \tag{7.95}$$

$$y(k) = x(k) + v_2(k). \tag{7.96}$$

It is assumed that the noise source $v_2(k)$ is a normally distributed white noise sequence with an intensity or variance $V_2(k)$. The initial condition for the variance is that $E\{Q(0)\} = q_0$.

The time development of the variance is given by equation (7.43). As $F = 1$ and $V_1 = 0$, it is clear that $q(k+1)^- = q(k)^+$. This means, using Eq. (7.90) in Eq. (7.89), it is possible to write

$$q(k)^+ = q(k)^- - q(k)^- C^T[Cq(k)^- C^T + V_2]^{-1} Cq(k)^- \tag{7.97}$$

which implies that

$$q(k)^+ = \frac{q(k)^-}{1 + \frac{q(k)^-}{V_2}} = \frac{q(k-1)^+}{1 + \frac{q(k-1)^+}{V_2}} \tag{7.98}$$

for this problem. Because of its simplicity and structure this difference equation can be solved directly by iteration:

$$\begin{aligned} q(1)^+ &= \frac{q_0}{1+\dfrac{q_0}{V_2}}, \\ q(2)^+ &= \frac{q_1}{1+\dfrac{q_1}{V_2}} = \frac{q_0}{1+2\dfrac{q_0}{V_2}}, \\ &\dots \\ q(k) = q(k)^+ &= \frac{q_0}{1+k\dfrac{q_0}{V_2}}. \end{aligned} \tag{7.99}$$

The Kalman gain is then

$$L(k) = \frac{q(k)}{V_2} \tag{7.100}$$

and the observer

$$\hat{x}(k+1) = \hat{x}(k) + L(k)[y(k) - \hat{x}(k)]. \tag{7.101}$$

For large times ($k$ large) it can be seen that the Kalman gain goes to zero and new measurements are effectively not processed as earlier for the continuous time analog. ❒

### 7.4.3 Closed Form of a DKF

It is possible to derive a closed form of the Kalman filter equations above. To accomplish this it is only necessary to substitute Eq. (7.76) and (7.88) into the right sides of equations (7.93) and (7.94) respectively. This gives

$$\begin{aligned} \hat{\mathbf{x}}(k) &= \mathbf{F}(k-1)\{\hat{\mathbf{x}}(k-1) + \mathbf{L}(k-1)[\mathbf{y}(k-1) - \mathbf{C}(k-1)\hat{\mathbf{x}}(k-1)]\} \\ &= \mathbf{F}(k-1)\hat{\mathbf{x}}(k-1) + \mathbf{F}(k-1)\mathbf{L}(k-1)[\mathbf{y}(k-1) - \mathbf{C}(k-1)\hat{\mathbf{x}}(k-1)] \end{aligned} \tag{7.102}$$

where $\hat{\mathbf{x}}(k) \equiv \hat{\mathbf{x}}(k)^-$ and

$$\begin{aligned} \mathbf{Q}(k) &= \mathbf{F}(k-1)[\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k-1)]\mathbf{Q}(k-1)\mathbf{F}^T(k-1) \\ &\qquad + \mathbf{G}_v(k-1)\mathbf{V}_1(k-1)\mathbf{G}_v^T(k-1) \\ &= \mathbf{F}(k-1)\mathbf{Q}(k-1)\mathbf{F}^T(k-1) \end{aligned}$$

$$\begin{aligned} &- \mathbf{F}(k-1)\mathbf{L}(k-1)\mathbf{C}(k-1)\mathbf{Q}(k-1)\mathbf{F}^T(k-1) \\ &\qquad + \mathbf{G}_v(k-1)\mathbf{V}_1(k-1)\mathbf{G}_v^T(k-1), \end{aligned} \tag{7.103}$$

where $\mathbf{Q}(k) \equiv \mathbf{Q}(k)^-$.

In the last equations if the Kalman gain is be redefined to be $\mathbf{L}'(k) = \mathbf{F}(k-1)\mathbf{L}(k-1)$ then the Kalman filter equations may be written as

$$\hat{\mathbf{x}}(k+1) = \mathbf{F}(k)\hat{\mathbf{x}}(k) + \mathbf{G}(k)\mathbf{u}(k) - \mathbf{L}'(k)[\mathbf{y}(k) - \mathbf{C}(k)\hat{\mathbf{x}}(k)] \tag{7.104}$$

$$\begin{aligned} \mathbf{Q}(k+1) = &\mathbf{F}(k)\mathbf{Q}(k)\mathbf{F}^T(k) + \mathbf{G}_v(k)\mathbf{V}_1(k)\mathbf{G}_v^T(k) \\ &- \mathbf{L}'(k)\mathbf{C}(k)\mathbf{Q}(k)\mathbf{F}^T(k), \end{aligned} \tag{7.105}$$

with a change in the time index. The second equation is the discrete time Riccati equation. In this case the Kalman gain is in this case given by

$$\mathbf{L}'(k) = \mathbf{F}(k)\mathbf{Q}(k)\mathbf{C}^T(k)[\mathbf{C}(k)\mathbf{Q}(k)\mathbf{C}^T(k) + \mathbf{V}_2(k)]^{-1}. \tag{7.106}$$

As in the case of the continuous time Kalman filter it is often convenient to use a steady state solution of the Riccati equation instead of a time dependent one for linear time invariant control objects. This is possible because in general the Riccati equation will approach a constant solution for large times. This comes about because the noise variances of the state and measurement noise sources are the only active inputs to the Riccati equation. This is made clear in the block diagram, Fig. 7.7. The closed form of the Riccati equation makes this straight forward. Table 7.2 on p. 458 shows the equations for the closed form of the discrete Kalman filter and compares these to the equations for the open form See Lewis (1992) for a different treatment.

It is often necessary to construct a suboptimal Kalman filter by using a constant Kalman gain. This is most often accomplished by assuming that $V_1$ and $V_2$ are constant and solving the Riccati equation for this assumption. The Riccati equation can then be solved for the stationary case by setting the error variance on the left hand side of equation (7.105) equal to $\mathbf{Q}(k)$,

$$\mathbf{Q}(k) = \mathbf{F}(k)\mathbf{Q}(k)\mathbf{F}^T(k) + \mathbf{G}_v(k)\mathbf{V}_1(k)\mathbf{G}_v^T(k) - \mathbf{L}'(k)\mathbf{C}(k)\mathbf{Q}(k)\mathbf{F}^T(k), \tag{7.107}$$

and inserting the Kalman gain from Eq. (7.106). This yields a set of coupled quadratic equations which can be solved for **Q**(k). Once the error variance has been found the actual value of the constant Kalman gain can be found. While the solution found in this way is suboptimal, it will in general be good enough to obtain stable and accurate observer characteristics around the design operating point.

**Table 7.2** Summary of the Discrete Kalman Filter Equations (Linear Systems Only)

| | Discrete Kalman Filter (closed form) | Discrete Kalman Filter (open form) |
|---|---|---|
| Control Object | $\mathbf{x}(k+1)=\mathbf{F}(k)\mathbf{x}(k)+\mathbf{G}(k)\mathbf{u}(k)+\mathbf{G}_v(k)\mathbf{v}_1(k)$ | $\mathbf{x}(k+1)=\mathbf{F}(k)\mathbf{x}(k)+\mathbf{G}(k)\mathbf{u}(k)+\mathbf{G}_v(k)\mathbf{v}_1(k)$ |
| Measurement Model | $\mathbf{y}(k)=\mathbf{C}(k)\mathbf{x}(k)+\mathbf{v}_2(k)$ | $\mathbf{y}(k)=\mathbf{C}(k)\ \mathbf{x}(k)+\mathbf{v}_2(k)$ |
| State Estimate Time Update | $\hat{\mathbf{x}}(k+1)=\mathbf{F}(k)\hat{\mathbf{x}}(k)+\mathbf{G}(k)\mathbf{u}(k)$<br>$+\mathbf{L}(k)[\mathbf{y}(k)-\mathbf{C}(k)\hat{\mathbf{x}}(k)]$ | $\hat{\mathbf{x}}(k)^-=\mathbf{F}(k-1)\mathbf{x}(k-1)^++\mathbf{G}(k-1)\mathbf{u}(k-1)$ |
| Covariance Time Update | $\mathbf{Q}(k+1)=\mathbf{F}(k)\mathbf{Q}(k)\mathbf{F}^T(k)+\mathbf{G}_v(k)\mathbf{V}_1(k)\mathbf{G}_v^T(k)$<br>$-\mathbf{L}'(k)\mathbf{C}(k)\mathbf{Q}(k)\mathbf{F}^T(k)$ | $\mathbf{Q}(k)^-=\mathbf{F}(k-1)\mathbf{Q}(k-1)^+\mathbf{F}^T(k-1)$<br>$+\mathbf{G}(k-1)\mathbf{V}_1(k-1)\mathbf{G}^T(k-1)$ |
| Initial Conditions for Time Update above | $E\{\mathbf{x}(0)\}=\hat{\mathbf{x}}(0)$ | $E\{\mathbf{x}(0)\}=\hat{\mathbf{x}}(0)$ |
| | $E\left\{(\mathbf{x}(0)-\hat{\mathbf{x}}(0))(\mathbf{x}(0)-\hat{\mathbf{x}}(0))^T\right\}=\mathbf{Q}(0)$ | $E\{(\mathbf{x}(0)-\hat{\mathbf{x}}(0))(\mathbf{x}(0)-\hat{\mathbf{x}}(0))^T\}=\mathbf{Q}(0)$ |
| Kalman Gain Matrix | $\mathbf{L}'(k)=\mathbf{F}(k)\mathbf{Q}(k)\mathbf{C}^T(k)[\mathbf{C}(k)\mathbf{Q}(k)\mathbf{C}^T(k)+\mathbf{V}_2]^{-1}$ | $\mathbf{L}(k)=\mathbf{Q}(k)^-\mathbf{C}^T(k)[\mathbf{C}(k)\mathbf{Q}(k)^-\mathbf{C}^T(k)+\mathbf{V}_2]^{-1}$ |
| State Estimate Measurement Update | Included in time update above | $\hat{\mathbf{x}}(k)^+=\hat{\mathbf{x}}(k)^-+\mathbf{L}(k)[\mathbf{y}(k)-\mathbf{C}(k)\hat{\mathbf{x}}(k)^-]$ |
| Covariance Measurement Update | Included in time update above | $\mathbf{Q}(k)^+=[\mathbf{I}-\mathbf{L}(t)\mathbf{C}(t))]\mathbf{Q}(k)^-$ |
| Other Assumptions | $E\{\mathbf{v}_1(j)\mathbf{v}_2^T(k)\}=0,$ for all j, k<br>$\mathbf{v}_1(k)\in N(0,\mathbf{V}_1),\ \mathbf{v}_2(k)\in N(0,\mathbf{V}_2)$ | $E\{\mathbf{v}_1(j)\mathbf{v}_2^T(k)\}=0,$ for all j, k<br>$\mathbf{v}_1(k)\in N(0,\mathbf{V}_1),\ \mathbf{v}_2(k)\in N(0,\mathbf{V}_2)$ |

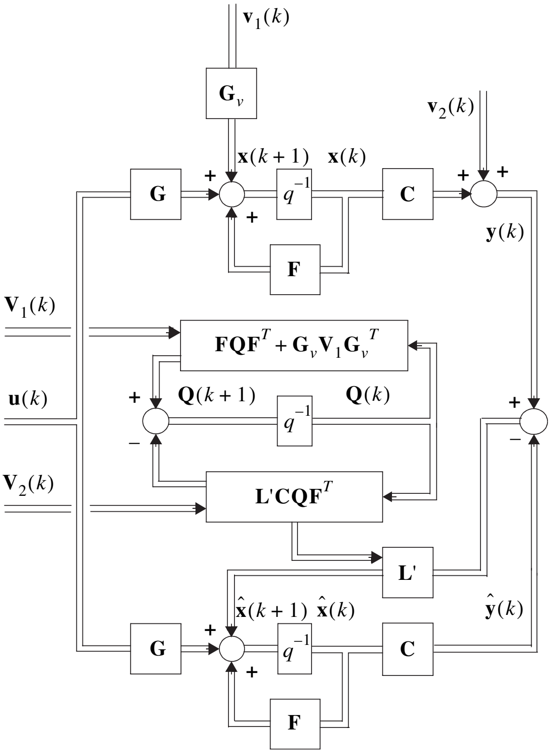


**Fig. 7.7** Block diagram of a closed form discrete Kalman filter

### *Example 7.7.* Discrete Kalman Filter for a Discrete Integrator

A digital integrator is described by the transfer function which emerges when a Z-transform is taken of a zeroth order hold network in series with an analog integrator. This is

$$\frac{y(z)}{u(z)} = \frac{Tz^{-1}}{1 - az^{-1}}. \tag{7.108}$$

The state equations and output equations for a digital integrator disturbed by state and measurement noise sources are thus

$$x(k+1) = ax(k) + Tu(k) + v_1(k), \tag{7.109}$$

$$y(k) = x(k) + v_2(k). \tag{7.110}$$

$a$ is a constant while T is the sampling time interval. It will be assumed that the state and measurement noise sources are white uncorrelated Gaussian noise with constant intensities $V_1$ and $V_2$ respectively.

Equation (7.105) is only a function of $k$ in steady state. This makes it possible to drop the time index and write:

$$l = aQ\,[Q + V_2]^{-1}. \tag{7.111}$$

$Q$ can be found from the equation,

$$\begin{aligned} Q &= [a - L]Qa + V_1 \\ &= [a - aQ(Q + V_2)^{-1}]Qa + V_1. \end{aligned} \tag{7.112}$$

This can be reduced to the quadratic equation,

$$Q^2 + Q[(1 - a^2)V_2 - V_1] - V_1 V_2 = 0, \tag{7.113}$$

which has to be solved for Q. The solution is given by

$$Q = \frac{1}{2}[V_1 - (1 - a^2)V_2] \pm \frac{1}{2}\sqrt{[V_1 - (1 - a^2)V_2]^2 + 4\,V_1 V_2}, \tag{7.114}$$

where the positive (definite) solution has to be selected. This quantity depends on the sign and magnitude of $a$ and the relative magnitudes of $V_1$ and $V_2$. When $Q$ has been found then it has to be substituted into the equation,

$$l = \frac{aQ}{Q + V_2}, \tag{7.115}$$

to find the required stationary value of the Kalman gain. ❐

***Example 7.8*. Kalman Filter for a First Order Discrete System**

Consider the first order system which was treated in Example 7.1. The z-transform of the state equation (including a zeroth order hold network) yields the expression,

$$\frac{x(z)}{u(z)} = \frac{b(1 - e^{-aT})z^{-1}}{1 - e^{-aT}z^{-1}}. \tag{7.116}$$

This transfer function is easily translated into a state space description which is

$$x(k + 1) = e^{-aT}x(k) + b(1 - e^{-aT})u(k). \tag{7.117}$$

Assuming noise sources like those in Example 7.1, the discretized model for the first order system becomes,

$$x(k + 1) = e^{-aT}x(k) + b(1 - e^{-aT})u(k) + Tv_1(k), \tag{7.118}$$

$$y(k) = x(k) + \frac{1}{T}v_2(k), \tag{7.119}$$

where the discretized versions of the noise sources in the earlier example have been used.

If $f = e^{-aT}$ and $g = b(1 - e^{-aT})$ then a Kalman filter for the system above can be expressed as:

$$\hat{\mathbf{x}}(k+1) = f\hat{x}(k) + g\,u(k) + l[y(k) - \hat{x}(k)], \tag{7.120}$$

where the Kalman gain is given by

$$l(k) = f q(k)\left[q(k) + \frac{\sigma_2^2}{T^2}\right]^{-1}. \tag{7.121}$$

$q(k)$ is found by solving the discrete Riccati equation,

$$q(k+1) = f^2\,q(k) + \sigma_1^2 T^2 - l(k) f q(k). \tag{7.122}$$

In the stationary state $q(k)$ becomes

$$\lim_{k\to\infty} q(k) \to -\frac{1}{2}\left(\frac{\sigma_2^2}{T^2}(1-f^2) - \sigma_1^2 T^2\right) \pm \frac{1}{2}\sqrt{\left(\frac{\sigma_2^2}{T^2}(1-f^2) - \sigma_1^2 T^2\right)^2 + 4\sigma_1^2\sigma_2{}^2}, \tag{7.123}$$

where the positive sign has to be selected in order that the error covariance is positive. Using the same input and the same values for the parameters in this case as in Fig. 7.2, $a = 1$, $b = 1$, $\sigma_1 = 0.5$ and $\sigma_2 = 0.1$, the results of simulating the discrete time Kalman filter with a continuous estimation object are shown in Fig. 7.8. The sample time is $T = 0.2$ sec.

These results should be compared to those shown in Fig. 7.2. It is apparently the case that the continuous and discrete filters have approximately the same performance given similar operating conditions. Note also that the innovation noise level corresponds to $\sigma_2 = 0.1/0.2 = 0.5$ in contrast to the earlier example. See Fig. 7.4. ❐

## 7.4.4 Discrete and Continuous Kalman Filter Equivalence

It is not difficult to show that the discrete and continuous Kalman filters can be derived from each other. In this section it will be shown that the continuous Kalman filter can be derived from the open form discrete Kalman filter.

Using the Euler integration approximation to the continuous time dynamics with a sampling period $T$, the system equations can be written

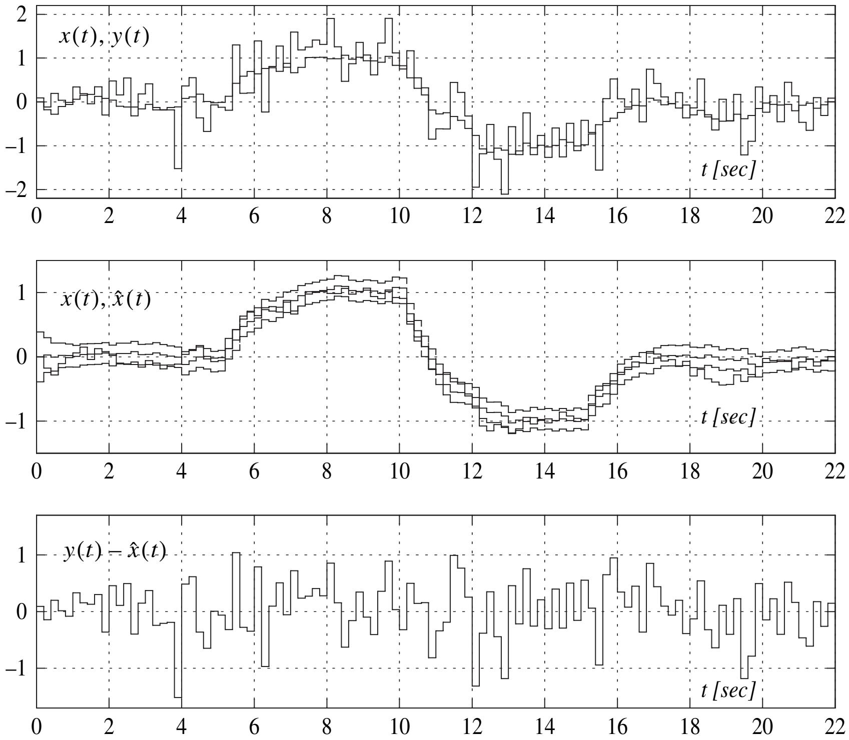


**Fig. 7.8** Simulation of a Kalman filter for a first order system. The input is the same as in Fig. 7.2 and results should be compared with those shown in this figure. See also Fig. 7.4

$$\mathbf{x}(k+1) = (\mathbf{I} + \mathbf{A}T)\mathbf{x}(k) + \mathbf{B}T\mathbf{u}(k) + \mathbf{B}_v\mathbf{v}_1(k), \tag{7.124}$$

$$\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) + \mathbf{v}_2(k), \tag{7.125}$$

where $\mathbf{v}_1(k) \in N(0, \mathbf{V}_1 T), \mathbf{v}_2(k) \in N\left(0, \frac{\mathbf{V}_2}{T}\right)$ with the standard initial conditions.

The state estimate discrete update equation can be written down using Eq. (7.87) and the fact that

$$\hat{\mathbf{x}}(k+1)^- = \mathbf{A}\hat{\mathbf{x}}(k) + \mathbf{B}\mathbf{u}(k). \tag{7.126}$$

This implies that

$$\begin{aligned} \hat{\mathbf{x}}(k+1)^- = {} & (\mathbf{I} + \mathbf{A}T)\hat{\mathbf{x}}(k)^- + \mathbf{B}T\mathbf{u}(k) \\ & + \mathbf{L}(k+1)[\mathbf{y}(k+1) - \mathbf{C}(\mathbf{I} + \mathbf{A}T)\hat{\mathbf{x}}(k) - \mathbf{C}\mathbf{B}T\mathbf{u}(k)]. \end{aligned} \tag{7.127}$$

Rearranging and dividing through by $T$ yields

$$\begin{aligned}\frac{\hat{\mathbf{x}}(k+1)-\hat{\mathbf{x}}(k)}{T} &= \mathbf{A}\hat{\mathbf{x}}(k)+\mathbf{Bu}(k)\\ &\quad+\frac{\mathbf{L}(k)}{T}[\mathbf{y}(k+1)-\mathbf{C}\hat{\mathbf{x}}(k)-\mathbf{C}(\mathbf{A}\hat{\mathbf{x}}(k)+\mathbf{Bu}(k))T].\end{aligned} \tag{7.128}$$

The Kalman gain is given by

$$\mathbf{L}(k+1)=\mathbf{Q}(k+1)^{-}\mathbf{C}^{T}\left(\mathbf{CQ}(k+1)^{-}\mathbf{C}^{T}+\frac{\mathbf{V}_2}{T}\right)^{-1}, \tag{7.129}$$

from Eq. (7.89) and thus from Eq. (7.129),

$$\frac{\mathbf{L}(k+1)}{T}=\mathbf{Q}(k+1)^{-}\mathbf{C}^{T}(\mathbf{CQ}(k+1)^{-}\mathbf{C}^{T}T+\mathbf{V}_2)^{-1}. \tag{7.130}$$

For $T \to 0$ one has that

$$\lim_{T\to 0}\frac{\mathbf{L}(k+1)}{T}\to \mathbf{L}(t)=\mathbf{Q}(t)\mathbf{C}^{T}\mathbf{V}_2^{-1}. \tag{7.131}$$

This results in the continuous time/measurement update for the state estimate or state mean value,

$$\begin{aligned}\dot{\hat{\mathbf{x}}}(t) &= \mathbf{A}\hat{\mathbf{x}}(t)+\mathbf{Bu}(t)+\mathbf{Q}(t)\mathbf{C}^{T}\mathbf{V}_2^{-1}[\mathbf{y}(t)-\mathbf{C}\hat{\mathbf{x}}(t)]\\ &= \mathbf{A}\hat{\mathbf{x}}(t)+\mathbf{Bu}(t)+\mathbf{L}(t)[\mathbf{y}(t)-\mathbf{C}\hat{\mathbf{x}}(t)],\end{aligned} \tag{7.132}$$

which is the same as Eq (7.7) (when it is remembered that $\hat{\mathbf{x}}(kT)=\hat{\mathbf{x}}(k)$ and $\mathbf{L}(kT)=(1/T)\mathbf{L}(k)$). The initial condition for integrating equation (7.132) is $\hat{\mathbf{x}}(0)=\mathbf{x}_0$.

The same approach as above can be used to derive the continuous convariance time/measurement update or Riccati equation. The discrete error covariance update equation is

$$\mathbf{Q}(k+1)^{+}=[\mathbf{I}-\mathbf{L}(k+1)\mathbf{C}]\mathbf{Q}(k+1)^{-}, \tag{7.133}$$

from Eq. (7.88) and

$$\mathbf{Q}(k+1)=(\mathbf{I}+\mathbf{A}T)\mathbf{Q}(k)(\mathbf{I}+\mathbf{A}T)^{T}+\mathbf{B}_v\mathbf{V}_1\mathbf{B}_v^{T}T, \tag{7.134}$$

from Eq. (7.92). From Eq. (7.134) it is clear that

$$\mathbf{Q}(k+1)^{-}=\mathbf{Q}(k)^{+}+(\mathbf{AQ}(k)+\mathbf{Q}(k)\mathbf{A}^{T}+\mathbf{B}_v\mathbf{V}_1\mathbf{B}_v^{T})T+\mathbf{0}(T^2). \tag{7.135}$$

If $\mathbf{Q}(k)^+$ is substituted into this expression from Eq. (7.133),

$$\begin{aligned}\mathbf{Q}(k+1)^- =&[\mathbf{I}-\mathbf{L}(k)\mathbf{C}]\mathbf{Q}(k)^- + [\mathbf{A}(\mathbf{I}-\mathbf{L}(k)\mathbf{C})\mathbf{Q}(k)^-]T \\ &+ [(\mathbf{I}-\mathbf{L}(k)\mathbf{C})\mathbf{Q}(k)^-\mathbf{A}^T + \mathbf{B}_v\mathbf{V}_1\mathbf{B}_v^T]T + \mathbf{0}(T^2).\end{aligned} \tag{7.136}$$

Dividing through by $T$ and rearranging this equation becomes

$$\begin{aligned}\frac{\mathbf{Q}(k+1)^- - \mathbf{Q}(k)^-}{T} &= \mathbf{A}\mathbf{Q}^-(k) + \mathbf{Q}(k)^-\mathbf{A}^T + \mathbf{B}_v\mathbf{V}_1\mathbf{B}_v^T \\ &- \mathbf{A}\frac{\mathbf{L}(k)}{T}T\mathbf{C}\mathbf{Q}(k)^- - \frac{\mathbf{L}(k)}{T}T\mathbf{C}\mathbf{Q}(k)^-\mathbf{A}^T - \frac{\mathbf{L}(k)}{T}\mathbf{C}\mathbf{Q}(k)^-.\end{aligned} \tag{7.137}$$

Now letting $T \to 0$ and remembering that $\mathbf{Q}(kT) = \mathbf{Q}(k)^-$ results in

$$\begin{aligned}\dot{\mathbf{Q}}(t) =&\mathbf{A}(t)\mathbf{Q}(t) + \mathbf{Q}(t)\mathbf{A}^T(T) \\ &+ \mathbf{B}_v(t)\mathbf{V}_1(t)\mathbf{B}_v^T(t) - \mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t)\mathbf{C}(t)\mathbf{Q}(t),\end{aligned} \tag{7.138}$$

which is the Riccati equation for continuous systems. The initial condition for integrating this equation is $\mathbf{Q}(0) = \mathbf{Q}_0$.

## 7.5 Stochastic Integral Quadratic Forms

In what follows the problem of combining Kalman filters with LQR regulators will be considered. In order to evaluate the error which might be involved in joining such systems together it is necessary to evaluate scalar indexes which have the form:

$$J = E\left\{\int_{t_0}^{t_1} \mathbf{x}^T(\tau)\mathbf{R}(\tau)\,d\tau + \mathbf{x}^T(t_1)\mathbf{P}(t)\mathbf{x}(t_1)\right\}. \tag{7.139}$$

Modified techniques of the same kind as above can be used to evaluate such indexes (where $\mathbf{R}(t)$ is symmetric and positive definite and $\mathbf{P}(t_1)$ is constant, symmetric and positive semidefinite). Indexes of this type often occur in optimal control and optimal observer control systems which contain stochastic signals.

Consider a linear dynamic system which has a form which is,

$$\dot{\mathrm{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{v}(t), \tag{7.140}$$

where $\mathbf{v}(t)$ is white noise with an intensity which is $\mathbf{V}(t)$ and where $\mathbf{x}(t)$ is a vector stochastic variable subject to the initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0$ and

$E\{\mathbf{x}_0\mathbf{x}_0^T\} = \mathbf{Q}_0$. If $\mathbf{R}(\tau)$ is a symmetric and positive semidefinite matrix (for $t_0 \leq t \leq t_1$) and $\mathbf{P}_1$ is symmetric, constant and positive semidefinite then it can be shown that

$$\begin{aligned} J &= E\left\{\int_{t_0}^{t_1} \mathbf{x}^T(\tau)\mathbf{R}(\tau)\mathbf{x}(\tau)\ d\tau + \mathbf{x}^T(t_1)\mathbf{P}(t_1)\mathbf{x}(t_1)\right\} \\ &= tr\left[\mathbf{P}(t_0)\mathbf{Q}_0 + \int_{t}^{t_1} \mathbf{B}(\tau)\mathbf{V}(\tau)\mathbf{B}^T(\tau)\mathbf{P}(\tau)\ d\tau\right], \end{aligned} \tag{7.141}$$

where

$$\mathbf{P}(t) = \int_{t}^{t_1} \Phi^T(\tau, t)\mathbf{R}(\tau)\Phi(\tau, t)\ d\tau + \Phi^T(t_1, t)\mathbf{P}_1\Phi(t_1, t), \tag{7.142}$$

where $\Phi(t, t_0)$ is the transition matrix of the system in Eq. (7.140). Moreover it can be shown by direct differentiation that $\mathbf{P}(t)$ satisfies the equation

$$-\dot{\mathbf{P}}(t) = \mathbf{A}^T(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A}(t) + \mathbf{R}(t) \tag{7.143}$$

with the terminal condition that $\mathbf{P}(t_1) = \mathbf{P}_1$.

When $t_0$ « $t_1 \to \infty$ and the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{V}$ and $\mathbf{R}$ are constant, Eq. (7.142) reduces to

$$P(t) = \int_{t}^{t_1} e^{\mathbf{A}^T(\tau-t)}\mathbf{R}e^{A(\tau-t)}d\tau + e^{\mathbf{A}^T(\tau-t)}\mathbf{P}_1e^{A(\tau-t)}. \tag{7.144}$$

In the limit when $t_1 \to \infty$ if $\mathbf{A}$ is stable one finds that

$$\lim_{t_1\to\infty} \mathbf{P}(t) = \mathbf{P}_\infty = \int_{t}^{\infty} e^{\mathbf{A}^T(\tau-t)}\mathbf{R}e^{A(\tau-t)}d\tau, \tag{7.145}$$

which is a constant matrix. Since $\mathbf{P}(t)$ satisfies Eq. (7.143) and large values of $t_1$ are being considered, it must be so that

$$\mathbf{0} = \mathbf{A}^T\mathbf{P}_\infty + \mathbf{P}_\infty\mathbf{A} + \mathbf{R} \tag{7.146}$$

which is the time independent Lyapunov equation. This algebraic equation has a unique solution and it is clear that for large $t_1$,

$$\begin{aligned} J &= E\left\{ \int\limits_{t_1}^{t_0} \mathbf{x}^T(t)\mathbf{R}(t)\mathbf{x}(t)dt + \mathbf{x}^T(t_1)\mathbf{P}(t)\mathbf{x}(t_1) \right\} \\ &= tr[\mathbf{P}_\infty \mathbf{Q}_0 + (t_1 - t_0)\mathbf{B}\mathbf{V}\mathbf{B}^T\mathbf{P}_\infty]. \end{aligned} \tag{7.147}$$

***Example 7.9.*** **Performance Index of a First Order Filter**

In Example 6.28 a low pass filter was considered. The index above for this simple system if it is driven by a white noise source with intensity $V$ and given that $R = 1$ and $Q_0 = 0$ is,

$$tr[BVB^T P_\infty] = B^2 V P_\infty = B^2 V\left(-\frac{1}{2A}\right) \tag{7.148}$$

$$= -\frac{1}{\tau^2}\frac{V}{2\left(-\frac{1}{\tau}\right)} = \frac{V}{2\tau} = \frac{\omega_0}{2}V, \tag{7.149}$$

which is the same result which was obtained in the earlier example. The corresponding index is

$$J = \int\limits_{t_0}^{t_1} x^T R x \; d\tau = \frac{\omega_0}{2}V(t_1 - t_0). \tag{7.150}$$

❒

## 7.6 Separation Theorem

Earlier is has been shown that if a full state feedback loop is combined with a full order observer then the two systems can be designed independently of each other. If this is done with a an optimal full state feedback system and a Kalman filter then an even more favorable situation obtains. In this case it can be shown that the system then becomes optimal overall, both with respect to the feedback system and the observer. This is fact known as the separation theorem. It is also known as the certainty equivalence principle.

Consider the index for a LQR regulator. It can be written

$$J = \frac{1}{2}\mathbf{x}^T(t_1)\mathbf{S}_1(t_1)\mathbf{x}(t_1) + \frac{1}{2}\int\limits_{t_0}^{t_1}[\mathbf{x}^T(\tau)\mathbf{R}_1(\tau)\mathbf{x}(\tau) + \mathbf{u}^T(\tau)\mathbf{R}_2(\tau)\mathbf{u}(\tau)]d\tau, \tag{7.151}$$

where $\mathbf{x}(t)$ is the controlled variable, $\mathbf{R}_1(t) \geq 0$, $\mathbf{R}_2(t) > 0$ and $\mathbf{S}(t) \geq 0$ are weighting matrices. Let $\hat{\mathbf{x}}(t)$ be the state estimate and calculate

$$\begin{aligned}E\{\mathbf{x}^T(t)\mathbf{R}_1(t)\mathbf{x}(t)\} &= \Big\{[\mathbf{x}(t)-\hat{\mathbf{x}}(t)+\hat{\mathbf{x}}(t)]^T\mathbf{R}_1(t)[\mathbf{x}(t)-\hat{\mathbf{x}}(t)+\hat{\mathbf{x}}(t)]\Big\}\\ &= E\Big\{[\mathbf{x}(t)-\hat{\mathbf{x}}(t)]^T\mathbf{R}_1(t)[\mathbf{x}(t)-\hat{\mathbf{x}}(t)]\Big\}\\ &\quad + 2E\Big\{[\mathbf{x}(t)-\hat{\mathbf{x}}(t)]^T\mathbf{R}_1(t)\hat{\mathbf{x}}(t)\Big\} + E\Big\{\hat{\mathbf{x}}^T(t)\mathbf{R}_1(t)\hat{\mathbf{x}}(t)\Big\}.\end{aligned} \tag{7.152}$$

From example 6.24 it is known that

$$E\Big\{[\mathbf{x}(t)-\hat{\mathbf{x}}(t)]^T\mathbf{R}_1(t)[\mathbf{x}(t)-\hat{\mathbf{x}}(t)]\Big\} = tr[\mathbf{R}_1(t)\mathbf{Q}(t)], \tag{7.153}$$

where $\mathbf{Q}(t)$ is the variance of the estimation error. Moreover the cross product terms involving the estimation error, $\mathbf{e}(t)$, and the state estimates are uncorrelated. Thus it is possible to write

$$E\{\mathbf{x}^T(t_1)\mathbf{S}(t_1)\mathbf{x}(t_1)\} = tr[\mathbf{S}(t_1)\mathbf{Q}(t_1)] + E\{\hat{\mathbf{x}}^T(t)\mathbf{S}(t_1)\hat{\mathbf{x}}(t)\}, \tag{7.154}$$

$$E\{\mathbf{x}^T(t_1)\mathbf{R}_1(t_1)\mathbf{x}(t_1)\} = tr[\mathbf{R}_1(t_1)\mathbf{Q}(t_1)] + E\{\hat{\mathbf{x}}^T(t)\mathbf{R}_1(t)\hat{\mathbf{x}}(t)\}. \tag{7.155}$$

Using these expressions, the index can be written

$$\begin{aligned}J = E\Bigg\{\int_0^{t_1}\big[E\{\hat{\mathbf{x}}^T(\tau)\mathbf{R}_1(\tau)\hat{\mathbf{x}}(\tau)\} + \mathbf{u}^T(\tau)\mathbf{R}_2(\tau)\mathbf{u}(\tau)\big]d\tau\Bigg\}\\ + E\{\hat{\mathbf{x}}^T(t_1)\mathbf{S}(t_1)\hat{\mathbf{x}}(t_1)\}\\ + tr\left[\int_{t_0}^{t_1}\mathbf{R}_1(\tau)\mathbf{Q}(\tau)d\tau + \mathbf{S}(t_1)\mathbf{Q}(t_1)\right].\end{aligned} \tag{7.156}$$

In this equation it can be seen that the last two terms are independent of the control applied to the system. It is also known that for the Kalman filter equation the innovations are a zero mean white noise process with an intensity $\mathbf{V}_2(t)$, corresponding to the measurement noise alone. Thus the problem of minimizing the criterion or index above is a stochastic optimal regulator problem where the complete state is observed.

It can be shown that the solution to the stochastic optimal regulator problem is to use a control which is,

$$\mathbf{u}(t) = -\mathbf{K}(t)\mathbf{x}(t), \tag{7.157}$$

with $\mathbf{K}(t) = \mathbf{R}_2^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t)$, where $\mathbf{P}(t)$ is the solution of the matrix Riccati equation,

$$-\dot{\mathbf{P}}(t) = \mathbf{R}_1(t) - \mathbf{P}(t)\mathbf{B}(t)\mathbf{R}_2^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t) + \mathbf{A}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A}^T(t), \qquad (7.158)$$

with the terminal condition $\mathbf{S}(t_1) = \mathbf{S}_1$. The minimum value of the criterion is in this case given by

$$J_{min} = tr\left[\mathbf{S}(t_1)\mathbf{Q}(t_1) + \int_{t_0}^{t_1} \mathbf{R}_1(\tau)\mathbf{Q}(\tau)d\tau\right]. \qquad (7.159)$$

Because of this, the overall optimal linear solution to the state feedback stochastic regulator problem is given with the linear control,

$$\mathbf{u}(t) = -\mathbf{K}(t)\hat{\mathbf{x}}(t), \qquad (7.160)$$

where $\mathbf{K}(t)$ is as given in Eq. (7.157). This is also the solution of the deterministic optimal regulator problem. A similar result can be shown to be true in the case of correlated state and measurement noise sources. The linear control which is obtained above is the same that would be found if there were no disturbances and if the state vector was known exactly: there is effectively no uncertainty in the feedback law. This explains the name certainty equivalence principle.

The solution above is the best linear solution. It can be shown that if the noise processes are Gaussian, it is the optimal linear solution, without qualification. There are other forms of the principle which are even less restrictive than expressed here. A block diagram of an LQR control system is shown in Fig. 7.9.

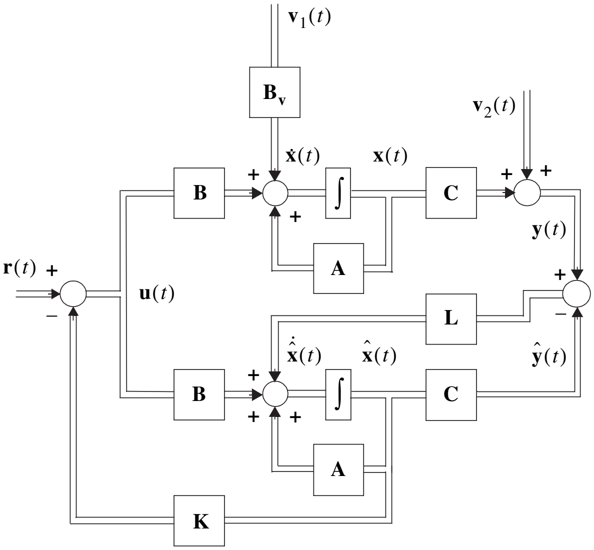


**Fig. 7.9** Block diagram of an LQG control system. The portion of the block diagram of the Kalman filter which solves the Riccati equation has been dropped for the sake of simplicity here

### 7.6.1 Evaluation of the Continuous LQG Index

As stated above, when a feedback control system is composed of a LQR regulator and a Kalman filter it is called an LQG regulator. It is often desirable to evaluate the performance index of such a system with the given noise inputs. This can be accomplished by calculating the noise variances of the state estimates and the estimate errors.

Consider first a system which is driven by white noise which has an intensity $\mathbf{V}_1(t)$ on which noise corrupted measurements are made (the measurement noise is white and has an intensity $\mathbf{V}_2(t)$,

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{v}_1(t),\ t \geq t_0, \tag{7.161}$$

where $\mathbf{x}(t_0) = \mathbf{x}_0$ is a stochastic vector with mean $\hat{\mathbf{x}}_0$ and variance $\mathbf{Q}_0$. The measurement variable is given by the expression,

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{v}_2(t),\ t \geq t_0. \tag{7.162}$$

The overall noise in the system is then has an intensity which is

$$\mathbf{I}_V = \begin{bmatrix} \mathbf{V}_1(t) & \mathbf{V}_{12}(t) \\ \mathbf{V}_{21}(t) & \mathbf{V}_2(t) \end{bmatrix}, t \geq t_0, \tag{7.163}$$

where obviously $\mathbf{V}_{21}(t) = \mathbf{V}_{12}^T(t)$.

An interconnection of a LQR regulator and a Kalman filter can be made by using the state estimates of the Kalman filter in the LQR regulator in such a way that

$$\mathbf{u}(t) = -\mathbf{K}(t)\hat{\mathbf{x}}(t), \tag{7.164}$$

where the Kalman state estimates are found as the solution of the equation,

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{A}(t)\hat{\mathbf{x}}(t) + \mathbf{B}(t)\mathbf{u}(t) + \mathbf{L}[\mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t)], \tag{7.165}$$

where the Kalman gain is found as

$$\mathbf{L}(t) = \mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t), \tag{7.166}$$

by solving the relevant Riccati equation (Eq. (7.16) with $\mathbf{B}_v(t) = \mathbf{I}$ and $\mathbf{Q}(t_0) = \mathbf{Q}_0$). The feedback gain matrix is obtained as

$$\mathbf{K}(t) = \mathbf{R}_2^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t), \tag{7.167}$$

where $\mathbf{P}(t)$ is found as the solution of the LQR Riccati equation with $\mathbf{P}(t_1) = \mathbf{P}_1$. Consider now the vector $[\hat{\mathbf{x}}(t),\ \mathbf{e}(t)] = [\hat{\mathbf{x}}(t),\ \mathbf{x}(t) - \hat{\mathbf{x}}(t)]$. Because the state feedback and the Kalman filter are interconnected the overall system is described by the state equation,

$$\begin{bmatrix} \dot{\mathbf{e}}(t) \\ \dot{\hat{\mathbf{x}}}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(t) - \mathbf{B}(t)\mathbf{L}(t) & \mathbf{0} \\ \mathbf{L}(t)\mathbf{C}(t) & \mathbf{A}(t) - \mathbf{K}(t)\mathbf{C}(t) \end{bmatrix} \begin{bmatrix} \mathbf{e}(t) \\ \hat{\mathbf{x}}(t) \end{bmatrix} + \begin{bmatrix} \mathbf{I} & -\mathbf{L}(t) \\ \mathbf{0} & \mathbf{L}(t) \end{bmatrix} \begin{bmatrix} \mathbf{v}_1(t) \\ \mathbf{v}_2(t) \end{bmatrix}, \tag{7.168}$$

with the initial condition that $[\mathbf{e}(t_0),\ \hat{\mathbf{x}}(t_0)]^T = [\mathbf{x}(t_0) - \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_0]^T$. To find the noise contributions of the error and state estimates the variance matrix of the system in Eq. (7.168) has to be found. This can be accomplished by direct calculation. If,

$$\begin{aligned} &\begin{bmatrix} \mathbf{Q}_{11}(t) & \mathbf{Q}_{12}(t) \\ \mathbf{Q}_{12}^T(t) & \mathbf{Q}_{22}(t) \end{bmatrix} \\ &= E\left\{ \begin{bmatrix} \mathbf{e}(t) - E\{\mathbf{e}(t)\} \\ \hat{\mathbf{x}}(t) - E\{\hat{\mathbf{x}}(t)\} \end{bmatrix} [\mathbf{e}(t) - E\{\mathbf{e}(t)\}]^T [\hat{\mathbf{x}}(t) - E\{\hat{\mathbf{x}}(t)\}]^T \right\}, \end{aligned} \tag{7.169}$$

then this is equivalent because of symmetry to three coupled matrix differential equations.

$$\begin{aligned} \dot{\mathbf{Q}}_{11}(t) = [\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]\mathbf{Q}_{11}(t) + \mathbf{Q}_{11}(t)[\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]^T \\ + \mathbf{V}_1(t) + \mathbf{L}(t)\mathbf{V}_2(t)\mathbf{L}^T(t), \end{aligned} \tag{7.170}$$

$$\begin{aligned} \dot{\mathbf{Q}}_{12}(t) = \ \mathbf{Q}_{11}(t)\mathbf{C}^T(t)\mathbf{L}^T(t) + \mathbf{Q}_{12}(t)[\mathbf{A}(t) - \mathbf{B}(t)\mathbf{K}(t)]^T \\ + [\mathbf{A}(t) - \mathbf{B}(t)\mathbf{L}(t)]\mathbf{Q}_{12}(t) - \mathbf{L}(t)\mathbf{V}_2(t)\mathbf{L}^T(t), \end{aligned} \tag{7.171}$$

$$\begin{aligned} \dot{\mathbf{Q}}_{22}(t) = \mathbf{Q}_{12}^T(t)\mathbf{C}^T(t)\mathbf{L}^T(t) + \mathbf{Q}_{22}(t)[\mathbf{A}(t) - \mathbf{B}(t)\mathbf{K}(t)]^T + \mathbf{L}(t)\mathbf{C}(t)\mathbf{Q}_{12}(t) \\ + [\mathbf{A}(t) - \mathbf{B}(t)\mathbf{K}(t)]Q_{22}(t) + \mathbf{L}(t)\mathbf{V}_2(t)\mathbf{L}^T(t), \end{aligned} \tag{7.172}$$

subject to the initial conditions $\mathbf{Q}_{11}(t_0) = \mathbf{Q}_0$, $\mathbf{Q}_{12}(t_0) = \mathbf{0}$, and $\mathbf{Q}_{22}(t_0) = \mathbf{0}$. If Eq. (7.170) is rewritten then it becomes

$$\begin{aligned} \dot{\mathbf{Q}}_{11} = \mathbf{A}(t)\mathbf{Q}_{11}(t) + \mathbf{Q}_{11}(t)\mathbf{A}^T(t) - \mathbf{L}(t)\mathbf{C}(t)\mathbf{Q}_{11}(t) - \mathbf{Q}_{11}(t)\mathbf{C}^T(t)\mathbf{L}^T(t) \\ + \mathbf{V}_1(t) + \mathbf{L}(t)\mathbf{V}_2(t)\mathbf{L}^T(t). \end{aligned} \tag{7.173}$$

The quantity $\mathbf{Q}_{11}$ is the variance of the state estimation error. Using the Kalman gain from Eq. (7.166), the last four terms on the right in this equation reduce to

$$\mathbf{V}_1(t) - \mathbf{Q}_{11}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t)\mathbf{C}(t)\mathbf{Q}_{11}(t), \tag{7.174}$$

using the Kalman gain from Eq. (7.15). But this means that Eq. (7.173) is just the Riccati equation used to determine the optimal observer feedback or Kalman gain for the system, which is of course satisfied. Using this result in Eq. (7.171) results in the expression,

$$\dot{\mathbf{Q}}_{12}(t) = \mathbf{Q}_{12}(t)[\mathbf{A}(t) - \mathbf{B}(t)\mathbf{K}(t)]^T + [\mathbf{A}(t) - \mathbf{B}(t)\mathbf{L}(t)]\mathbf{Q}_{12}(t), \tag{7.175}$$

which is homogeneous. Given the initial condition that $\mathbf{Q}_{12}(t_0) = \mathbf{0}$ it can immediately be seen that

$$\mathbf{Q}_{12}(t) = \mathbf{0}, \ \ t \geq t_0, \tag{7.176}$$

which shows that the estimation error, $\mathbf{e}(t)$, and the state estimate, $\hat{\mathbf{x}}(t)$, are uncorrelated.

The quantity $\mathbf{Q}_{22}$ is the variance of the state estimate and it is given by the differential equation (7.172) above with the correct expression for the Kalman gain inserted. With this substition it reduces to the equation,

$$\begin{aligned}\dot{\mathbf{Q}}_{22}(t) = [\mathbf{A}(t) - \mathbf{B}(t)\mathbf{K}(t)]\mathbf{Q}_{22}(t) + \mathbf{Q}_{22}(t)[\mathbf{A}(t) - \mathbf{B}(t)\mathbf{K}(t)]^T \\ + \mathbf{L}(t)\mathbf{V}_2(t)\mathbf{L}^T(t),\end{aligned} \tag{7.177}$$

with the initial condition that $\mathbf{Q}_{22}(t_0) = \mathbf{0}$.

Because the state is equal to it's estimate plus the error, $\mathbf{x}(t) = \hat{\mathbf{x}}(t) + \mathbf{e}(t)$

$$E\{\mathbf{x}(t)\mathbf{x}^T(t)\} = E\left\{\hat{\mathbf{x}}(t)\hat{\mathbf{x}}^T(t) + \hat{\mathbf{x}}(t)\mathbf{e}^T(t) + \mathbf{e}(t)\hat{\mathbf{x}}(t) + \mathbf{e}(t)\mathbf{e}^T(t)\right\} \tag{7.178}$$

and the error and state estimates are uncorrelated, it is clear that,

$$\begin{aligned}E\{\mathbf{x}^T(t)\mathbf{W}_x\mathbf{x}(t)\} &= tr[\mathbf{W}_x E\{\mathbf{x}(t)\mathbf{x}^T(t)\}] \\ &= tr[\mathbf{W}_x\{\mathbf{Q}_{22}(t) + \bar{\mathbf{x}}(t)\bar{\mathbf{x}}^T(t) + \mathbf{Q}_{11}(t)\}],\end{aligned} \tag{7.179}$$

where $\mathbf{W}_x$ is some weighting matrix used to select the relevant states. In the same way it is possible to find the variance of (or power in) the input variable(s) because of the way the LQG feedback is defined in Eq. (7.164):

$$\begin{aligned}E\{\mathbf{u}^T(t)\mathbf{W}_u\mathbf{u}(t)\} &= E\left\{\hat{\mathbf{x}}^T(t)\mathbf{K}^T(t)\mathbf{W}_u\mathbf{K}(t)\hat{\mathbf{x}}(t)\right\} \\ &= tr[\mathrm{K}^T(t)\mathbf{W}_u\mathbf{K}(t)\mathrm{E}\{\hat{\mathbf{x}}(t)\hat{\mathbf{x}}^T(t)\}] \\ &= tr[\mathbf{K}^T(t)\mathbf{W}_u\mathbf{K}(t)\{Q_{22}(t) + \bar{\mathbf{x}}(t)\bar{\mathbf{x}}^T(t)\}],\end{aligned} \tag{7.180}$$

where $\mathbf{W}_u$ is again a (input) weighting matrix. Equations (7.179) and (7.180) make it possible to find the power in the states as well as the input power once the Kalman gain and LQR gain matrices are known. This is obviously useful for practical system design. Equation (7.173) gives the power in the estimation error. Thus it is possible to find the value of the indexes used to design the LQR regulator.

Typically an index or criteria for an LQG regulator is expressed in the form

$$J(t_1) = E\left\{ \int_{t_0}^{t_1} [\mathbf{x}^T(\tau)\mathbf{R}_1(\tau)\mathbf{x}(\tau) + \mathbf{u}^T(\tau)\mathbf{R}_2(\tau)\mathbf{u}(\tau)]\; d\tau + \mathbf{x}^T(t_1)\mathbf{P}_1\mathbf{x}(t_1) \right\}, \tag{7.181}$$

which is to be minimized over a time horizon $t_0 \le t \le t_1$ where $\mathbf{R}_1(t) > \mathbf{0}$, $\mathbf{R}_2(t) > \mathbf{0}$ and $\mathbf{P}_1 \ge \mathbf{0}$ are symmetric weighting matrices. Using the results derived above the minimum value of this performance index can be shown to be either of the expressions below:

$$\begin{aligned} J(t_1) &= \bar{\mathbf{x}}^T(t_0)\mathbf{P}(t_0)\bar{\mathbf{x}}(t_0) \\ &\quad + tr\left[ \int_{t_0}^{t_1} [\mathbf{P}(\tau)\mathbf{L}(\tau)\mathbf{V}_2(\tau)\mathbf{L}^T(\tau) + \mathbf{Q}(\tau)\mathbf{R}_1(\tau)]\; d\tau + \mathbf{P}_1\mathbf{Q}(t_1) \right] \\ &= \bar{\mathbf{x}}^T(t_0)\mathbf{P}(t_0)\bar{\mathbf{x}}(t_0) \\ &\quad + tr\left[ \mathbf{P}_0\mathbf{Q}(t_0) + \int_{t_0}^{t_1} [\mathbf{P}(\tau)\mathbf{V}_1(\tau) + \mathbf{Q}(\tau)\mathbf{K}(\tau)\mathbf{R}_1(\tau)\mathbf{K}^T(\tau)]d\tau \right], \end{aligned} \tag{7.182}$$

where $\mathbf{P}(t)$ and $\mathbf{Q}(t)$ are the solutions to the LQR and Kalman filter Riccati equations respectively. This value of the index will be obtained if the assumptions behind the derivation of the LQR and Kalman gains are fulfilled.

In the special case of a time invariant system in the stationary state these expressions reduce to:

$$J_\infty = E\{\mathbf{x}^T(t)\mathbf{R}_1 x(t) + \mathbf{u}^T(t)\mathbf{R}_2\mathbf{u}(t)\} \tag{7.183}$$

$$= tr[\mathbf{P}_\infty\mathbf{L}_\infty\mathbf{V}_2\mathbf{L}_\infty^T + \mathbf{Q}_\infty\mathbf{R}_1] \tag{7.184}$$

$$= tr[\mathbf{P}_\infty\mathbf{V}_1 + \mathbf{Q}_\infty\mathbf{K}_\infty^T\mathbf{R}_2\mathbf{K}_\infty], \tag{7.185}$$

where the infinity subscripts mean that these are the stationary variables of these variables.

It should be pointed out here that certainty equivalence applies to a large set of linear dynamic systems with different noise sources apart from the simple systems treated in this book. It also applies to systems with non-Gaussian state and measurement noise, as well as those with correlated state and measurement noise. In addition it applies to systems which have colored (filtered) Gaussian state and measurement noise.

The results above can be shown using the methods discussed at the beginning of Sect. 7.5. For a more detailed treatment see Kwakernaak and Sivan (1972) and Friedland (1987).

***Example 7.10.*** **LQG Controller for the Hydraulic Servo**

Consider the problem of combining the Kalman filter of Example 7.2 with an LQR regulator. This regulator will be designed here. If the damping of the original model is assumed to be small, the hydraulic servo becomes a harmonic oscillator. The dynamic and input matrices of the system are:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ \bar{\omega}_0^2 \end{bmatrix}, \tag{7.186}$$

where $\bar{\omega}_0^2 = 2\beta A_c/(VM) = 1400$ and $\omega_0^2 = 2\beta A_c^2/(VM) = 2.1 \times 10^5$, using the units of the example. The output matrix of the system is $\mathbf{C} = [\,1 \quad 0\,]$. The state noise scaling matrix will be assumed to be $\mathbf{I}$ for simplicity here.

First it must be shown that the approximate model of the equation above yields nearly the same eigenfrequencies predicted by the earlier linearized model. Using units of the example the eigenfrequency of the approximate model is $\omega_0 = 458.26$ while that of the linearized model is $\omega_0 = 458.31$ (no units are given here because of the unusual mix of units used in the example). The input matrix element of the approximate model is $\bar{\omega}_0 = 37.42$. The linearized model has an input matrix element which is 46.67. The approximate model's parameters are thus fairly close to those of the linearized model.

The LQR regulator Riccati equation for this second order system is:

$$0 = \mathbf{R}_1 - \mathbf{P}\mathbf{B}\mathbf{R}_2^{-1}\mathbf{B}^T\mathbf{P} + \mathbf{P}\mathbf{A} + \mathbf{A}^T\mathbf{P} \quad \text{or}$$

$$\begin{aligned} 0 = {} & \begin{bmatrix} R_{111} & 0 \\ 0 & R_{122} \end{bmatrix} - \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} 0 \\ \bar{\omega}_0^2 \end{bmatrix} \frac{1}{R_2} \begin{bmatrix} 0 & \bar{\omega}_0^2 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \\ & + \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -\omega_0^2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}. \end{aligned} \tag{7.187}$$

With the weighting factors, the following coupled set of nonlinear second order equations results:

$$\begin{aligned} 0 &= R_{111} - \frac{\bar{\omega}_0^4}{R_2} p_{12}^2 - 2\omega_0^2 p_{12}, \\ 0 &= -\frac{\bar{\omega}_0^4}{R_2} p_{11} p_{22} - \omega_0^2 p_{22} + p_{11}, \\ 0 &= R_{122} - \frac{\bar{\omega}_0^4}{R_2} p_{22}^2 + 2p_{12}. \end{aligned} \tag{7.188}$$

These equations can be solved for the matrix elements of the $\mathbf{P}$ matrix:

$$p_{12} = R_2\left(\frac{\omega_0^2}{\bar{\omega}_0^4} + \frac{1}{\bar{\omega}_0^4}\sqrt{\frac{\omega_0^2}{\bar{\omega}_0^4} + \frac{R_{111}}{R_2}}\right),$$

$$p_{22} = \frac{\sqrt{R_2}}{\bar{\omega}_0^2}\sqrt{R_{122} + 2p_{12}}, \tag{7.189}$$

$$p_{11} = \left(\omega_0^2 + \frac{\bar{\omega}_0^4}{R_2}p_{11}\right)p_{22}.$$

The LQR state feedback matrix is then

$$\mathbf{K} = \mathbf{R}_2^{-1}\mathbf{B}^T\mathbf{P} = \frac{1}{R_2}\begin{bmatrix} 0 & \bar{\omega}_0^{-2} \end{bmatrix}\begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} = \frac{\bar{\omega}_0^2}{R_2}[p_{12} \quad p_{22}]. \tag{7.190}$$

Now the units of Example 3.26 can be used to find the optimal LQR gain for the hydraulic servo with the weighting matrices, $\mathbf{R}_1 = diag\begin{bmatrix} 10^5 & 0.11 \end{bmatrix}$ and $R_2 = 0.4$ This unusual selection of weights is necessary because the control object is very powerful and at the same time small. This means that large forces are involved and this makes the system very fast. This is shown by the resonance frequency of the bare control object, $\omega_0 = 458.26$. It is desired that the closed loop system be about twice this fast, well damped and have complex eigenfrequencies: this is the reason for the selection of the weight matrices above.

With the design specification above, the gain of the LQR regulator is found to be $\mathbf{K} = \begin{bmatrix} 8.612 \cdot 10^2 & 1.526 \end{bmatrix}$ giving closed loop eigenfrequencies of $\lambda_c = (-1.0685 \mp j0.5233) \cdot 10^3$. These gains and eigenfrequencies were obtained using both the matrix elements of the P matrix in equations (7.189–7.190) and the MATLAB routine `lqr: [K,S,Ereg] = lqr(A,B,R1,R2,0).`

To design the Kalman filter the state and measurement noise intensities must be specified. Here the noise intensities $\mathbf{V}_1 = diag\begin{bmatrix} 1 \cdot 10^5 & 500 \end{bmatrix}$ and $V_2 = 1$ were selected. This gave the Kalman gain of $\mathbf{L} = \begin{bmatrix} 3.162 \cdot 10^2 & 1.527 \end{bmatrix}^T$. Again the numerical qualities of the model gave the large asymmetry of the different gains. The eigenfrequencies of the Kalman filter were determined to be $\lambda_o = (-1.581 \pm j4.301) \cdot 10^2$. These results were found using either the MATLAB routine, `lqe: [L,P,Eest] = lqe(A,G,C,V1,V2,N), G = eye2, N = [0 0]'`, or equations (7.38) and (7.39) above. The two results were nearly identical, differences were due to the numerical accuracy of the calculation.

The Kalman filter estimation error was found with the MATLAB routine `lqe` and the equations above the results using both methods gave

$$\mathbf{Q}_{11} = \begin{bmatrix} 3.163 \cdot 10^{-5} & 0 \\ 0 & 6.641 \end{bmatrix} \cdot 10^7.$$

This result was confirmed emprically using Simulink simulation. In the same way $\mathbf{Q}_{22}$ was calculated mathematically using the two different methods, with nearly equal results, namely

$$\mathbf{Q}_{22} = \begin{bmatrix} 9.888 \cdot 10^{-6} & 0.005 \\ 0.005 & 3.312 \end{bmatrix} \cdot 10^{7}$$

Simulation proved less reliable. The results using Simulink were incorrect with a factor of two. The difficulty is the large spread in the numbers in the problem and the difficulties involved in integrating numerically the white noise in the dynamic system in Matlab/Simulink. ❐

### 7.6.2 Evaluation of the Discrete LQG Index

The evaluation of the index for a discrete time LQG system can be carried out in the same way as for the analogous continuous time system. The structure of discrete time controllers and Kalman filters and their equation representation has already been reviewed in detail above so there is no reason to repeat these here. What must be accomplished here is to find the discrete equivalents to equations (7.184) and (7.185).

As before the focus is on finding the equations which describe the variances of the state estimation error and the state estimates. The state estimate error is defined by

$$\mathbf{e}(k) = \mathbf{x}(k) - \hat{\mathbf{x}}(k). \tag{7.191}$$

The discrete version of Eq. (7.168) is:

$$\begin{bmatrix} \mathbf{e}(k+1) \\ \hat{\mathbf{x}}(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(k) - \mathbf{L}(k)\mathbf{C}(k) & \mathbf{0} \\ \mathbf{K}(k)\mathbf{C}(k) & \mathbf{A}(k) - \mathbf{B}(k)\mathbf{K}(k) \end{bmatrix} \begin{bmatrix} \mathbf{e}(k) \\ \hat{x}(k) \end{bmatrix} + \begin{bmatrix} \mathbf{I} & -\mathbf{K}(k) \\ 0 & \mathbf{K}(k) \end{bmatrix} \begin{bmatrix} \mathbf{v}_1(k) \\ \mathbf{v}_2(k) \end{bmatrix}, \tag{7.192}$$

which are to be evaluated with the initial conditions, $[\mathbf{e}(k_0), \hat{\mathbf{x}}(k_0)]^T = [\mathbf{x}(t_0) - \hat{\mathbf{x}}_0, \hat{\mathbf{x}}_0]^T$. The matrix which gives the variance matrix of $[\mathbf{e}(k), \hat{\mathbf{x}}(k)]^T$ is

$$\begin{bmatrix} \mathbf{Q}_{11}(k) & \mathbf{Q}_{12}(k) \\ \mathbf{Q}_{12}^T(k) & \mathbf{Q}_{22}(k) \end{bmatrix} = E\left\{ \begin{bmatrix} \mathbf{e}(k) - E\{\mathbf{e}(k)\} \\ \hat{\mathbf{x}}(k) - E\{\hat{\mathbf{x}}(k)\} \end{bmatrix} [\mathbf{e}(k) - E\{\mathbf{e}(k)\}]^T [\hat{\mathbf{x}}(k) - E\{\hat{\mathbf{x}}(k)\}]^T \right\}, k \geq k_0. \tag{7.193}$$

As in the continuous case $\mathbf{Q}_{11}(k)$ is the variance of the state error which is calculated with the Kalman filter design equation, (7.107). $\mathbf{Q}_{12}(k) = \mathbf{0}$ is zero as the estimation error and the state estimate are not correlated with each other as for the continuous case. $\mathbf{Q}_{22}(k)$ has to be found by evaluating the difference equation,

$$\begin{aligned}\mathbf{Q}_{22}(k+1) = {} & [\mathbf{A}(k) - \mathbf{B}(k)\mathbf{K}(k)]\mathbf{Q}_{22}(k)[\mathbf{A}(k) - \mathbf{B}(k)\mathbf{K}(k)]^T \\ & + \mathbf{K}(k)[\mathbf{C}(k)\mathbf{Q}_{11}(k)\mathbf{C}^T(k) + \mathbf{V}_2(k)]\mathbf{K}^T(k),\end{aligned} \tag{7.194}$$

with the initial condition, $\mathbf{Q}_{22}(k_0) = \mathbf{0}$. The stationary state value of this variance can of course be obtained by setting the equation above equal to $\mathbf{Q}_{22}(k)$ and solving the resulting equation for $\mathbf{Q}_{22}$.

The value of the index can be written in the time varying case:

$$\begin{aligned}J(k) = {} & \bar{\mathbf{x}}_0^T\mathbf{P}(k_0)\bar{\mathbf{x}}_0 + tr\{\mathbf{P}_1\mathbf{Q}_{11}(k_1)\} \\ & + \sum_{k=k_0}^{k_1-1} tr\{\mathbf{R}_1(k+1)\mathbf{Q}_{11}(k+1) + [\mathbf{P}(k+1) + \mathbf{R}_1(k+1)]\mathbf{L}(k) \\ & \times [\mathbf{C}(k)\mathbf{Q}_{11}(k)\mathbf{C}^T(k) + \mathbf{V}_2(k)]\mathbf{L}^T(k)\},\end{aligned} \tag{7.195}$$

which can also be written in an alternative form:

$$\begin{aligned}J(k) = {} & \bar{\mathbf{x}}_0^T\mathbf{P}(k_0)\bar{\mathbf{x}}_0 + tr\{\mathbf{P}_0\mathbf{Q}_1 1(k_0)\} \\ & + \sum_{k=k_0}^{k_1-1} tr\{[\mathbf{R}_1(k+1) + \mathbf{P}(k+1)]\mathbf{V}_1(k) + \mathbf{Q}_{11}(k)\mathbf{K}(k) \\ & \times \{\mathbf{R}_2(k) + B^T(k)[\mathbf{R}_1(k+1) + \mathrm{P}(k+1)]\mathbf{B}(k)\}\mathbf{B}(k)\mathbf{K}(k)\}.\end{aligned} \tag{7.196}$$

For a time invarient system the index is

$$\mathbf{J}_\infty = \lim_{k_0 \to -\infty} E\{\mathbf{x}^T(k+1)\mathrm{R}_1(k+1)\mathbf{x}(k+1) + \mathbf{u}^T(k)\mathbf{R}_2\mathbf{u}(k)\} \tag{7.197}$$

$$= tr\left\{\mathbf{R}_1\mathbf{Q}_{11\infty} + (\mathbf{P}_\infty + \mathbf{R}_1)\mathbf{L}_\infty(\mathbf{C}\mathbf{Q}_{11\infty}\mathbf{C}^T + \mathbf{V}_2)\mathbf{L}_\infty^T\right\} \tag{7.198}$$

$$= tr\{(\mathbf{R}_1 + \mathbf{P}_\infty)\mathbf{V}_1 + \mathbf{Q}_{11\infty}\mathbf{K}_\infty^T[\mathbf{R}_2 + \mathbf{B}^T(\mathbf{R}_1 + \mathbf{P}_\infty)\mathbf{B}]\mathbf{K}_\infty\}. \tag{7.199}$$

In the last three equations above with the infinity subscripts of course mean that these are the steady state values of these variables. For a more detailed treatment see Kwakernaak and Sivan (1972) and Stengel (1986).

## 7.7 Summary

The main subject of this chapter is the derivation of Kalman filters both for continuous time and for discrete time linear systems. A secondary aim is to show that stochastic optimal observers can with advantage be used for the construction of full state feedback systems which are optimal in a wider sense than would be the case if only pole placement or LQR design methods are used.

The advantage of Kalman filters in relation to ordinary observers is that they can be shown to be optimal in the sense that they have the best performance of any other observer, linear or nonlinear for a given system. Moreover the design procedure is constructive: given models for the system and the noise sources, a Kalman filter automatically emerges from the procedure which is optimal for the system in question. Optimality is here understood to be in the sense of best signal power to noise power ratio. If a suboptimal solution is selected for various subsidiary reasons, a Kalman filter can be used for comparison in order to understand what a compromise costs in terms of overall performance.

Kalman filters for continuous systems have the same structure as the pole placement observers which were designed in Chap. 4. The main difference between Kalman filters and ordinary observers is that they take account of the noise built into the control object both in terms of its internal dynamics and in terms of the noise which must accompany any measurement made on it. This is done by keeping track of mean values of the states and the error covariance of the observer itself, the two moments of the assumed Gaussian system noise. In detail this accomplished by calculating the state noise power which propagates through the control object and comparing this to the noise which is associated with the measurement. This comparison leads to a selection of the observer- or Kalman- gain matrix which is the best compromise between the inherent system noise and the measurement noise. It has been shown that the gain matrix is obtained by solving the time dependent Riccati equation. The Riccati equation emerges naturally from the theory as the time dependent Lyapunov equation appended with a term which embodies the uncertainty reduction due to measurement. This is true both for continuous time and discrete time systems.

Discrete time Kalman filters have a structure which is close to that of continuous Kalman filters and they work on the same principle of state/measurement power noise compromise. But there are some differences which are due to the fact of sampled time operation. In particular the time update of the mean values of the states and the error covariance is separate from the measurement update. One of the positive differences is that there is some flexibility in handling the method in which the measurement data is used, leading to different filter algorithms. A negative difference is that measurements are only available at sample times and this introduces an inavoidable time delay into the system. Moreover for this reason, continuous measurements cannot be used even if they are available. This means that for a given system, a continuous Kalman filter will nearly always have a better performance that a discrete Kalman filter, independent of what sampling time is used. Discrete time Kalman filters (and indeed also discrete time regulators) have however the advantage of being immediately compatible with conventional digital computers and are more easily implemented in them, at least at the present time.

The final section of this chapter shows that when the optimal or LQR regulators of Chap. 6 are combined with Kalman filters a very unusual situation obtains. In part it can be shown that the LQR regulator and Kalman filter can be designed independently of each other and in part what emerges from this construction is nevertheless overall optimal in itself. This separation theorem is the stochastic equivalent of the separation principle presented in Chap. 4 for deterministic systems. It simplifies observer/controller design enormously and is one of the prime reasons why LQG controllers have found such wide use in many different control applications.

## 7.8 Notes

### *7.8.1 Background for Kalman Filtering*

In 1959 Swerling (1959) introduced a unbiased minimum variance estimation algorithm for linear systems similar to that later published by Kalman (1960). It is Kalman's work that drew the greatest attention and thus the name 'Kalman filtering' is applied to this kind of filtering or estimation algorithm.

What Kalman and later Kalman and Bucy did (Kalman and Bucy (1961)) was to produce a recursive algorithm which is optimal in the sense of minimum variance and least square error, and which can at the same time be easily implemented and solved on a digital computer. The optimally of the filter is important as it gives the designer confidence that the system found is the best that can be achieved without tedious experimentation or simulation. The filter is thus very practical to realize for a number of different applications especially since the price of small digital computers continues to fall.

Another way of describing the filter is to say that it circumvents the problem of solving the Wiener-Hopf integral equation (Wiener (1949) and Kolmogrov (1941)) by working with the equivalent differential equation. Then recognizing that no analytic solution is actually required, the differential equation is solved recursively to find the state estimates. The large computational burden incurred in this way is placed on a digital computer (Sorenson and Stubberund (1970)).

Through a parallel line of development due to Stratonovich (1960), it was later found possible to apply the Kalman filter to nonlinear processes without doing violence to its structure or to its basic principles of operation (Schwartz (1970)), though the sense in which it is optimal is changed. Such a filter is called an extended Kalman filter (EKF). There are continuous (CEKF), continuous- (propogation) discrete (measurement) (CDEKF) and discrete (DEKF) versions of such filters for nonlinear systems (Meditch (1969), Jazwinski (1970) and Gelb (1974), Kushner (1967), Maybeck (1979), Safonov and Athers (1977)).

## 7.9 Problems

### *Problem 7.1*

A plant consists of a D.C. motor which has and angular velocity ω(*t*) which is driven by an input voltage $V_a$(t). The system is described by the scalar state equation

$$\dot{\omega}(t) = -\alpha\omega(t) + \beta V_a(t) + \omega_1(t),$$

where α and β are positive constants. $\omega_1(t)$ is due to a torque disturbance which acts on the shaft and it is white noise with a mean value 0 and an intensity $V_1$. The velocity of the D.C. motor is observed with an additive noise component, $\omega_2(t)$, which is white noise with a mean value 0 and a noise intensity $V_2$. $\omega_1(t)$ and $\omega_2(t)$ are uncorrelated. Both $V_1$ and $V_2$ are constants.

a. What is a reasonable measurement model for the system?
b. Design a Kalman filter for the system which is capable of estimating the speed optimally in the steady state. The Kalman gain is to be given in terms of the system parameters and constants.
c. If the signal to noise ratio of the system is poor, what is the optimal observer for the system assuming that the model of it is accurate? What is such a degenerate 'observer' called?
d. If the signal to noise ratio is good (the measurement noise is small relative to the plant noise), how does the Kalman filter make use of this?

### *Problem 7.2*

Discretize the system of Problem 7.1 assuming a sampling time of *T*. Design a Kalman filter for the resulting system which is optimal in the steady state.

a. What are the advantages and disadvantages of this filter with respect to the filter designed in Problem 7.1?

### *Problem 7.3*

The D.C. motor of Problem 7.1 is to be used in an angular position control for an actuator. This means that the state equation in the original problem formulation has to be augmented with an equation which describes the angular position, θ(t), of the system.

a. Write down the overall state equations of the position control system, including the torque disturbance. The state vector of the augmented system is to be $[\theta\ \omega]^T$.
b. Can one design a Kalman filter for this system based on using the measurement of the angular velocity alone assuming steady state operation? Explain.

A more direct measurement of the system performance is desired. For this reason the angular position of the system is to be measured. This measurement is also corrupted with noise. The measurement model is thus

$$y(t) = [1\ 0]\begin{bmatrix} \theta(t) \\ \omega(t) \end{bmatrix} + w_\theta(t),$$

$w_\theta(t)$ is measurement noise and is characterized by the expression $w_\theta(t) \in N(0, V_2)$ where $V_2$ is the noise intensity (constant).

c. Design a Kalman filter for this system based on using the measurement of the position of the system alone. Again only stationary performance of the filter is of interest.
d. Which of the Kalman filters above will give the best overall performance? Why?

***Problem 7.4***

Consider a continuous system which has a state space form

$$\begin{aligned} \dot{x}(t) &= -x(t) + u(t) + v(t), \\ y(t) &= x(t) + z(t), \end{aligned}$$

where $x(t)$ is the system state, $u(t)$ is its control input, $v(t)$ is low pass filtered white noise with zero mean and an intensity which is 1, $y(t)$ is the output of the system and $z(t)$ is low pass filtered white noise. It is assumed that this input comes from the system

$$\dot{z}(t) = -100z(t) + 100\,w(t).$$

Here $w(t)$ is Gaussian, zero mean, white noise with an intensity 0.01. It is desired that $x(t)$ be estimated (as $\hat{x}(t)$) given the output measurement $y(t)$.

a. Find the stationary variance of the estimation error $e(t) = x(t) - \hat{x}(t)$, if the estimate $\hat{x}(t) = y(t)$.

The measurement noise source is now to be approximated with a white noise source which has an intensity which corresponds to the spectral density of $z(t)$ at the frequency $\omega = 0$.

b. Find a full order stationary Kalman filter which minimizes the stationary estimation error's variance given that the approximate noise source above is used.
c. Find the stationary variance of the estimation error when the Kalman filter of question b and the calculation still uses the simplified measurement noise model. Compare the result of question c with that of b. Examine the approximation that white noise is an acceptable approximation. Explain the answer found.
d. Find the exact value of the estimation error's stationary variance when the coloring of the measurement noise is taken into consideration, but where the simplified full order Kalman filter of question b is still used.

***Problem 7.5***

Consider a discrete system which is given in terms of a transfer function,

$$Z\left\{G_{ho}(s)\frac{1}{s}\right\} = \frac{Tz^{-1}}{1 - z^{-1}},$$

where s is the Laplace operator, Z is the Z-transform, $G_{ho}(s)$ is the transfer function of a zeroth order hold network and $T$ is the sampling time.

a. Write down the state equation of the system.

The state of the system has state noise which has an intensity, $V$ (a constant). The state of the system is to be measured with an additive noise source. This noise source has a noise intensity which is $W$ (a constant) and is zero mean.

b. Is the assumption of broad band white noise reasonable under the circumstances? What additional assumptions are necessary to use such a model?
c. Design a Kalman filter for this system to estimate its state.

The Kalman filter is to be used in conjunction with an ordinary state feedback system.

d. Design a pole placement regulator for the system. In what limit kan one find a minimum variance (or deadbeat) regulator?
e. Design an LQR regulator for the system which can minimize an index having the form

$$J = \sum_{i=0}^{\infty}[x^2(i) + \rho u^2(i)],$$

where ρ is a weighting factor.

f. In what limit can one find a minimum variance regulator?
g. When the Kalman filter is used to estimate the state of the system and an LQR regulator is used with it to control the state, what kind of control system is obtained? This is a fairly complex and thus expensive control system. Is there any particular reason then to use it?

***Problem 7.6***

A disc drive positioning actuator is described by the differential equation,

$$ma = K_f I,$$

where $m$ is the moving mass, $a$ is the acceleration of the head, $K_f$ is the force constant and $I$ is the driving current Bell et al. (1984). The position of the head is given by $x$ and the velocity by $v$. Both the position and velocity of the head are subject to disturbances which can be modelled as white noise such that the state noise on $x$ and $v$ are given by $w_x(t) \in N(0, W_x)$ and $w_v(t) \in N(0, W_v)$ respectively.

Because of noise in the disc drive dynamics it is desired to establish an observer to estimate its position. This estimation is to be made on the basis of a noisy position measurement:

$$y(t) = x(t) + w(t),$$

where $w(t)$ is white noise such that $w(t) \in N(0, W)$.

a. Write down the differential equations which describe the disc drive system, including the disturbances.
b. Design a Kalman filter to estimate the position of the head. The Kalman gain should be expressed in terms of the parameters and constants of the system.
c. Design an LQR regulator for this system in order to use the estimates of the Kalman filter which has been found.
d. Under what circumstances would it be advantageous to use a continuous observer/regulator for this system?

### *Problem 7.7*

Assume that there is only state noise on the position component of the system in Problem 7.6. The system is to be sampled with a sampling time $T$.

a. Design a discrete Kalman filter to estimate the position of the head drive in Problem 7.6.
b. Design a regulator for the system which will insure that the overall system performance is optimal.
c. Can the performance of the discrete system be as good as that for a corresponding continuous system? In what limit?
d. What are the advantages of using a discrete observer/control system instead of a continuous one?

### *Problem 7.8*

The height of a hot air ballon with a mass $M$ and a ballon volume $V$ can be controlled with the addition of a heating power $q$. The air in the ballon has a temperature T and a density ρ. Physical effects apart from those which are explicitly mentioned in the exercise are to be ignored.

The equation which controls the air temperature in the ballon is

$$\frac{dT}{dt} = a(T - T_a) + cq.$$

$a$ and $c$ are constants. $T_a$ and $\rho_a$ are the temperature and density of the ambient air and are related to the temperature and density of the air in the ballon by the equation

$$\rho T = \rho_a T_a.$$

If $v$ is the vertical velocity of the ballon then Newton's second law requires that

$$M\frac{dv}{dt} = Vg(\rho_a - \rho) - Mg - bv,$$

where $g$ is the acceleration of gravity and $b$ is the air resistance coefficient. $M$, $V$, $b$ and $g$ are constants while $T$, $T_a$, $\rho$ and $\rho_a$ are variables.

a. Set up the nonlinear state equations for the system, choosing the states as $T$ and $v$, the input as $q$ and the state disturbance as $T_a$.

Assume now that the ballon is hovering with a constant altitude.

b. Find the required stationary heating power, $q_0$, expressed in terms of the stationary values of $T_a$, $\rho_a$ and the system constants.
c. Linearize the system around its stationary state and write down the state equation of the linearized system.
d. Investigate whether or not the linearized system is stable. Investigate also whether or not it is controllable.

In the final portion of the exercise only that portion of the system which is concerned with the temperature will be considered. That is to say the portion of it which has the form:

$$\Delta\dot{T} = A\Delta T + B\Delta q + B_v\Delta T_a.$$

e. Find $A$, $B$ and $B_v$ as functions of the system constants.
f. Solve the LQR Riccati equation which minimizes the index,

$$J = \int_0^\infty (\Delta T^2 + \rho\Delta q^2)d\tau,$$

and determine the corresponding feedback gain, $K$.

$\Delta T_a$ in the equation above (question d) is now to be considered as a white noise process with $E\{\Delta T_a\} = 0$ and $E\{\Delta T_a^2\} = V$. The temperature $\Delta T$ is to be measured in such a way that the system output equation is

$$\Delta y = \Delta T + w,$$

where w is white noise process with $E\{w\} = 0$ and $E\{w^2\} = W$. The state and process noise are independent and uncorrelated, i.e. $E\{\Delta T_a\ w\} = 0$.

g. Determine the Kalman gain $L$ for a Kalman filter which can estimate the state of the temperature subsystem and write down the equations for it.

### *Problem 7.9*

This exercise deals with a nonlinear optical fiber pulling process at NKT A/S (Nordic Cable and Wire Company, Inc.), Denmark. In an earlier project a linear control for this system was derived: Hendricks et al. (1985). It has become clear that because of newer products which require a variation of the fiber diameter and high accuracy that a digital control of the process would be advantageous.

In the figure below (Fig. 7.10) a schematic drawing of the pulling process is shown. Optical fibers are drawn from a quarts rod or preform. This is done by heating the preform to a high temperature (about 2000 C) and pulling a fiber from the semi-molten zone of the preform called the neck down region. The

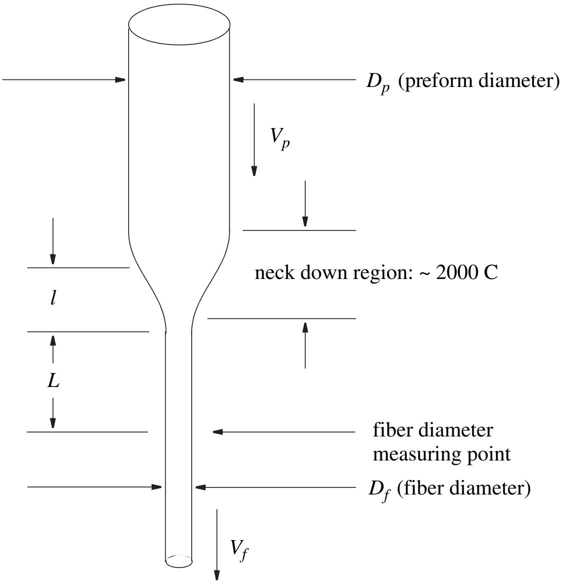


**Fig. 7.10** Schematic drawing of an optical fiber pulling process

control object is to control the diameter of the fiber, $D_\mathrm{f}$, (usually 0.125–1 mm) by changing the drawing speed, $V_\mathrm{f}$, (usually 0.2–1 m/s).

The fiber is formed over a distance which is $l$ but its diameter can be measured (optically) only after it has traveled a further distance $L$.

The process is such that it obeys the mass conservation law,

$$D_f^2 V_f = D_p^2 V_p,$$

where $D_p$ is the preform diameter (about 10–15 mm) and $V_p$ is the preform advance velocity. Both $D_p$ and $V_p$ are approximately constant.

The dynamic process model has been identified earlier and is in the form of a transfer function for the process.

$$\frac{d_f(s)}{v_f(s)} = \frac{-K_v e^{-s\tau_d}}{1 + \tau_v s},$$

where $s$ is the Laplace operator, $d_f$ is the incremental fiber diameter, $v_f$ is the incremental fiber pulling velocity, $\tau_d$ is the measurement time delay $(= L/V_f)$, $\tau_v$ is the process relaxation time $(= l/V_f)$ and $K_v$ is the process amplification. $\tau_d$, $\tau_v$ and $K_v$ all depend on the large signal pulling velocity, $V_f$. The process gain is give by the expression,

$$K_v = \frac{kD_f}{2V_f},$$

where $k$ is a constant.

a. Write down an differential equation which describes the dynamics of the system when it is assumed that $\tau_d$, $\tau_v$ and $K_v$ are constant.
b. What is the time constant and eigenfrequency of the system?
c. The tolerance for $D_f$ is ±1%. Show using the mass conservation law that $V_f$ is approximately constant in time.

Now the system is to be sampled in such a way that n samples are to be taken in the least of the characteristic times $\tau_d$ or $\tau_v$, i.e., $T =$ (least time constant)$/n$, where $T$ is the sampling period and $n$ is a whole number. As both $\tau_d$ and $\tau_v$ are scalars with respect to $1/V_f$, this leads to the state difference equation,

$$d(k+1) = a\,d(k) + (1-a)\,v(k-n),$$

where $a$ is a constant and $d(k) = d_f(k)$.

d. Show that the difference equation above is correct and give expressions for $a$ and $v$ in terms of the original system parameters and constants. Also select a reasonable sampling time (or $n$) for the system and explain this choice.

Assume now that $n = 1$. The additive state noise can be described as $v(k) = N(0, V)$ and that measurements of $d(k) = d_f(k)$ are made subject to measurement noise which is $w(k) = N(0, W)$.

e. Write down the state equations of this reduced system.
f. Design a Kalman filter for this system.
g. What are the advantages of using a non-constant sampling time in the system description? Name the technical disadvantages of such a choice.

### *Problem 7.10*

This problem concerns an idle speed control system for a gasoline engine. When an operator takes his foot off the accelerator pedal of a car, the engine speed generally falls to below 1200 rpm (revolutions per minute), and an idle speed control system takes over control of the engine speed. The idle speed control system keeps the engine speed constant at about 900 rpm in spite of other engine loads such as the generator, power steering pump, air conditioning compressor, etc. The throttle plate position is regulated in and around the nominal engine speed by a D.C. motor, usually via a gear reduction box. See Fig. 7.11.

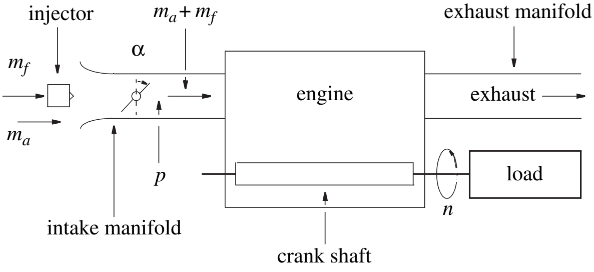


**Fig. 7.11** Schematic drawing of a single point injection, spark ignition engine. The arrow shows the positive direction for the throttle angle

The gear box drives the throttle plate directly. Usually a return spring is provided to return the throttle to its closed position (not shown) when there is no throttle input from the operator. The closed throttle angle is usually between 5 and 15 degrees.

The equation which describes the acceration of the inertia of the engine (plus its load) is

$$\frac{d}{dt}\left(\frac{1}{2}In^2\right) = In\dot{n} = -(P_l + P_b) + H_u\eta_i m_f,$$

where $n$ is the crank shaft speed, $I = 0.25(\mathrm{Kgm}^2) \cdot (2\pi/60)/1000$ is the engine inertia, $H_u = 43 \cdot 10^3\,\mathrm{Kg/kg}$ is the heating value of the fuel, $\eta_i \approx 0.33$ (at 900 rpm) is the engine thermal (indicated) efficiency, $m_f$(kg/s) is the fuel mass flow and $P_b$ is the load power. Here

$$P_l = n(k_{f0} + k_{f1}n) + k_{p0}n,$$

with $k_{f0} = 0.4500 \cdot 10^{-3}$ (kW/rpm), $k_{f1} = 0.2321 \cdot 10^{-6}$ (kW/rpm) and $k_{p0} = 0.9326 \cdot 10^{-3}$ (kW/rpm).

The air supply to the engine comes through the intake manifold which is described by the equation,

$$\dot{p} = -k_1\eta_{vol}p\,n + k_2 m_a(\alpha, p),$$

where $0 \le p \le 1$ is the normalized manifold pressure (i.e., $p = p_{man}/p_{am\hat{b}} =$ manifold pressure/ambient pressure). $k_1 = 15.118 \cdot 10^{-3}$ (1/rpm), $k_2 = 1.3822 \cdot 10^3$ (s/kg) and $m_a$ (kg/s) is the air mass flow to the engine. Given that $\alpha$ is the throttle angle (degrees), the air mass flow is given by

$$m_a(\alpha, p) = m_{a0} + m_{a1}\Phi(p)(1 - \cos\alpha),$$

and $m_{a0} = 1.0938 \cdot 10^{-3}$ (kg/s), $m_{a1} = 0.7854$ (kg/s) and $\Phi(p)$ is the flow coefficient for a compressible gas through a converging/diverging nozzle:

$$\Phi(p) = \begin{cases} \sqrt{\frac{2}{\kappa - 1}}\sqrt{p^{\frac{2}{\kappa}} - p^{\frac{\kappa + 1}{\kappa}}}, & \text{if} \quad p \ge \left(\frac{2}{\kappa + 1}\right)^{\frac{\kappa}{\kappa - 1}} \\ \sqrt{\left(\frac{2}{\kappa + 1}\right)^{\frac{\kappa + 1}{\kappa - 1}}}, & \text{if} \quad p < \left(\frac{2}{\kappa + 1}\right)^{\frac{\kappa}{\kappa - 1}} \end{cases}$$

and $\kappa = 1.4$ for air. If $p < (2/(\kappa + 1))^{\frac{\kappa}{\kappa - 1}}$ then the air flow is sonic and is limited by the shock wave around the throttle plate.

The volumetric efficiency is given by the equation,

$$\eta_{vol}(p) = \eta_0 + \eta_1 p,$$

with $\eta_0 = 0.5966$ and $\eta_1 = 0.2363$. Both of these constants are dimensionless.

Finally it is necessary that modern engines run with a control which gives a normalized stochiometric air/fuel ratio, $\lambda$, as a Three Way Catalyst (TWC) has to be used because of emission restrictions. The normalized air/fuel ratio is

$$\lambda = \frac{m_a}{m_f L_{th}},$$

where $L_{th} = 14.67$. This subsidiary control loop is not under consideration here but keeps $\lambda$ constant and equal to 1 over the entire operating range of the engine.

Before one can design a regulator for the idle speed, it is necessary to find the stationary operating point of the engine in idle. As an aid in determining this point, it is given that $n$ can be considered to be constant, that $n_0 = 900$ rpm and that the effective time constant for the manifold pressure is very small in relation to the effective crank shaft speed time constant.

a. Write down the gasoline engine's nonlinear state equations.
b. What is the fuel mass flow at the desired idle speed in the stationary state? What is the corresponding air mass flow? The engine is not loaded in idle.
c. Find an expression for the normalized manifold pressure p in terms of the known constants and variables at the desired operating point in the stationary state. Find its value in idle.
d. If one has calculated $m_a$ in idle it is possible to find the numerical value of $p$ at this point, $p_0$. What is it? What is the value of $\Phi(p_0)$? What is the throttle angle in idle?

As a gasoline engine and its control system are a higher order system, it is necessary to introduce some approximations in order to reduce the order of the system for hand calculations. The manifold pressure reaches its equilibrium rapidly so that one can assume that its dynamics can be neglected.

e. Write down the approximate nonlinear state equations for the engine control system alone when the dynamics of the intake manifold pressure are neglected.
f. Write down the corresponding linearized state equation for this system.

The throttle plate is driven by a gear box by the D.C. motor (see Fig. 2). The D.C. motors transfer function is

$$\omega(s) = \frac{K_{vm}}{1 + s\tau_m} V_a(s) - \frac{\tau_m / J}{1 + s\tau_m} M_{be}(s)$$

where $\omega$ is the axel's angular speed, $V_a$ is the rotor voltage, $K_{vm} = 20.76$ ((rad/s)/volt) and $\tau_m = 6.6323 \cdot 10^{-3}$ s. $M_{be}$ is the load torque (Nm) calculate for the axel of the D.C. motor. The gear ratio between the motor axel and the throttle plate axel is $R_a = 0.0019$. Because the D.C. motors time constant is very small compared to that of the gasoline engine, it is possible to neglect the dynamics of the D.C. motor drive system. Again, a reduction of the order of the system model is required.

g. Write down the resulting approximate state equations for the throttle control system which are compatible with the gasoline engine's state equation. These model is to include the torque load on the D.C. motor.

h. Design an optimal feedback matrix for the idle speed control system which is able to keep the crank shaft speed constant in idle when disturbances are ignored.

The stationary solution is desired so that the criteria which has to be minimized is

$$J = \lim_{t_1 \to \infty} \int_{t_0}^{t_1} [(n - n_0)^2 + \rho(\Delta V_a)^2] d\tau$$

where ρ is an arbitrary weighting factor. To simplify these calculations it is given that α and $n$ can be measured and that the system state equations have the general form

$$\begin{bmatrix} \Delta\dot{\alpha} \\ \Delta\dot{n} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \Delta\alpha \\ \Delta n \end{bmatrix} + \begin{bmatrix} b_1 \\ 0 \end{bmatrix} \Delta V_a.$$

The $\Delta'$s in front of the variables above indicates that they are incremental variables. This problem can be solved explicitly, without calculation by using the results of earlier examples.

i. How can the parameters of this example be arranged in such a way that this is possible? It is not required that the Riccati equation be actually solved here.

j. If part g. of this problem has been solved then it is possible to express $a_{21}$, $a_{22}$ and $b_1$ in terms of the gasoline engine and D.C. actuator system parameters and constants. Give expressions for these quantities.

In order to close the throttle plate when the accelerator pedal is released, the throttle plate is spring loaded (see the second figure in this problem). This results in a constant torque loading of the D.C. motor. Also the gasoline engine itself is loaded with a dynamo and perhaps and air conditioning system. It is also obviously necessary to take account of other accessories and the fact that the engine and its components will be worn and will age. Moreover the control system has to work on all vehicles in a particular production series in spite of production tolerances.

k. Under what conditions is it possible to obtain a low frequency compensation for loading of both the D.C. motor and gasoline engine? Are these conditions satisfied for the system which has been designed above?

l. Give a well supported solution of the loading problem if the engine is to run with approximately constant speed in idle. A complete solution is required with a control law and other detail taken into account. If a Riccati equation solution is required for this purpose, it is not to be solved here. Only a symbolic solution is desired.

### *Problem 7.11*

A test mass accelerometer is a measuring system for a vehicle (for example a ship, an aircraft of a car. In this problem a test mass accerometer is to be used to

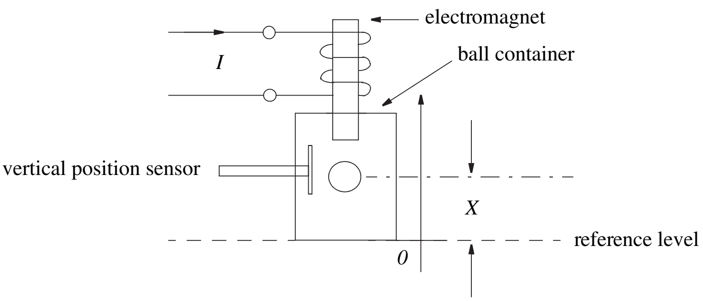


**Fig. 7.12** Drawing of a test mass accelerator

measure vertical accelerations. Figure 7.12 shows a simplified drawing of the accelerometer. The test mass is a steel ball with a mass $m = 10\,\text{g}$. The ball is held floating in a magnetic field by the electromagnet above it. The vertical position of the ball is measured by the position sensor on the left side to the ball container.

The force in Newtons which keeps the ball floating is given by the equation:

$$F(X, I) = 1.805 \times 10^3 X^2 + 14.44\,X\,I - 6.498\,X + 0.02888\,I^2 + 0.3740\,I - 0.1742,$$

where $X$ is the ball's position (in meters) in relation to the reference point (see the Fig. 7.13) and $I$ is the current (in Amperes). The function *F(X, I)* is pictured on the graph below. The ball is also influenced by air resistance. The force which air resistance exerts on the ball is

$$F(V) = -V(c_1 + c_2|V|)$$

where $c_1 = 1.55 \times 10^{-6}\,\text{N(m/s)}$, $c_2 = 2.2 \times 10^{-4}\,\text{N(m/s)}^2$ and the ball's velocity, $V$, is measured in m/s.

The accelerometer is under the influence of an external vertical acceleration, $A$.

a. Show that the nonlinear, large signal, movement of the ball can be described by the equation

$$\frac{d^2X}{dt^2} = \frac{1}{m}F(V) + \frac{1}{m}F(X, I) + A - g,$$

where $g = 9.81\,\text{m/s}^2$ is the acceleration of gravity.

b. The system in exercise a. above is to be linearized around a working point which is $X = 2.5\,\text{mm}$. What is the corresponding controlling current in the electromagnet at this operating point? At most two decimals accuracy is required. A graphic solution is possible.

c. Show that the system's linearized state equations can be written in the form

$$\dot{x} = a_{12}v,$$
$$\dot{v} = a_{21}x + a_{22}v + a + bi,$$

where $x$, $v$, $a$ and $i$ are the incremental state and input variables corresponding to $X$, $V$, $A$ and $I$, and $a_{12}$, $a_{21}$, $a_{22}$ and $b$ are constants. Show that $a_{12} = 1$, $a_{21} > 0$ and $a_{22} < 0$. Numerical results are not necessary at this stage but might be helpful later.

For a measuring system the accelerometer has unusual dynamics.

d. Calculate the transfer function of the accelerometer and show this. A numerical result is not necessary at this stage. Why is feedback necessary in order for the system to work as an accelerometer?

It is desired to design a full feedback system which is capable of minimizing the index

$$J = \lim_{t\to\infty} \int_0^t (x^2 + \rho i^2) d\tau$$

where ρ is a weighting factor.

e. Set up the Riccati equation which has to be solved in order to find the optimal feedback matrix. Solve the Riccati equation. Write down the optimal gain matrix in terms of the solution of the LQR Riccati equation.

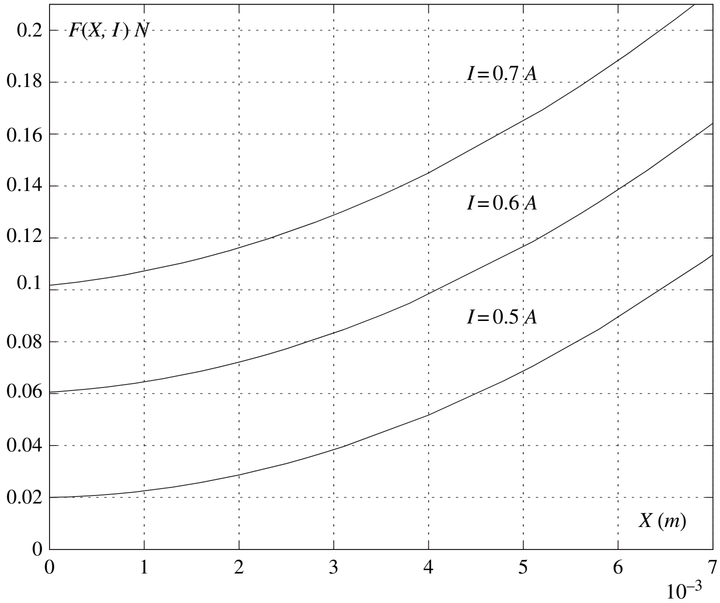


**Fig. 7.13** The function *F(X, I)* with the electromagnet current as a parameter

f. Draw a complete block diagram of the overall system.

g. If the feedback system in question c. is used to measure the unknown acceleration, $a$, is it possible to measure $a$ directly? What does one then measure and why? Is such a measurement accurate? The effects of noise are to be ignored in answering the question.

In order to find the unknown acceleration, $a$, more accurately it is helpful to estimate it with the use of a Kalman filter. That is to say that $a$ is to be assumed to be modelled as integrated white noise with an intensity $V_1$. The position measurement is noisy and the measurement model is thus

$$y = x + v_2$$

where $V_2$(t) is assumed to be white noise with an intensity $V_2$.

h. Write down the Kalman filter's state equations in such a way that one can estimate $a$ as a slowly varing constant in the stationary state.

i. Write down the Kalman filter Riccati equation which has to be solved in order to find the Kalman filter gain matrix. Write down the gain matrix in terms of the solution of the Riccati equation. Finally the equation is to be solved in order to find the explicit Kalman gain matrix.

# Appendix A
# Static Optimization

In optimization one is interested in finding the extreme values of a function of the one or more variables under certain constraints. In order to formulate the necessary equations the function to be optimized (maximized or minimized) has to be defined. Such a function is usually called an optimization or a performance index.

## A.1 Optimization Basics

Assume that $J$ is a real function of an $n$-dimensional vector $\mathbf{u} \in \Re^n$:

$$J = J(\mathbf{u}). \tag{A.1}$$

The value of $\mathbf{u}$ where the minimum of $J$ is obtained has to be found among the solutions of the equation:[†]

$$\frac{\partial J}{\partial \mathbf{u}} = \mathbf{0}. \tag{A.2}$$

The solutions of this equation give either a minimum of $J(\mathbf{u})$, a maximum of $J(\mathbf{u})$ or neither (a saddle point).

The method to determine the nature of the solution is to look at the second derivative of $J$, the Hessian matrix. A sufficient condition for a minimum is that the Hessian matrix is positive definite,

$$\frac{\partial^2 J}{\partial \mathbf{u} \partial \mathbf{u}} = \begin{bmatrix} \dfrac{\partial^2 J}{\partial u_1 \partial u_1} & \cdots & \dfrac{\partial^2 J}{\partial u_1 \partial u_n} \\ \vdots & & \vdots \\ \dfrac{\partial^2 J}{\partial u_n \partial u_n} & \cdots & \dfrac{\partial^2 J}{\partial u_n \partial u_n} \end{bmatrix} > 0.$$

[†] The gradient of a real valued function $f(\mathbf{x})$, $\mathbf{x} \in \Re^n$ is defined as a column vector $\nabla_{\mathbf{x}} f = \frac{\partial f}{\partial \mathbf{x}} = f_{\mathbf{x}} = \left[ \frac{\partial f}{\partial x_1} \frac{\partial f}{\partial x_2 .} \cdots \frac{\partial f}{\partial x_n} \right]^T$.

A maximum has a negative definite Hessian and if the Hessian is neither positive nor negative definite the solution to equation (A.2) is a saddle point.

***Example A.1*. Minimum of a Multivariable Function**

The real-valued function $J(\mathbf{u})$ is given by

$$J(\mathbf{u}) = \frac{1}{2}(x^2 + y^2 + x(y-1)). \tag{A.3}$$

The shape of the function is indicated in Fig. A.1. Here $\mathbf{u} = [x, y]^T \in \Re^2$. The gradient of the function $J$ is

$$\frac{\partial J}{\partial \mathbf{u}} = \begin{bmatrix} \dfrac{\partial J}{\partial x} \\ \dfrac{\partial J}{\partial y} \end{bmatrix} = \begin{bmatrix} x + \dfrac{1}{2}(y-1) \\ y + \dfrac{1}{2}x \end{bmatrix}. \tag{A.4}$$

This gradient is only zero in one point, which is

$$\mathbf{u} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \dfrac{2}{3} \\ -\dfrac{1}{3} \end{bmatrix}. \tag{A.5}$$

At this point the function has a global minimum, since the Hessian matrix is positive definite:

$$\frac{\partial^2 J}{\partial \mathbf{u} \partial \mathbf{u}} = \begin{bmatrix} \dfrac{\partial^2 J}{\partial x^2} & \dfrac{\partial^2 J}{\partial x \partial y} \\ \dfrac{\partial^2 J}{\partial x \partial y} & \dfrac{\partial^2 J}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 1 & \dfrac{1}{2} \\ \dfrac{1}{2} & 1 \end{bmatrix}. \tag{A.6}$$

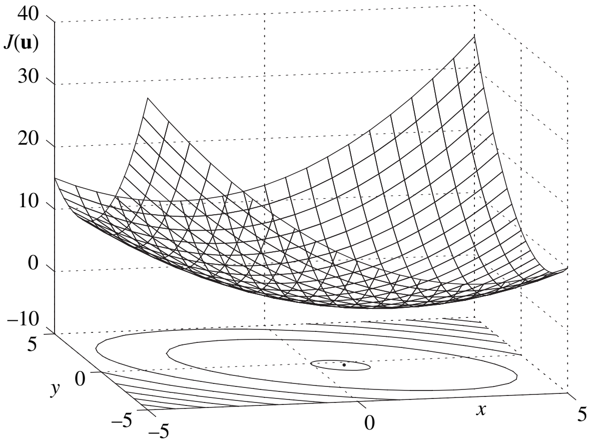


**Fig. A.1** The function $J(\mathbf{u})$ from Example A.1

To check for positive definiteness the following test is made:

$$\frac{\partial^2 J}{\partial x^2} = 1 > 0, det\left(\frac{\partial^2 J}{\partial \mathbf{u} \partial \mathbf{u}}\right) = \frac{3}{4} > 0. \tag{A.7}$$

❐

It is desired to generalize this method for more complicated situations. Therefore the simple static optimization problem is formulated in a way which lends itself to the more complicated problems.

Again let $J(\mathbf{u})$ be a real-valued function of the vector variable $\mathbf{u} \in \Re^n$. Let $\eta \in \Re^n$ be an arbitrary vector and $\varepsilon \in \Re$ a small real number. The function,

$$J(\mathbf{u} + \varepsilon\eta),$$

is now considered as a function of the real value $\varepsilon$. Then, according to basic calculus the minimum of $J$ has to be found among the solutions $\mathbf{u}$ to the equation

$$\left.\frac{d}{d\varepsilon} J(\mathbf{u} + \varepsilon\eta)\right|_{\varepsilon=0} = 0. \tag{A.8}$$

Application of the chain rule gives for this expression:

$$\left.\frac{d}{d\varepsilon} J(\mathbf{u} + \varepsilon\eta)\right|_{\varepsilon=0} = \left(\frac{\partial}{\partial \mathbf{u}} J(\mathbf{u})\right)^T \cdot \eta.$$

Since this condition has to hold for all values of $\eta \in \Re^n$ the result in equation (A.2) is obtained again. This method will be used in the later sections.

***Example A.2.* Minimization by Using Variational Techniques**

For the function in example A.1 the insertion of $\varepsilon$ gives for arbitrary $\eta = [\eta_x \eta_y]^T$,

$$\begin{aligned} J(\varepsilon) = J(\mathbf{u} + \varepsilon\eta) = \\ = \frac{1}{2}((x + \varepsilon\eta_x)^2 + (y + \varepsilon\eta_y)^2 + (x + \varepsilon\eta_x)(y + \varepsilon\eta_y - 1)). \end{aligned} \tag{A.9}$$

When differentiating this function with respect to $\varepsilon$ one finds

$$\lim_{\varepsilon \to 0} \frac{d}{d\varepsilon} J(\varepsilon) = \frac{1}{2}((2x + y - 1)\eta_x + (2y + x)\eta_y). \tag{A.10}$$

For arbitrary $\eta_x$ and $\eta_y$ this will only be zero if $x$ and $y$ are the same as in equation (A.5). ❐

### *A.1.1 Constrained Static Optimization*

The minimization above is valid if the values of $\mathbf{u}$ are unconstrained and unbounded. However in the case of a constraint on the values that the variable can attain this procedure must be modified. Assume that the task is to find the value of $\mathbf{u}$ that optimizes $J$ and at the same time obeys the constraint equation:

$$\mathbf{f}(\mathbf{u}) = \mathbf{0}. \tag{A.11}$$

This is known as an equality constraint. Here, $\mathbf{f} \in \Re^p$. Assume that $p < n$, which means that the constraint equation (A.11) is the equation of a $p$-dimensional surface in $\Re^n$.

In the optimum (where $J_\mathbf{u} = \mathbf{0}$) where the constraint is fulfilled the following differentials must be zero:

$$dJ = \left(\frac{\partial J}{\partial \mathbf{u}}\right)^T d\mathbf{u} = J_\mathbf{u}^T d\mathbf{u} = 0,$$

$$d\mathbf{f} = \frac{\partial \mathbf{f}}{\partial \mathbf{u}} d\mathbf{u} = \mathbf{f}_\mathbf{u} d\mathbf{u} = 0.$$

These are $p + 1$ linear equations in the $n$ variables $d\mathbf{u}$, which in matrix form are

$$\begin{bmatrix} dJ \\ d\mathbf{f} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial J}{\partial u_1} & \cdots & \dfrac{\partial J}{\partial u_n} \\ \dfrac{\partial f_1}{\partial u_1} & \cdots & \dfrac{\partial f_1}{\partial u_n} \\ \vdots & \cdots & \vdots \\ \dfrac{\partial f_p}{\partial u_1} & \cdots & \dfrac{\partial f_p}{\partial u_n} \end{bmatrix} \begin{bmatrix} du_1 \\ \vdots \\ du_n \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}.$$

If this set of linear equations is to have non-trivial solutions the matrix must have less than full rank or the rows must be linearly dependent, i.e., there must exist a $p + 1$ – dimensional vector of the form $[1, \lambda_1, \lambda_2, \ldots, \lambda_p]$ such that

$$[1, \lambda_1, \lambda_2, , \lambda_p] \begin{bmatrix} J_\mathbf{u}^T \\ \mathbf{f}_\mathbf{u} \end{bmatrix} = \mathbf{0}$$

The vector $\lambda = [\lambda_1, \lambda_2, , \lambda_p]^T$ is called a Lagrange multiplier and it is seen that by appropriate transposition the above matrix equation becomes:

$$J_\mathbf{u} + \mathbf{f}_\mathbf{u}^T \lambda = \mathbf{0}.$$

Inspecting this set of equations the Hamilton function is introduced,

$$H(\mathbf{u}, \lambda) = J(\mathbf{u}) + \lambda^T \mathbf{f}(\mathbf{u}). \tag{A.12}$$

The differential of $H$ is given by

$$\begin{aligned} dH &= H_{\mathbf{u}}^T d\mathbf{u} + H_{\lambda}^T d\lambda \\ &= (J_{\mathbf{u}} + \mathbf{f}_{\mathbf{u}}^T \lambda)^T d\mathbf{u} + \mathbf{f}^T d\lambda \\ &= 0. \end{aligned}$$

This is true, because $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ and because $J_{\mathbf{u}} + \mathbf{f}_{\mathbf{u}}\lambda = \mathbf{0}$. It is now possible to draw the following conclusion:

A necessary condition for the optimal solution of equation (A.1) under the constraint in (A.11) is that $\mathbf{u}$ and $\lambda$ obey the two equations:

$$\begin{aligned} \frac{\partial H}{\partial \mathbf{u}} &= \mathbf{0}, \\ \frac{\partial H}{\partial \lambda} &= \mathbf{0}. \end{aligned} \tag{A.13}$$

Here $H(\mathbf{u}, \lambda)$ is given in Eq. (A.12). It is seen that the problem has been reduced to solving a set of unconstrained equations at the price of introducing $p$ new variables, the Lagrange multipliers. This is the way in which equality constraints in the optimal control problem are to be treated.

### *Example A.3*. Minimization with Constraints

The function $J(\mathbf{u})$ from Example A.1 is now to be optimized under the equality constraint:

$$f(\mathbf{u}) = x - 2y = 0. \tag{A.14}$$

Introducing a scalar Lagrange multiplier, $\lambda$, gives the un-constrained optimization problem,

$$\begin{aligned} H(\mathbf{u}, \lambda) &= (J(\mathbf{u}) + \lambda \cdot f(\mathbf{u})) \\ &= \frac{1}{2}(x^2 + y^2 + x(y-1)) + \lambda(x - 2y). \end{aligned} \tag{A.15}$$

The un-constrained optimization gives the three equations,

$$\begin{bmatrix} \dfrac{\partial H}{\partial \mathbf{u}} \\ \dfrac{\partial H}{\partial \lambda} \end{bmatrix} = \begin{bmatrix} x + \frac{1}{2}(y-1) + \lambda \\ y + \frac{1}{2}y - 2\lambda \\ x - 2y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{A.16}$$

There is one solution to this set of coupled equations,

$$[x\,y\,\lambda] = \begin{bmatrix} \frac{2}{7} & \frac{1}{7} & \frac{1}{7} \end{bmatrix}.$$

This simple example could be solved by inserting the constraint $x - 2y = 0$ into the function $J$, i.e., if $x = 2y$ the function becomes

$$J(x,y)|_{x=2y} = g(y) = \frac{1}{2}(4y^2 + y^2 + 2y(y-1)) = \frac{1}{2}(7y^2 - 2y).$$

The minimum of the function $g(y)$ is the solution of

$$\frac{dg}{dy} = 0 \Leftrightarrow y = \frac{1}{7}.$$

This is of course the same as above, but the above method can also be used in more complex optimization problems. ❒

## A.2 Problems

***Problem A.1***

Find the extrema of the function:

$$J(x,y) = e^{-r} \cdot \cos(r).$$

Here $(x,y) = (r\cos\theta, r\sin\theta)$. Show that origin is a local maximum.

***Problem A.2***

Find the extrema of the function,

$$J(x,y) = x^2 - xy + y^2 + 3x,$$

under the constraint,

$$f(x,y) = 0$$

Here $f(x,y) = x - 4y$. Show that the extremum found is a minimum.

### *Problem A.3*

A rectangle has side lengths $x$ and $y$ respectively.

a. Find the rectangle with the largest area given that its perimeter is $p$.

b. Find the rectangle with smallest perimeter given that its area is $a^2$.

### *Problem A.4*

Minimize the cost function,

$$L(\mathbf{x},\mathbf{u}) = \frac{1}{2}\mathbf{x}^T\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}\mathbf{x} + \frac{1}{2}\mathbf{u}^T\begin{bmatrix}2 & 1\\ 1 & 1\end{bmatrix}\mathbf{u},$$

subject to the constraint,

$$\mathbf{x} = \begin{bmatrix}1\\ 3\end{bmatrix} + \begin{bmatrix}2 & 2\\ 1 & 0\end{bmatrix}\mathbf{u}.$$

At the minimum find the values of $\mathbf{x}$, $\mathbf{u}$, $\lambda$ and $L$.

# Appendix B
# Linear Algebra

This appendix gives a short introduction to the basic matrix properties necessary for this book. The reader is referred to the literature for a more in depth treatment of linear algebra, e.g. Golub and Van Loan, 1993.

## B.1 Matrix Basics

A $n \times m$ matrix is a collection of real or complex numbers, $a_{ij}$ organized in a rectangular array with $n$ rows and $m$ columns:

$$\mathbf{A} = \{a_{ij}\} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix}. \tag{B.1}$$

Addition of two matrices is accomplished by adding their elements. Multiplication of two matrices, e.g. $\mathbf{A} = \mathbf{B} \cdot \mathbf{C}$ is done by combining the rows of $\mathbf{C}$ with the columns of $\mathbf{B}$ according to the formula:

$$a_{ij} = \sum_{k=1}^{p} b_{ik} \cdot c_{kj}, \mathrm{i} = 1, \dots, n, \mathrm{j} = 1, \dots, m. \tag{B.2}$$

For this to be possible $\mathbf{B}$ has to have the same number of columns as $\mathbf{C}$ has rows, i.e., if the size of $\mathbf{B}$ is $\dim[\mathbf{B}] = n \times p$ and the size of $\mathbf{C}$ is $\dim[\mathbf{C}] = p \times m$ the product has the size $\dim[\mathbf{A}] = n \times m$. Generally this multiplication does not obey the commutative law, i.e.,

$$\mathbf{AB} \neq \mathbf{BA}. \tag{B.3}$$

For a matrix $\mathbf{A} = (a_{ij})$ a number of related manipulations are defined. The transposed matrix is obtained by interchanging columns and rows,

$$\left(\mathbf{A}^T\right)_{ij} = a_{ji}. \tag{B.4}$$

A symmetric matrix, $\mathbf{A}_S$, is equal to its transpose, i.e.,

$$\mathbf{A}_S{}^T = \mathbf{A}_S, \tag{B.5}$$

and likewise an anti-symmetric matrix, $\mathbf{A}_A$ changes sign when transposed,

$$\mathbf{A}_A^T = -\mathbf{A}_A. \tag{B.6}$$

The trace of a quadratic matrix (a matrix with equal number of rows and columns) is the sum of the diagonal elements, i.e.,

$$tr(A) = \sum_{i=1}^{n} a_{ii}. \tag{B.7}$$

For two quadratic matrices one has: $tr(\mathbf{AB}) = tr(\mathbf{BA})$.

The identity matrix is

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}. \tag{B.8}$$

The inverse matrix $\mathbf{A}^{-1}$ of a quadratic matrix $\mathbf{A}$ obeys the equation,

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \tag{B.9}$$

A quadratic matrix can only be inverted if it is non-singular. To determine if a matrix is non-singular the determinant of a matrix is introduced which is defined as

$$\det\mathbf{A} = \sum_{j=1}^{n} a_{ij}\Upsilon_{ij}. \tag{B.10}$$

On the right hand side of equation (B.10) the index can be any number in the interval $[1, n]$. $\gamma_{ij}$ is the $(i,j)$ the *cofactor* of $\mathbf{A}$ defined by:

$$\gamma_{ij} = (-1)^{i+j}\det\tilde{\mathbf{A}}_{ij}. \tag{B.11}$$

The matrix $\tilde{\mathbf{A}}_{ij}$ has the same elements as $\mathbf{A}$ except that the $i$'th row and the $j$'th column have been removed so that $\tilde{\mathbf{A}}_{ij}$ has the dimension $(n-1)\times(n-1)$. It is seen that the determinant is defined and can be calculated recursively.

The *complementary matrix* of **A** is defined as

$$\mathbf{C_A} = \{\gamma_{ij}\}. \tag{B.12}$$

The transpose of $\mathbf{C_A}$ is called the *adjoint matrix* of **A**:

$$\mathrm{adj}(\mathbf{A}) = \mathbf{C}_\mathbf{A}^T \tag{B.13}$$

and this allows a closed formula for the inverse of a non-singular matrix to be given:

$$\mathbf{A}^{-1} = \frac{\mathrm{adj}(\mathbf{A})}{\det\mathbf{A}}. \tag{B.14}$$

This formula is most convenient for calculating the inverse symbolically. For numerical computation better and more efficient methods can be found in the literature.

The rank of a matrix is defined as the number of linearly independent columns or rows, i.e., if **A** is a $n \times m$ – dimensional matrix, then $r$ is the rank if any $r+1$ columns or rows of **A** are linearly dependent but any $r$ columns or rows are linearly independent. This means that if $[\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r]$ are any $r$ columns or rows then the following implication is true:

$$\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \ldots + \alpha_r\mathbf{a}_r = \mathbf{0} \Rightarrow \alpha_i = 0, \quad \forall i. \tag{B.15}$$

If the rank is $r$ then all $(r+1) \times (r+1)$ – dimensional complementary matrices will be singular whereas at least one $r \times r$ – dimensional one will not be.

A diagonal matrix is a quadratic matrix, where all off-diagonal elements are zero,

$$\mathrm{diag}(\lambda_1, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}. \tag{B.16}$$

The determinant of the diagonal matrix is the product of the diagonal elements.

## B.2 Eigenvalues and Eigenvectors

The characteristic polynomial for a quadratic $n \times m$ matrix **A** is defined as:

$$a(s) = \det(s\mathbf{I} - \mathbf{A}) = s^n + a_{n-1}s^{n-1} + \ldots + a_1 s + a_0. \tag{B.17}$$

The $n$ roots of the characteristic polynomial are called the eigenvalues of $\mathbf{A}$. For each of the eigenvalues $\lambda_i, i = 1, \ldots, n$ there exists a non-zero vector $\mathbf{v}_i$, the eigenvector, such that the eigenvector equation is satisfied:

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i. \tag{B.18}$$

If $\mathbf{v}_i$ is an eigenvector then any multiplication of $\mathbf{v}_i$ with a non-zero scalar is also an eigenvector. If the characteristic equation has no two eigenvalues equal, then all the corresponding eigenvectors are linearly independent and will therefore span all of $\Re^n$. If all the eigenvectors are linearly independent the matrix is said to be simple. The matrix consisting of the eigenvectors as columns is called the modal matrix:

$$\mathbf{M} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]. \tag{B.19}$$

For simple matrices the modal matrix will be a non-singular similarity transformation and from (B.18) one has

$$\mathbf{A}\mathbf{M} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)\mathbf{M} \tag{B.20}$$

or

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n). \tag{B.21}$$

In the space spanned by the eigenvectors the matrix becomes a diagonal matrix with the eigenvalues in the diagonal. This is very convenient but unfortunately this only holds for simple matrices. If the characteristic polynomial has multiple roots (which is not unusual) the eigenvectors can not be guaranteed to be linearly independent. Such a matrix is known as being defective. In this case another set of $n$ linearly independent vectors has to be chosen. Assume that a given eigenvalue $\lambda_i$ has a multiplicity of $r_i$, then one can define $r_i$ vectors $\mathbf{v}_1^i, \mathbf{v}_2^i, \ldots, \mathbf{v}_{r_i}^i$ recursively with the following formula:

$$\mathbf{A}\mathbf{v}_{k+1}^i = \lambda_i \mathbf{v}_{k+1}^i + \mathbf{v}_k^i, \ k = 0, \ldots, r_i. \tag{B.22}$$

Define $\mathbf{v}_0^i = \mathbf{0}$, then the formula is valid for $k = 0$ and the first vector $\mathbf{v}_1^i$ is the eigenvector corresponding to $\lambda_i$. These vectors are called generalized eigenvectors and are linearly independent. All these generalized eigenvectors form a set of $n$ linearly independent basis vectors and may be collected in the matrix,

$$\mathbf{M} = [\mathbf{v}_{r_1}^1, \ldots, \mathbf{v}_1^1, \mathbf{v}_{r_2}^2, \ldots, \mathbf{v}_1^2, \ldots, \mathbf{v}_{r_l}^l, \ldots, \mathbf{v}_1^l]. \tag{B.23}$$

Here $l$ is the number of different eigenvectors, such that $r_1 + r_2 + \ldots + r_1 = n$. Multiplying with $\mathbf{A}$ gives

$$\mathbf{AM} = \mathbf{JM}. \tag{B.24}$$

The so-called *Jordan normal form* is obtained for the matrix in the basis spanned by the column vectors in **M**. The Jordan normal form is a block diagonal matrix in the form:

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & 0 & \dots & 0 \\ 0 & \mathbf{J}_2 & \dots & 0 \\ 0 & & \dots & 0 \\ 0 & \dots & 0 & \mathbf{J}_l \end{bmatrix}. \tag{B.25}$$

Here, the individual blocks are:

$$\mathbf{J}_i = \begin{bmatrix} \lambda_i & 1 & \dots & 0 \\ 0 & \lambda_i & 1 & \vdots \\ \vdots & \vdots & \vdots & 0 \\ 0 & \dots & 0 & \lambda_i \end{bmatrix}. \tag{B.26}$$

The blocks in the Jordan normal form have the eigenvalues in the diagonal and either 1's or 0's over the diagonal.

Apart from the eigenvectors which are also sometimes called the right eigenvectors the left eigenvectors may be defined as the solution to the linear equation:

$$\mathbf{wA} = \lambda\mathbf{w}. \tag{B.27}$$

Forming the modal matrix from the left eigenvectors one has in a way similar to Eq. (B.21):

$$\begin{aligned} \mathbf{N} &= [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n], \\ \mathbf{NA} &= \text{diag}(\lambda_1, \dots, \lambda_n)\mathbf{N}. \end{aligned} \tag{B.28}$$

If **A** is simple the left modal matrix is non-singular and therefore on has the relation:

$$\mathbf{NAN}^{-1} = (\lambda_1, \dots, \lambda_n) = \mathbf{M}^{-1}\mathbf{AM}. \tag{B.29}$$

This means that $\mathbf{N} = \mathbf{M}^{-1}$ and therefore the left and right eigenvectors are orthogonal:

$$\mathbf{w}_i^{-1}\mathbf{v}_j = \mathbf{w}_i\mathbf{v}_j^{-1} = \begin{cases} 1 & \text{for} \quad i = j \\ 0 & \text{for} \quad i \neq j \end{cases}. \tag{B.30}$$

For a quadratic matrix $\mathbf{A}$ with the characteristic polynomial, $a(s)$, given by Eq. (B.17) the so-called Cayley-Hamilton theorem states that

$$a(\mathbf{A}) = \mathbf{A}^n + a_{n-1}\mathbf{A}^{n-1} + \ldots + a_1\mathbf{A} + a_0\mathbf{I} = \mathbf{0}. \tag{B.31}$$

That $\mathbf{A}$ is a solution of in its own characteristic equation.

## B.3 Partitioned Matrices

A block diagonal matrix is defined in the form,

$$\mathbf{D} = \begin{bmatrix} \mathbf{A}_{11} & 0 & \ldots & 0 \\ 0 & \mathbf{A}_{22} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & \mathbf{A}_{nn} \end{bmatrix} \tag{B.32}$$

and it can also be written in the form:

$$\mathbf{D} = \mathrm{diag}(\mathbf{A}_{11}, \mathbf{A}_{22}, \ldots, \mathbf{A}_{nn}). \tag{B.33}$$

A (lower) lower triangularupper triangular block matrix has all the elements (above) below the diagonal equal to zero:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \ldots & \mathbf{A}_{1n} \\ 0 & \mathbf{A}_{22} & \ldots & \mathbf{A}_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathbf{A}_{nn} \end{bmatrix}. \tag{B.34}$$

Its determinant is the product of the determinants of the diagonal matrices,

$$\det(\mathbf{A}) = \det(\mathbf{A}_{11})\det(\mathbf{A}_{22})\ldots\det(\mathbf{A}_{nn}). \tag{B.35}$$

For the block partitioned matrix,

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \tag{B.36}$$

the Schur complement of $\mathbf{A}_{11}$ is defined to be:

$$\mathbf{D}_{11} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}, \tag{B.37}$$

and the Schur complement of $\mathbf{A}_{22}$,

$$\mathbf{D}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}. \tag{B.38}$$

Then depending on whether $\det(A_{11}) \neq 0$ or $\det(A_{22}) \neq 0$ or both, it can be shown that the inverse of the $\mathbf{A}$ is,

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{D}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{D}_{22}^{-1} \\ -\mathbf{D}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{D}_{22}^{-1} \end{bmatrix} \text{ for } \det(A_{11}) \neq 0,$$

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{D}_{11}^{-1} & -\mathbf{D}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{D}_{11}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{D}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix} \text{ for } \det(A_{22}) \neq 0, \tag{B.39}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{D}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{D}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{D}_{11}^{-1} & \mathbf{D}_{22}^{-1} \end{bmatrix} \text{ for } \det(A_{11}) \neq 0 \text{ and } \det(A_{11}) \neq 0.$$

This is most easily verified by doing the multiplications, $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. From these forms the well-known matrix inversion lemma can be derived:

$$(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}. \tag{B.40}$$

## B.4 Quadratic Forms

For any square matrix, $\mathbf{Q}$ the quadratic form can be constructed:

$$\mathbf{x}^T\mathbf{Q}\mathbf{x}. \tag{B.41}$$

This gives a mapping from $\Re^n$ to $\Re$. All matrices may be divided into a symmetric, $\mathbf{Q}_S$, and an anti-symmetric part, $\mathbf{Q}_A$, with the following trick:

$$\mathbf{Q} = \frac{1}{2}(\mathbf{Q} + \mathbf{Q}^T) + \frac{1}{2}(\mathbf{Q} - \mathbf{Q}^T) = \mathbf{Q}_S + \mathbf{Q}_A. \tag{B.42}$$

In the quadratic form however the anti-symmetric component gives no contribution, since

$$\mathbf{x}^T\mathbf{Q}\mathbf{x} = (\mathbf{x}^T\mathbf{Q}\mathbf{x})^T = \mathbf{x}^T\mathbf{Q}^T\mathbf{x} \Rightarrow \mathbf{x}^T\mathbf{Q}_A\mathbf{x} = 0. \tag{B.43}$$

Therefore one can always assume that the matrices in quadratic forms are symmetric without loss of generality.

For a given matrix, **Q**, the following may be defined:

- **Q** is positive definite, ($\mathbf{Q} > 0$), if $\mathbf{x}^T\mathbf{Q}\mathbf{x} > 0$ for all non-zero **x**,
- **Q** is negative definite, ($\mathbf{Q} < 0$) if $\mathbf{x}^T\mathbf{Q}\mathbf{x} < 0$ for all non-zero **x**,
- **Q** is positive semi-definite, ($\mathbf{Q} \geq 0$) if $\mathbf{x}^T\mathbf{Q}\mathbf{x} \geq 0$ for all **x**,
- **Q** is negative semi-definite, ($\mathbf{Q} \leq 0$) if $\mathbf{x}^T\mathbf{Q}\mathbf{x} \leq 0$ for all **x**.

For a positive definite $n \times n$ matrix, **Q**, the corresponding **Q**-norm can be defined from the expression,

$$\|\mathbf{x}\|_{\mathbf{Q}} = \sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}}. \tag{B.44}$$

This has all the qualities of a norm on the vector space $\Re^n$. It is a positive number except when **x** is zero. If **Q** is the identity matrix one obtains the normal Euclidean norm.

For a positive semi-definite matrix the square root matrix, **S**, is defined such that

$$\mathbf{Q} = \mathbf{S}^T\mathbf{S} \text{ or } \mathbf{Q} = \mathbf{S}\mathbf{S}^T, \tag{B.45}$$

and one writes $\mathbf{S} = \sqrt{\mathbf{Q}}$. In general there will be many different square root matrices for a given positive definite matrix and they are all non-singular.

To check for definiteness, the easiest way is to calculate all the eigenvalues, $\lambda_i$, of **Q**. The rule is:

$$\begin{aligned} \mathbf{Q} &> 0 \text{ if all } \lambda_i > 0, \\ \mathbf{Q} &\geq 0 \text{ if all } \lambda_i \geq 0, \\ \mathbf{Q} &\leq 0 \text{ if all } \lambda_i \leq 0, \\ \mathbf{Q} &< 0 \text{ if all } \lambda_i < 0, \end{aligned} \tag{B.46}$$

An alternative test can be made by calculating the leading minors of $\mathbf{Q} = [q_{ij}]$:

$$\begin{aligned} m_1 &= q_{11}, \\ m_2 &= \det\begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}, \\ m_2 &= \det\begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}, \\ &\vdots \\ m_n &= \det\mathbf{Q}. \end{aligned} \tag{B.47}$$

Then if $\mathbf{Q} > 0$, is all the leading minors are positive, if $\mathbf{Q} < 0$ then the minors alternate between positive and negative values. As far as semi-definiteness is concerned all minors (not just the leading ones) have to be non-negative and non-positive respectively.

## B.5 Matrix Calculus

For a quadratic matrix $\mathbf{A}$ the exponential is defined by the infinite sum,

$$e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + \frac{1}{3!}\mathbf{A}^3 + \ldots \tag{B.48}$$

It can be shown that this sum always converges for any $\mathbf{A}$. The exponential has the following properties:

$$\begin{aligned} \frac{d}{dt}e^{\mathbf{A}t} &= \mathbf{A}e^{\mathbf{A}t}, \\ e^{\mathbf{A}t_1}e^{\mathbf{A}t_2} &= e^{\mathbf{A}(t_1+t_2)}. \end{aligned} \tag{B.49}$$

and for a non-singular matrix $\mathbf{A}$ one has further that

$$\left(e^{\mathbf{A}}\right)^{-1} = e^{-\mathbf{A}}. \tag{B.50}$$

On the other hand the exponentials do not generally commute, i.e.,

$$e^{\mathbf{A}}e^{\mathbf{B}} \neq e^{\mathbf{A}+\mathbf{B}} \neq e^{\mathbf{B}}e^{\mathbf{A}}, \tag{B.51}$$

unless $\mathbf{A}$ and $\mathbf{B}$ commute, i.e., $\mathbf{AB} = \mathbf{BA}$.

For a real-valued function $L(\mathbf{x}) \in \Re$, where $\mathbf{x} \in \Re^n$ the gradient is defined as the column vector:

$$\frac{\partial L}{\partial \mathbf{x}} = L_{\mathbf{x}} = \begin{bmatrix} \frac{\partial L}{\partial x_1} \\ \frac{\partial L}{\partial x_2} \\ \vdots \\ \frac{\partial L}{\partial x_n} \end{bmatrix}. \tag{B.52}$$

The total differential of $L$ is then:

$$dL = \left(\frac{\partial L}{\partial \mathbf{x}}\right)^T d\mathbf{x} = \sum_{k=1}^{n} \frac{\partial L}{\partial x_i} dx_i. \tag{B.53}$$

If $L$ were a function of two variables, $\mathbf{x}$ and $\mathbf{y}$, the total differential is:

$$dL = \left(\frac{\partial L}{\partial \mathbf{x}}\right)^T d\mathbf{x} + \left(\frac{\partial L}{\partial \mathbf{y}}\right)^T d\mathbf{y}. \tag{B.54}$$

The Hessian for the function $L(\mathbf{x})$ is defined as the $n \times n$-matrix:

$$L_{\mathbf{xx}} = \frac{\partial^2 L}{\partial \mathbf{x}^2} = \left(\frac{\partial^2 L}{\partial x_i \partial x_j}\right) \tag{B.55}$$

For a vector valued function $f(\mathbf{x}) \in \Re^p$ the Jacobian (the Jacoby matrix) is defined as the $p \times n$ – dimensional matrix:

$$f_{\mathbf{x}} = \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_p}{\partial x_1} & \frac{\partial f_p}{\partial x_2} & \cdots & \frac{\partial f_p}{\partial x_n} \end{bmatrix}. \tag{B.56}$$

For $\mathbf{A}$ and $\mathbf{y}$ of appropriate dimensions the following rules hold true:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}}(\mathbf{y}^T\mathbf{x}) &= \frac{\partial}{\partial \mathbf{x}}(\mathbf{y}\mathbf{x}^T) = \mathbf{y}, \\ \frac{\partial}{\partial \mathbf{x}}(\mathbf{y}^T\mathbf{A}\mathbf{x}) &= \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}^T\mathbf{y}) = \mathbf{A}^T\mathbf{y}, \\ \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) &= \mathbf{A}^T\mathbf{x} + \mathbf{A}\mathbf{x}, \\ \frac{\partial}{\partial \mathbf{x}}(\mathbf{y}^T f(\mathbf{x})) &= \frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x})^T\mathbf{y}) = f_{\mathbf{x}}^T\mathbf{y}. \end{aligned} \tag{B.57}$$

If $f$ and $g$ are vector functions the product rule gives the following Jacobian:

$$\frac{\partial}{\partial \mathbf{x}}(f^T g) = \frac{\partial}{\partial \mathbf{x}}(g^T f) = g_{\mathbf{x}}^T f + g^T f_{\mathbf{x}}. \tag{B.58}$$

# Appendix C
# Continuous Riccati Equation

This appendix deals with the details of deriving the optimal solution to the estimator problem used in Chap. 7. These results have been judged to be too complex to be inserted into the text of the chapter at the proper point. For further details see Kwakernaak and Sivan (1972) and Friedland (1987).

## C.1 Estimator Riccati Equation

In order to use the LQR control solution to derive the optimal estimator solution, it is necessary to reverse the time axis in this solution.

### *C.1.1 Time Axis Reversal*

It is easy to show (using a simple change of variables) that the two differential equations,

$$-\frac{dp(t)}{dt} = f[p(t), t' - t], t \leq t_1, \tag{C.1}$$

with an initial condition $p(t_1) = p_1$ and

$$\frac{dq(t)}{dt} = f[q(t), t], t \geq t_0, \tag{C.2}$$

with an initial condition $q(t_0) = q_0$ and where

$$t' = t_0 + t_1, t_0 < t_1, \tag{C.3}$$

have the solutions which have the follow relationships:

$$\begin{aligned} p(t) &= q(t' - t), t \leq t_1, \\ q(t) &= p(t' - t), t \geq t_0. \end{aligned} \tag{C.4}$$

These relationships will be used in what follows.

### C.1.2 Using the LQR Solution

In Chap. 7 the following equation was derived

$$\begin{aligned}\dot{\mathbf{Q}}'(t) &= [\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]\mathbf{Q}'(t) + \mathbf{Q}'(t)[\mathbf{A}(t) - \mathbf{L}(t)\mathbf{C}(t)]^T \\ &\quad + \mathbf{B}_v(t)\mathbf{V}_1(t)\mathbf{B}_v^T(t) + \mathbf{L}(t)\mathbf{V}_2(t)\mathbf{L}^T(t).\end{aligned} \tag{C.5}$$

The initial condition for integration of this equation is $\mathbf{Q}'(t_0) = \mathbf{Q}_0'$.

What is required now is an optimal value of $\mathbf{L}(t)$ which will allow the estimation error covariance matrix in equation (C.5) to be minimized. To do this the time axis in equation (C.5) has to be reversed in order to be able to use the solution of the optimal regulator problem (Sect. 5.3, Eq. (5.37)). This is effectively only a change of variables as shown above.

Letting $t' = t_1 + t_0$, $\mathbf{Q}'(t) = \mathbf{P}'(t' - t)$ and the regulator terminal condition be $\mathbf{P}'(t_1) = \mathbf{Q}_0'$, the time reversal of the LQR Riccati equation (C.5) leads to

$$\begin{aligned}-\dot{\mathbf{P}}'(t) &= [\mathbf{A}^T(t' - 1) - \mathbf{C}^T(t' - t)\mathbf{L}^T(t' - 1)]^T\mathbf{P}'(t) \\ &\quad + \mathbf{P}'(t)[\mathbf{A}^T(t' - t) - \mathbf{C}^T(t' - t)\mathbf{L}^T(t' - t)] \\ &\quad + \mathbf{B}_v(t' - t)\mathbf{V}_1(t' - t)\mathbf{B}_v^T(t' - t) + \mathbf{L}(t' - t)\mathbf{V}_2(t)\mathbf{L}^T(t' - t),\end{aligned} \tag{C.6}$$

where $t \leq t_1$, $t' = t_0 + t_1$ and here $\mathbf{V}_1(t) = \mathbf{R}_1(t)$ and $\mathbf{V}_2(t) = \mathbf{R}_2(t)$ in equation (5.36).

In this form, Eq. (C.6) can be identified with the solution to the optimal regulator problem. This is given by the Riccati equation,

$$\begin{aligned}-\dot{\mathbf{P}}(t) &= [\mathbf{A}(t) - \mathbf{B}(t)\mathbf{K}(t)]^T\mathbf{P}(t) + \mathbf{P}(t)[\mathbf{A}(t) - \mathbf{B}(t)\mathbf{K}(t)] \\ &\quad + \mathbf{R}_1(t) - \mathbf{K}(t)\mathbf{R}_2(t)\mathbf{K}^T(t),\end{aligned} \tag{C.7}$$

with the terminal condition $\mathbf{P}(t_1) = \mathbf{P}_1,\ t_0 \leq t \leq t_1$. The optimal regulator feedback gain for equation (C.7) has be shown to be given by

$$\mathbf{K}(t) = \mathbf{R}_2^{-1}(t)\mathbf{B}^T(t)\mathbf{P}(t). \tag{C.8}$$

From the previous section it is clear that

$$\mathbf{Q}'(t) = \mathbf{P}'(t' - t), \tag{C.9}$$

where the terminal condition is now $\mathbf{P}'(t_1) = \mathbf{Q}_0$.

Using the results from the LQR control solution, it is clear that the optimal solution to equation (C.6) is

$$\mathbf{L}(t' - \tau) = \mathbf{V}_2^{-1}(t' - \tau)\mathbf{C}(t' - \tau)\mathbf{P}(\tau), \tag{C.10}$$

in the interval $t \leq \tau \leq t_1$. With this substitution for $\mathbf{L}(t)$, Eq. (C.6) becomes

$$\begin{aligned}-\dot{\mathbf{P}}(t) = {} & \mathbf{P}(t)\mathbf{A}^T(t'-t) + \mathbf{A}(t'-t)\mathbf{P}(t) \\ & + \mathbf{B}_v(t'-t)\mathbf{V}_1(t'-t)\mathbf{B}_v^T(t'-t) \\ & - \mathbf{P}(t)\mathbf{C}^T(t'-t)\mathbf{V}_2^{-1}(t'-t)\mathbf{C}(t'-t)\mathbf{P}(t),\end{aligned} \tag{C.11}$$

for $t \leq t_1$, with the terminal condition $\mathbf{P}(t_1) = \mathbf{Q}_0$ . $\mathbf{P}(t)$ has been minimized in the sense that $\mathbf{P}(t) \leq \mathbf{P}'(t)$, $t \leq t_1$.

Reversing time back again in equation (C.11), one obtains the desired equation which is

$$\begin{aligned}\dot{\mathbf{Q}}(t) = {} & \mathbf{A}(t)\mathbf{Q}(t) + \mathbf{Q}(t)\mathbf{A}^T(t) \\ & + \mathbf{B}_v(t)\mathbf{V}_1(t)\mathbf{B}_v^T(t) - \mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t)\mathbf{C}(t)\mathbf{Q}(t),\end{aligned} \tag{C.12}$$

where it is clear that $\mathbf{Q}'(t) \geq \mathbf{Q}(t)$, $t \geq t_0$. The optimal stochastic observer or Kalman gain is then

$$\mathbf{L}(t) = \mathbf{Q}(t)\mathbf{C}^T(t)\mathbf{V}_2^{-1}(t). \tag{C.13}$$

The initial condition for this equation is $\mathbf{Q}(t_0) = \mathbf{Q}_0$. Because $\mathbf{L}(t)$ is time dependent, the filter is in fact optimal in time at every instant.

# Appendix D
# Discrete Time SISO Systems

This appendix gives a short overview of the classic analysis methods for treating discrete time systems and systems with continuous as well as discrete time signals ( *sampled data systems*).

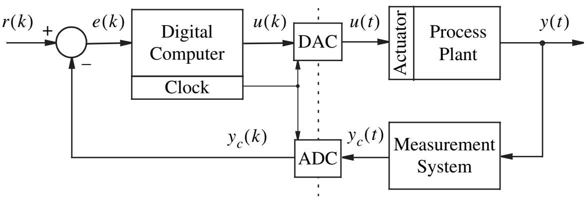
## D.1 Introduction

Many simple feedback control systems are of SISO nature: they have one input and one output. Even control systems integrated in larger and more complex process plants are sometimes considered as SISO systems although they actually interact with other feedback loops in a more or less complicated way. In some cases the influence from the other control loops can be interpreted as disturbances to the SISO system. But that means of course that it is no longer a true SISO systems but a MISO system.

In this overview such problems will not be explored in depth but rather some useful general results for analysis and design of discrete time systems will be presented. Some of the methods are derived directly as discrete time equivalents to the continuous time techniques.

Figure D.1 shows a typical SISO control loop where the controller is implemented as an algorithm in a digital computer. The process plant and the measurement system are continuous time elements and the computer is a discrete time element. The two parts of the system are divided by a *digital-to-analog converter* (DAC) and an *analog-to-digital converter* (ADC). The ADC acts as a sampler which provides the computer with measured values of the output at discrete instances, the *sampling instances*, which are determined by a clock within the computer. At similar instances the digital control values from the computer are converted to a continuous ('analog') control signal (usually a voltage or a current) which is fed to the actuator inputs of the process plant. Note that the time arguments for the signals are the continuous time *t* and the discrete time *k* respectively. *k* is convenient abbreviation for *kT* where *T* is the (usually constant) *sampling period.*

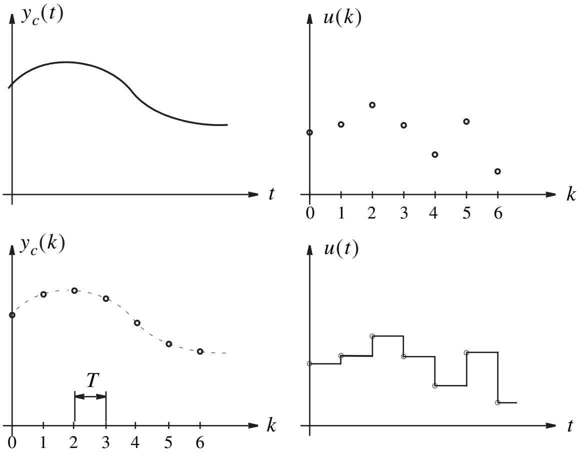
**Fig. D.1** Sampled data control system



The system in Fig. D.1 resembles that of the simplest continuous time SISO control systems. The difference is that the continuous controller has been replaced by a computer and that, consequently, the AD and DA converters have been added.

It is important to note that the discrete time signals ($r(k)$, $e(k)$, $u(k)$) are only defined at the sampling instants, whereas the continuous time signals are defined at all times. Figure D.2 shows examples of signals in the system on Fig. D.1.

**Fig. D.2** Discrete and continuous time signals



Note that the conversion from the discrete values $u(k)$ to the continuous control signal $u(t)$ implies that the $u(t)$ signal is maintained constant between the sampling instants. Since this way of generating the continuous (or rather piecewise continuous) control signal can be called a zero-order interpolation, the DAC is said to contain a *zero-order-hold* ( ZOH).

## D.2 The Sampling Process

If the two left hand curves on Fig. D.2 are combined it is possible to construct a mathematical description of the process of sampling.

Consider Dirac's unit impulse function $\delta(t)$. The function is zero everywhere except at $t = 0$. The impulse function $\delta(t - kT)$ is a similar impulse which 'arrives' at the time $t = kT$, i.e., the $k$'th sampling instant. If the signal $y_c(k)$ is seen as train of pulses arriving at the sampling instants $\ldots(k-1)T,\ kT,\ (k+1)T\ldots$, the discrete signal can be expressed as:

$$y_c^*(t) = y_c(t) \sum_{k=-\infty}^{\infty} \delta(t - kT), \tag{D.1}$$

where the *–notation shows that the entire signal for all $k$ in the interval $[0, \infty]$ is considered.

Since the impulse train function $\sum \delta(t - kT)$ is periodic with the period $T$, it can be expanded into a Fourier-series. It can be shown that the Fourier-series of $y_c^*(t)$ can be written:

$$\sum_{k=-\infty}^{\infty} \delta(t - kT) = \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{jn\omega_s^t}, \tag{D.2}$$

where $\omega_s$ is the sampling frequency in [rad/s],

$$\omega_s = \frac{2\pi}{T}. \tag{D.3}$$

If (D.2) is inserted into (D.1), the expression can be Laplace-transformed (assuming that the functions are Laplace-transformable):[†]

$$\begin{aligned} y_c^*(s) = \mathscr{L}\{y_c^*(t)\} &= \int_{-\infty}^{\infty} y_c(t) \left( \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{jn\omega_s^t} \right) e^{-st} dt \\ &= \frac{1}{T} \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} e^{jn\omega_s^t} y_c(t) e^{-st} dt = \frac{1}{T} \sum_{n=-\infty}^{\infty} \mathscr{L}(e^{jn\omega_s^t} y_c(t)) \\ &= \frac{1}{T} \sum_{n=-\infty}^{\infty} Y_c(s - jn\omega_s). \end{aligned} \tag{D.4}$$

If the Laplace-operator $s$ is replaced by $j\omega$ the frequency spectrum of the sampled signal $y_c^*(t)$ is obtained:

$$Y_c^*(j\omega) = \frac{1}{T} \sum_{n=-\infty}^{\infty} Y_c(j(\omega - n\omega_s)). \tag{D.5}$$

[†] The Laplace-transform used her is the two-sided transform where the integral is taken on the interval $[-\infty, \infty]$. In the more often used one-sided transform the interval is $[0, \infty]$.

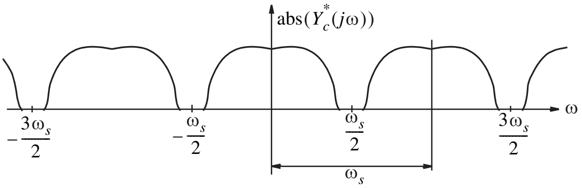


**Fig. D.3** Amplitude spectrum of sampled signal

Thus the spectrum is a sum of the spectra:

$$Y_c^*(j\omega) = \frac{1}{T}(\ldots + Y_c(j(\omega - \omega_s)) + Y_c(j\omega) + Y_c(j(\omega + \omega_s)) + \ldots), \qquad \text{(D.6)}$$

where the 'center-spectrum' $Y_c(j\omega)$ is the spectrum one would get if no sampling had taken place and only the continuous time signal $y_c(t)$ had been considered.

If the numerical value $(Y_c^*(j\omega))$ is calculated the amplitude spectrum is obtained it will look like Fig. D.3. It is obvious that the spectrum is periodic with the period $\omega_s$. The center spectrum is the spectrum of the continuous signal (multiplied by the gain $1/T$) and the side spectra are by products of the sampling process.

If it is desired to reconstruct the continuous signal (i.e., convert $y_c(k)$ back to $y_c(t)$, it will necessary to filter $y_c(k)$ to cut off all the side spectra. The filter required would have to have the amplitude response on Fig. D.4. The filter has the transfer function $H(s)$. Note that for physical reasons only positive frequencies are considered.

Apart from the fact that this ideal filter cannot be realized, it is also clear that the reconstuction of the continuous signal will only be possible if the original continuous signal spectrum is *band limited*. If this not the case, the situation on Fig. D.5a may arise where there is spectural overlap. It will not be possible to reconstruct the original signal here *even if* the ideal filter could physically be made. If the continuous signal contains frequencies higher than half the sample frequency, it will *never* be possible to reconstruct the signal after the sampling because the side spectra will 'spill' amplitude content into the frequency interval

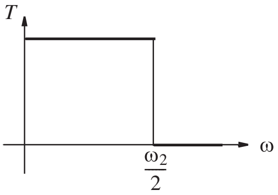
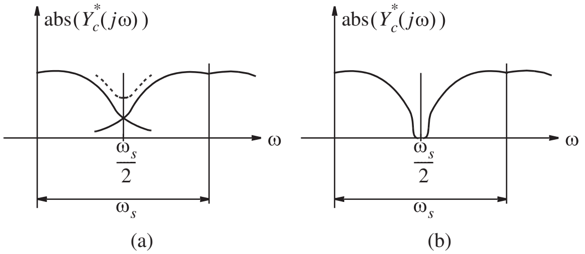


**Fig. D.4** Ideal reconstruction filter transfer function

**Fig. D.5** The effect of non band limited signals



$[0, \omega_s/2]$ and there is no way to remove this content. This phenomenon is called *aliasing*. The frequency,

$$\omega_N = \omega_s/2, \tag{D.7}$$

is called the *Nyquist frequency*. So it is seen that the continuous signal must not contain frequencies above the Nyquist frequency. This statement is often called the *Shannon sample theorem*.

The reconstruction filter usually applied is the zero-order-hold. Figure D.6 shows the impulse response of the ZOH. The impulse response is a *unit pulse*. The time response is composed of two unit step functions, the last one negative and with the delay $T$:

$$y(t) = h(t) - h(t - T). \tag{D.8}$$

.

The Laplace transform of the impulse response is the transfer function and is readily found as

$$G_h(s) = \frac{1}{s} - \frac{1}{s}e^{-Ts} = \frac{1 - e^{-Ts}}{s}. \tag{D.9}$$

The frequency response of this filter can be determined in the usual way. If $s$ is replaced by $j\omega$ it is seen that

$$G_h(j\omega) = T\frac{\sin\frac{\omega T}{2}}{\frac{\omega T}{2}}e^{-j\frac{\omega T}{2}}. \tag{D.10}$$

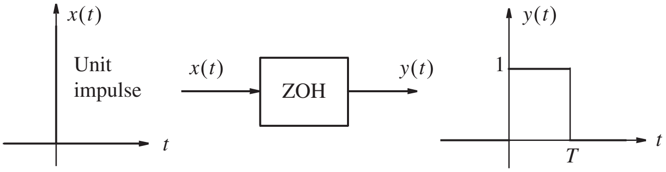


**Fig. D.6** Zero-order-hold (ZOH) time response

The amplitude ratio is the absolute value of this complex number:

$$A(\omega) = \text{abs}(G_h(j\omega)) = T\frac{\text{abs}\left[\sin\frac{\omega T}{2}\right]}{\frac{\omega T}{2}}. \tag{D.11}$$

$A(\omega)$ as a function of the normalized frequency $\omega/\omega_s$ can be seen on Fig. D.7.

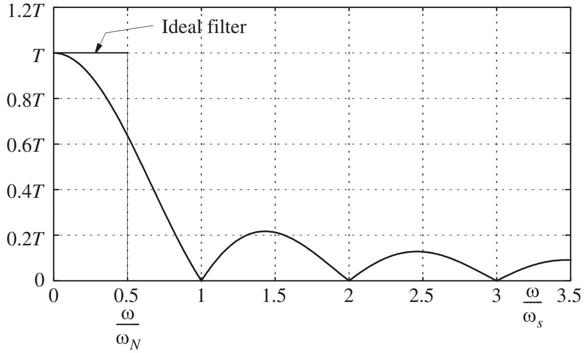


**Fig. D.7** Amplitude response of a zero-order-hold

The gain is seen to deviate significantly from that of an ideal filter.

The problem of aliasing can be solved (or at least reduced) by the application of an *antialiasing filter*. This is a continuous time filter (an 'analog' filter) that is inserted in the measurement channel in front of the sampler, i.e., between the measurement system and the ADC, see Fig. D.8. The filter should have a steep roll off starting below the Nyquist frequency, assuring that the frequencies above $\omega_N$ are removed or only remain present with very small amplitudes so that the more favorable situation shown on figure D.5b can be achieved.

In the time domain it is easy to see what effect aliasing has on the reconstructed signal. Figure D.9 shows a sinusoidal signal with the frequency $\omega_c = 5.65\,\text{rad/sec}$ ($f_c = 0.9\,\text{Hz}$). The signal is sampled with the sampling frequency $\omega_s = 6.28\,\text{rad/sec}$ which means that the Nyquist frequency is $\omega_N = 3.14\,\text{rad/sec}$. Since $\omega_c > \omega_N$, the sampling theorem is clearly violated.

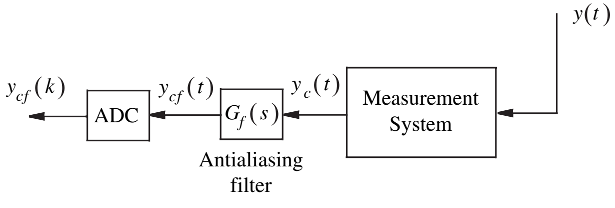


**Fig. D.8** Application of an antialiasing filter

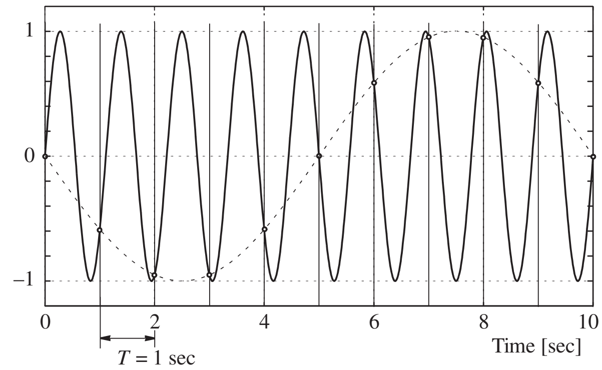


**Fig. D.9** Sampling of a sinusoidal signal with aliasing

The frequency detected in the sampled signal (see the dots on figure D.9) is 0.628 rad/sec (0.1 Hz). This is what one should expect because the signal frequency is 'folded back' into the interval [0, $\omega_N$] symmetrically relative to $\omega_N$, see Fig. D.10.

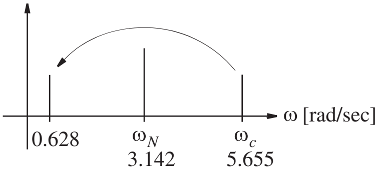


**Fig. D.10** 'Folding back' of the signal frequency above the Nyquist frequency (aliasing)

## D.3 The *Z*-Transform

If a continuous time signal $y(t)$ is sampled with the sample period $T$, the Laplace transform of the sampled signal can be written (see (D.1) and (D.4)):

$$\begin{aligned} y^*(s) = \mathscr{L}\{y^*(t)\} &= \int_{-\infty}^{\infty} y(t) \sum_{k=-\infty}^{\infty} \delta(t-kT)e^{-st}dt \\ &= \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} y(t)\delta(t-kT)e^{-st}dt. \end{aligned} \tag{D.12}$$

The impulse function has the property that

$$\int_{-\infty}^{\infty} \delta(t-kT)x(t)dt = x(kT). \tag{D.13}$$

Applying this rule to (D.12) yields:

$$Y^*(s) = \sum_{n=-\infty}^{\infty} y(kT)e^{-skT}, \tag{D.14}$$

where the last relation is valid because only function values at the sampling instants $kT$ are defined for the sampled signal.

Now a new complex variable is introduced. It is defined as

$$z = e^{Ts}. \tag{D.15}$$

This allows a new notion to be established: the $z$-transform of the sampled signal. It is derived directly from (D.14) and (D.15):

$$Z\{y^*(t)\} = Y(z) = \sum_{k=-\infty}^{\infty} y(kT)z^{-k} \tag{D.16}$$

If $y(t) = 0$ for $t < 0$, which is often the case, (D.16) is reduced to:

$$Y(z) = \sum_{k=0}^{\infty} y(kT)z^{-k}. \tag{D.17}$$

Expansion of (D.17) gives the result:

$$Y(z) = y(0) + y(T)z^{-1} + y(2T)z^{-2} + \ldots + y(nT)z^{-n} + \ldots. \tag{D.18}$$

***Example D.1.* Discretization of Time Functions**

The step function on Fig. D.11a can easily be $z$-transformed using the definition (D.17) and applying a sum-formula for infinite series,

$$X_1(z) = \sum_{k=0}^{\infty} 1 \cdot z^{-k} = 1 + z^{-1} + z^{-2} + \ldots = \frac{z}{z-1}. \tag{D.19}$$

The pulse function on Fig. D.11b with the period $T$ will have exactly the same $z$-transform. This will be the case for all continuous signals which coincide at the sampling instants. ❐

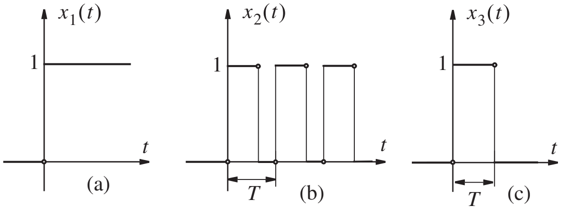


**Fig. D.11** Step and pulse functions

***Example D.2*. Discretization of Pulse Time Functions**

For a unit pulse as shown on Fig. D.11c is found that:

$$X_3(z) = \sum_{k=0}^{\infty} 1 \cdot z^{-k} = 1 + 0 \cdot z^{-1} + 0 \cdot z^{-2} + \ldots = 1. \tag{D.20}$$

❐

***Example D.3*. Discretization of a First Order Low Pass Filter**

For the exponential function,

$$y(t) = \begin{cases} 0 & \text{for } \ \mathrm{t}<0 \\ e^{-at} & \text{for } \ \mathrm{t}\geq 0\text{’} \end{cases} \tag{D.21}$$

is known that the Laplace transform is

$$Y(s) = \frac{1}{s+a}. \tag{D.22}$$

The $z$-transform can be calculated to be

$$Y(z) = \sum_{k=0}^{\infty} e^{-akT} z^{-k} = 1 + e^{-aT} z^{-1} + e^{-2aT} z^{-2} + \ldots = \frac{z}{z - e^{-aT}}. \tag{D.23}$$

❐

A short table of $z$-transforms and the corresponding Laplace transforms are given in Table D.1.

From the definition expression it is possible to deduce the following properties of the $z$-transform.

1. The constant rule:

   $$Z\{af(t)\} = aF(z).$$

2. Linearity rule:

   $$Z\{af(t) + bg(t)\} = aF(z) + bG(z).$$

3. Backward shift or delay theorem:

   $$Z\{f(t - nT)\} = z^{-n}F(z) + \sum_{k=0}^{n-1} f((k-n)T)z^{-k}.$$

4. Forward shift theorem:

   $$Z\{f(t - nT)\} = z^{n}F(z) - z^{n} \sum_{k=0}^{n-1} f(kT)z^{-k}.$$

**Table D.1** Laplace and Z-transform Pairs

| | $f(t)$ or $f(k)$ | $F(s)$ | $F(z)$ |
|---|---|---|---|
| 1 | Unit impulse $\delta(t)$ | $1$ | $1$ |
| 2 | $\delta(t-nT)$ | $e^{-nTs}$ | $z^{-n}$ |
| 3 | Unit step $h(t)$ | $\frac{1}{s}$ | $\frac{z}{z-1}$ |
| 4 | Unit ramp $t$ | $\frac{1}{s^2}$ | $\frac{Tz}{(z-1)^2}$ |
| 5 | $t^2$ | $\frac{2}{s^3}$ | $\frac{T^2z(z+1)}{(z-1)^3}$ |
| 6 | $e^{-at}$ | $\frac{1}{s+a}$ | $\frac{z}{z-e^{-aT}}$ |
| 7 | $\frac{1}{a}(1-e^{-at})$ | $\frac{1}{s(s+a)}$ | $\frac{(1-e^{-aT})z}{a(z-1)(z-e^{-aT})}$ |
| 8 | $e^{-at}-e^{-bt}$ | $\frac{b-a}{(s+a)(s+b)}$ | $\frac{(e^{-aT}-e^{-bT})z}{(z-e^{-aT})(z-e^{-bT})}$ |
| 9 | $a^k$ | | $\frac{z}{z-a}$ |
| 10 | $a^k \cos k\pi$ | | $\frac{z}{z+a}$ |
| 11 | $\sin \omega t$ | $\frac{\omega}{s^2+\omega^2}$ | $\frac{z \sin \omega T}{z^2-2z\cos \omega T+1}$ |
| 12 | $\cos \omega t$ | $\frac{s}{s^2+\omega^2}$ | $\frac{z(z-\cos \omega T)}{z^2-2z\cos \omega T+1}$ |
| 13 | $e^{-at}\sin \omega t$ | $\frac{\omega}{(s+a)^2+\omega^2}$ | $\frac{e^{-aT}z\sin \omega T}{z^2-2e^{-aT}z\cos \omega T+e^{-2aT}}$ |
| 14 | $e^{-at}\cos \omega t$ | $\frac{s+a}{(s+a)^2+\omega^2}$ | $\frac{z^2-e^{-aT}z\cos \omega T}{z^2-2e^{-aT}z\cos \omega T+e^{-2aT}}$ |

5. Initial value theorem:

$$f(0)=\lim_{z\to\infty} F(z).$$

6. Final value theorem:

$$\lim_{k\to\infty} f(k)=\lim_{z\to 1}(1-z^{-1})F(z).$$

This rule is only valid if all poles of $(1-z^{-1})F(z)$ are inside the unit circle.

## D.4 Inverse $Z$-Transform

The simplest method for generating the discrete time signal $x(k)$ from the $z$-transform $X(z)$ is to use the result (D.18).

***Example D.4.*** **Discrete Transfer Function Step Time Response**

Consider the $z$-transform,

$$X(x) = \frac{0.2858z}{z^2 - 0.4500z + 0.03020}. \tag{D.24}$$

Dividing the numerator and denominator by the highest power of $z$,

$$X(z) = \frac{0.2858z^{-1}}{1 - 0.4500z^{-1} + 0.03020z^{-2}},$$

and carrying out a polynomial division,

$$0.2858Z^{-1}/(1 - 0.4500z^{-1} + 0.03020z^{-2}) = 0.2858z^{-1} + 0.1286z^{-2}$$
$$+0.04923z^{-3} + 0.01827z^{-4} + 0.006734z^{-5} + 0.002478z^{-6} + \ldots.$$

According to (D.18) the coefficients of the right hand polynomial in $z^{-1}$ are the values of $x(k)$ at the sample instants indicated by the negative power of $z$.

The initial value theorem shows that the value $x(0)$ is,

$$x(0) = \lim_{z\to\infty} \frac{0.2858z^{-1}}{1 - 0.4500z^{-1} + 0.03020z^{-2}} = 0,$$

and the sequence of $x(k)$-values are:

$$\begin{aligned} x(0) &= 0, \\ x(1) &= 0.2858, \\ x(2) &= 0.1286, \\ x(3) &= 0.04923, \\ x(4) &= 0.01827, \\ x(5) &= 0.006734, \\ x(6) &= 0.002478, \\ &\vdots \end{aligned} \tag{D.25}$$

This sequence can be seen on Fig. D.12.

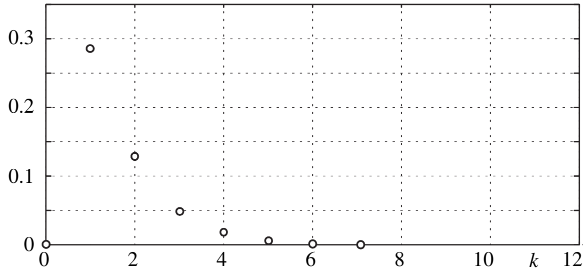


**Fig. D.12** The sequence $x(k)$ of equation (D.25)

The $z$-transform (D.24) is actually the transform in entry 8 of Table D.1 with $a = 2$, $b = 5$ and $T = 0.5$ s. If the time instants $t = 0, T, 2T, 3T, \ldots$ are inserted into the left hand continuous function $e^{-at} - e^{-bt}$ in the table, exactly the same function values as equation (D.25) will be found. ❐

An alternative method for inverse $z$-transform is based on backwards application of the $z$-transform table. The method is also known from calculation of the Laplace transform.

***Example D.5*. Discrete Transfer Function Explicit Time Response**

The $z$-transform of equation (D.24) is considered again. The first step is a partial fraction expansion into terms which can be found in the $z$-transform table.

Since almost all $z$-transforms have the factor $z$ in the numerator, it makes the fraction expansion a little easier if $X(z)/z$ is expanded instead of $X)z)$. In this case one obtains:

$$\frac{X(z)}{z} = \frac{0.2858}{z^2 - 0.4500z + 0.03020} = \frac{0.2858}{(z - 0.3679)(z - 0.08209)}$$
$$= \frac{1}{z - 0.3679} - \frac{1}{z - 0.08209},$$

which means that

$$X(z) = \frac{z}{z - 0.3679} - \frac{z}{z - 0.08209}.$$

The two last terms can be found in Table D.1, row 6. When the sample period is known ( $T = 0.5$ s), it is immediately seen that

$$a = -\frac{\log_e 0.3679}{0.5} = 2 \text{ and } b = -\frac{\log_e 0.08209}{0.5} = 5,$$

and that the values *at the sampling instants* are

$$x(kT) = e^{-2kT} - e^{-5kT}.$$

❐

## D.5 Discrete Transfer Functions

An input-output relationship for a linear time invariant continuous time system is usually expressed as a differential equation of the form:

$$\begin{aligned} &a_n y^{(n)}(t) + a_{n-1} y^{(n-1)}(t) + \ldots + a_0 y(t) \\ &\quad = b_m u^{(m)}(t) + b_{m-1} u^{(m-1)}(t) + \ldots + b_0 u(t). \end{aligned} \tag{D.26}$$

Laplace transforming this equation leads to the continuous transfer function:

$$G(s) = \frac{Y(s)}{U(s)} = \frac{b_m s^m + \ldots + b_1 s + b_0}{a_n s^n + \ldots + a_1 s + a_0}. \tag{D.27}$$

For a discrete time system a difference equation is the natural way to represent a system model,

$$\begin{aligned} &y(k) + a_1 y(k-1) + a_2 y(k-2) + \ldots + a_n y(k-n) \\ &\quad = b_0 x(k) + b_1 x(k-1) + b_2 x(k-2) + \ldots + b_l x(l-l). \end{aligned} \tag{D.28}$$

This equation is $z$-transformed by applying the backward shifting theorem to each term:

$$\begin{aligned} &Y(z) + a_1 z^{-1} Y(z) + a_2 z^{-2} Y(z) + \ldots + a_n z^{-n} Y(z) \\ &\quad = b_0 X(z) + b_1 z^{-1} X(z) + b_2 z^{-2} X(z) + \ldots + b_l z^{-l} X(z). \end{aligned} \tag{D.29}$$

Factoring Y(z) X(z) out and dividing on both sides of the equal sign gives

$$\frac{Y(z)}{X(z)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_l z^{-l}}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_n z^{-n}} = H(z) \tag{D.30}$$

.

Multiplying numerator and denominator by $z^n$ gives the alternative form,

$$\frac{Y(z)}{X(z)} = \frac{b_0 z^n + b_1 z^{n-1} + b_2 z^{n-2} + \ldots + b_l z^{n-l}}{z^n + a_1 z^{n-1} + a_2 z^{n-2} + \ldots + a_n} = \frac{B(z)}{A(z)} = H(z). \tag{D.31}$$

$H(z)$ is called the *discrete time transfer function* for the system modelled by equation (D.28). The transfer function is usually a rational fraction with polynomials $B(z)$ and $A(z)$ as numerator and denominator.

The solutions of the equation,

$$B(z) = 0, \tag{D.32}$$

are called the *zeros* of the transfer function and the solutions to

$$A(z) = 0 \tag{D.33}$$

are the *poles*. The denominator polynomial $A(z)$ is called the *characteristic polynomial* of $H(z)$ and Eq. (D.33) is the *characteristic equation*.

Just as in the continuous case it is now possible to express an input–output relationship using the transfer function,

$$Y(z) = H(z)X(z). \tag{D.34}$$

## D.6 Discrete Systems and Difference Equations

The output expression (D.34) can be written:

$$Y(Z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_l z^{-l}}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_n z^{-n}} X(z). \tag{D.35}$$

Simple manipulation of this equation leads to

$$\begin{aligned} Y(z) = {} & (b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_l z^{-1}) X(z) \\ & - (a_1 z^{-1} + a_2 z^{-2} + \ldots + a_n z^{-n}) Y(z). \end{aligned} \tag{D.36}$$

Inverse $z$-transformation gives (only values at the sampling instants $kT$ are relevant),

$$\begin{aligned} y(k) = {} & b_0 x(k) + b_1 x(k-1) + b_2 x(k-2) + \ldots + b_l x(k-l) \\ & - a_1 y(k-1) - a_2 y(k-2) - \ldots - a_n y(k-n) \end{aligned} \tag{D.37}$$

Equation (D.37) is a difference equation which allows computation of $y$ at the time $t = kT$ when $y$ is known at the times $t = (k-1)T$, $(k-2)T, \ldots, (k-n)T$ and $x$ is known at $t = kT$, $(k-1)T$, $(k-2)T, \ldots, (k-l)T$.

If the transfer function is considered as a discrete (or 'digital') filter then the difference equation is very easy to implement on a computer. The computation at each sampling instant requires that at the most $n$ previous $y$-values and $l$ previous $x$-values are stored in the computer's memory. If $b_0 \neq 0$ the $x$-value must be known at the time of the calculation.

## D.7 Discrete Time Systems with Zero-Order-Hold

In many cases it is required to describe the entire continuous part of the hybrid system on Fig. D.1 in the discrete time domain. This means that it is necessary to generate a discrete model equivalent to the continuous time part of Fig. D.1. This is shown in a slightly simplified form on Fig. D.13a. The discrete equivalent is shown on Fig. D.13b.

The problem is to find the z-transform of the continuous transfer function preceded by a ZOH. The transfer function of the ZOH is (D.9). So what has to be determined is

$$H_{ZOH}(z) = Z\left\{\frac{1 - e^{-Ts}}{s} G(s)\right\}. \tag{D.38}$$

One finds by direct calculation that

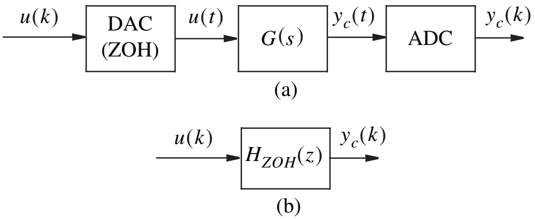


**Fig. D.13** Continuous and discrete time model equivalents

$$
\begin{aligned}
H_{ZOH}(z) &= Z\left\{\frac{G(s)}{s} - e^{-Ts}\frac{G(s)}{s}\right\} = Z\left\{\frac{G(s)}{s}\right\} - Z\left\{e^{-Ts}\frac{G(s)}{s}\right\} \\
&= Z\left\{\frac{G(s)}{s}\right\} - z^{-1}Z\left\{\frac{G(s)}{s}\right\},
\end{aligned}
\tag{D.39}
$$

where the last result is due to the fact that $e^{-Ts}$ gives a delay of one sample period, and therefore the backward shift theorem can be used. The final result is

$$
H_{ZOH}(z) = (1 - z^{-1})Z\left\{\frac{G(s)}{s}\right\}. \tag{D.40}
$$

## D.8 Transient Response, Poles and Stability

The transient response of a discrete system can be calculated from (D.34).

The simplest case arises if the input function $X(t)$ is the unit impulse function on figure D.11c because its $z$-transform is $X(z) = 1$. In this case the transient response becomes the transfer function itself and the discrete time function is found simply by inverse transformation as in examples D.4 and D.5. So the response on Fig. D.12 is the unit pulse response of a system with the transfer function (D.24).

***Example D.6*. Continuous System Response with Zeroth Order Hold**

Given the continuous system with transfer function,

$$
G(s) = \frac{b - a}{(s + a)(s + b)} = \frac{3}{(s + 2)(s + 5)}, \tag{D.41}
$$

with a zero-order-hold added the discrete transfer function is obtained,

$$
H_{ZOH}(z) = (1 - z^{-1})Z\left\{\frac{3}{s(s + 2)(s + 5)}\right\}.
$$

The unit step response is found from

$$\begin{aligned}Y(z) &= H_{ZOH}(z)X(s) = (1-z^{-1})Z\left\{\frac{3}{s(s+2)(s+5)}\right\}\frac{z}{z-1}\\ &= Z\left\{\frac{3}{s(s+2)(s+5)}\right\} = Z\left\{\frac{0.3}{s}-\frac{0.5}{s+2}+\frac{0.2}{s+5}\right\}\\ &= 0.3\frac{z}{z-1}-0.5\frac{z}{z-e^{-2T}}+0.2\frac{z}{z-e^{-5T}}.\end{aligned}$$

With the sample period $T = 0.2$ s it is found that

$$Y(z) = \frac{0.03842z^2+0.02410z}{z^3-2.0382z^2+1.2848z-0.2466}.$$

The inverse $z$-transform can be found by polynomial division:

$$\begin{aligned}y(0) &= 0,\\ y(1) &= 0.03842,\\ y(2) &= 0.1024,\\ y(3) &= 0.1594,\\ y(4) &= 0.2027,\\ &\vdots\end{aligned}$$

The result is seen on Fig. D.14. On the same figure is drawn the unit step response for the continuous system (D.41).

It is clearly seen that the two responses coincide at the sampling instants. ❐

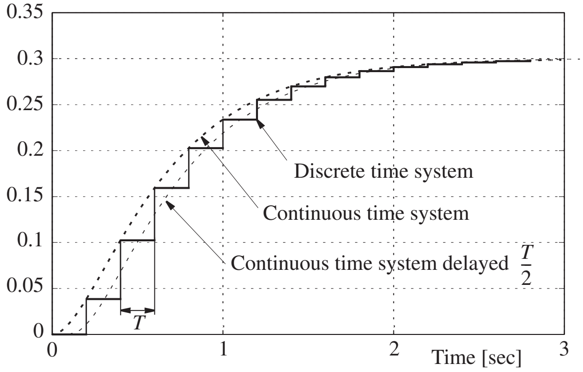


**Fig. D.14** Unit step response for discrete and continuous system

As seen from the $z$-transform table there is a simple correspondence between the poles of the Laplace and the $z$-transforms.

If the continuous signal (or system) has a pole at $s_p$ and the equivalent discrete signal (or system) has a pole at $z_p$, then they are related as follows:

$$z_p = e^{Ts_p} \Leftrightarrow s_p = \frac{1}{T}\log_e z_p. \tag{D.42}$$

This is in agreement with the definition of the $z$-transform (D.15). The zeros of the signals and transfer functions do not follow this simple rule.

If the poles are real, $s_p = r$ , the discrete poles will also be real, and the relation is simply

$$z_p = e^{Tr} \Leftrightarrow r = \frac{1}{T}\log_e z_p. \tag{D.43}$$

If the poles are complex,

$$\begin{aligned} s_p &= \sigma \pm j\omega, \\ z_p &= \alpha \pm j\beta, \end{aligned} \tag{D.44}$$

then the relation will be

$$\begin{gathered} z_p = e^{T\sigma} \cdot e^{jT\omega} \\ \text{or} \\ \alpha \pm j\beta = e^{T\sigma}\cos T\omega \pm je^{T\sigma}\sin T\omega \Leftrightarrow \sigma \pm j\omega = \frac{1}{T}\log_e \sqrt{\alpha^2+\beta^2} \pm j\frac{1}{T}\arctan\frac{\beta}{\alpha}. \end{gathered} \tag{D.45}$$

The upper expression in (D.45) shows that if $\sigma < 0$ then Mod $z_p = \sqrt{\alpha^2+\beta^2} < 1$. In other words, if the continuous poles are in the left half plane, then the discrete poles are inside the unit circle disc.

This leads to a very simple stability criterion for discrete time systems:

A discrete time system given by the transfer function (D.30) and (D.31) is asymptotically stable if and only if all poles are inside the unit circle.

A mapping of the continuous $s$-plane on the discrete $z$-plane can be carried out from the above relations. It is shown in Fig. D.15.

Note that the curves in the $z$-plane for constant damping ratio are logarithmic spirals.

A more accurate picture of the curves for constant damping ratio and for constant natural frequency is shown on Fig. D.16

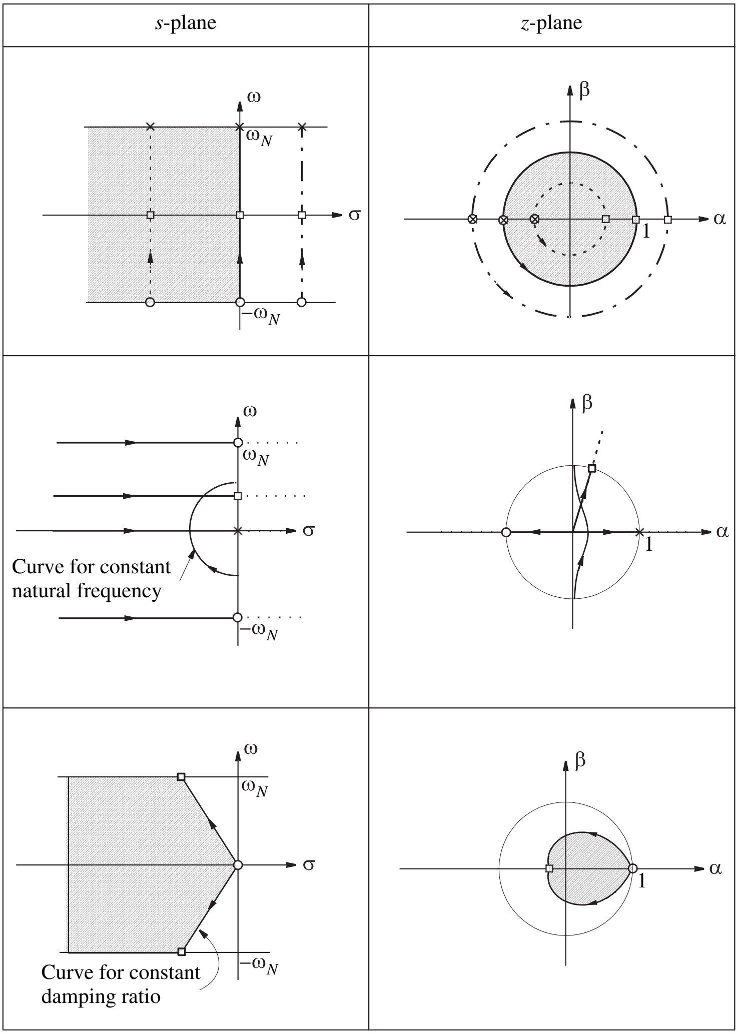

s-plane
z-plane
ω
ω_N
σ
−ω_N
β
α
1
ω
ω_N
σ
Curve for constant natural frequency
−ω_N
β
α
1
ω
ω_N
σ
Curve for constant damping ratio
−ω_N
β
α
1

**Fig. D.15** The correspondence between s-plane and z-plane (z-transform)

## D.9 Frequency Response

For continuous systems the frequency response can be found by replacing the Laplace operator $s$ by $j\omega$ in the transfer function $G(s)$ and then calculating the complex number for appropriate frequencies in the interval $[0, \infty]$, i.e., $s$ is given values on the positive half of the imaginary axis. The amplitude ratio and the phase are $\mathrm{Abs}(G(j\omega))$ and $\mathrm{Arg}(G(j\omega))$ respectively.

The procedure is quite similar for discrete time systems. In the transfer function $H(z)$ the operator $z$ is replaced by $e^{jT\omega}$. The frequencies must now be kept

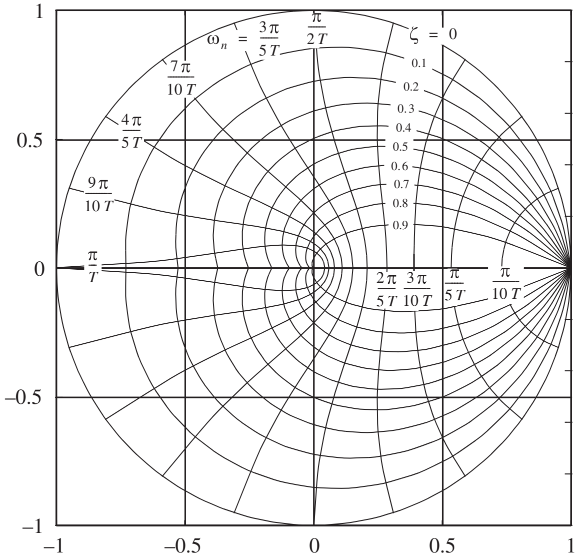


**Fig. D.16** Discrete time (z-plane) curves for constant damping ratio and natural frequency

to the interval $\omega \in [0, \omega_s/2]$ since it does not make sense to expose a discrete time system to frequencies higher than the Nyquist frequency. In other words, $z$ is given values on the upper half of the unit circle. See also Fig. D.15.

### *Example D.7.* Bode Plots for a Continuous and a Discrete System

The continuous system from Example D.6 has the transfer function,

$$G(s) = \frac{3}{(s+2)(s+5)}. \tag{D.46}$$

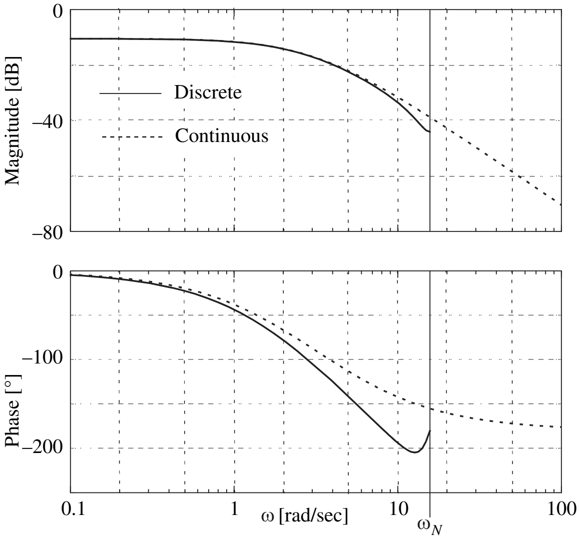
With a ZOH the discrete time transfer function was found to be

$$\begin{aligned} H_{ZOH}(z) &= 0.03843\frac{(z-1)(z^2+0.6274z)}{z(z^3-2.0382z^2+1.2848z-0.2466)} \\ &= 0.03842\frac{z+0.6274}{z^2-1.0382z+0.2466}, \end{aligned} \tag{D.47}$$

with $T = 0.2$ s and $\omega_N = \frac{\omega_s}{2} = 15.17$ rad/s rad/s.

On Fig. D.17 can be seen the Bode plot for the discrete system (D.47) and the continuous system (D.46). The magnitude curves only deviate slightly from each other close to the Nyquist frequency. For the phase the situation is different. The discrete system has a larger phase lag than the continuous system. A little below the Nyquist frequency more than 200° is seen as opposed to 155°. As a matter of fact this is what could be expected. Looking at the step responses

**Fig. D.17** Bode plots for discrete and continuous system



for the same systems on Fig. D.14, it is clearly seen that the ‘best’ continuous approximation to the discrete staircase response of the discrete system is the thin dashed curve shown on the figure. This curve is the continuous response *delayed* half the sample period, $T/2$. So the sampling and the zero-order-hold process can be *approximated* with a delay of $T/2$. A delay generates a phase lag proportional to the frequency. The extra phase lag for a time lag of $T/2$ is,

$$\varphi_d = \frac{\omega T}{2}, \tag{D.48}$$

which provides a reasonable explanation to the result on Fig. D.17. ❐

## D.10 Discrete Approximations to Continuous Transfer Functions

Design of discrete time controllers for simple feedback systems can be accomplished by application of root locus or frequency response methods to the discrete transfer functions. However these techniques are easier to use and they are also better known for the continuous domain. It is therefore common to design standard controllers (P, PI, PID) as continuous controllers and then to ‘translate’ the transfer functions in $s$ to an equivalent transfer function in $z$ which allows easy implementation as a computer algorithm as discussed in Sect. D.6.

In principle, the translation could be done by replacing $s$ by $z$ according to (D.15) but unfortunately that would lead to non-rational transfer functions for which simple inverse transforms cannot be found.

## D.10.1 Tustin Approximation

Instead of direct substitution a rational approximation to the exponential function is used. A well known example of such a simple approximation is the first order Padé-approximation:

$$z = e^{Ts} \cong \frac{1 + Ts/2}{1 - Ts/2}. \tag{D.49}$$

The inverse expression,

$$s \cong \frac{2}{T}\frac{z-1}{z+1}, \tag{D.50}$$

Is called the *Tustin approximation*.
The continuous transfer function,

$$G(s) = \frac{\prod_{i=1}^{m}(s + s_{zi})}{\prod_{i=1}^{n}(s + s_{pi})}, \tag{D.51}$$

has $m$ zeros $-s_{z1}, -s_{z2}, \ldots, -s_{zm}$ and $n$ poles $-s_{p1}, -s_{p2}, \ldots, -s_{pn}$.
The corresponding approximate discrete transfer function is

$$H_T(z) \cong G(s)|_{s=\frac{2z-1}{Tz+1}} = \left(\frac{T}{2}(z+1)\right)^r \frac{\prod_{i=1}^{m}\left[\left(1 + \frac{T}{2}s_{zi}\right)z - \left(1 - \frac{T}{2}s_{zi}\right)\right]}{\prod_{i=1}^{n}\left[\left(1 + \frac{T}{2}s_{pi}\right)z - \left(1 - \frac{T}{2}s_{pi}\right)\right]} \tag{D.52}$$

where $r = n - m \geq 0$ is called *the relative order* of the transfer function. If $G(s)$ has no zeros ($M = 0$), the numerator of (D.52) is replaced by the constant numerator of $G(s)$. Note that $H_T(z)$ has always the same number poles and zeros (namely $n$) and that $r$ of the zeros are in the same position: $z = -1$.

All the poles and the $m$ zeros of (D.52) can be found directly from the continuous poles and zeros by using (D.49). The additional discrete zeros in $z = -1$ of (D.52) cannot be found in (D.51). If one insists on translating these zeros with (D.50), one finds that $s_{zi} \cong \infty$. In this context it is sometimes said that (D.51) has *r zeros at infinity*.

The mapping of the $s$-plane into the $z$-plane looks like Fig. D.18 (similar to Fig. D.17 for the $z$-transform).

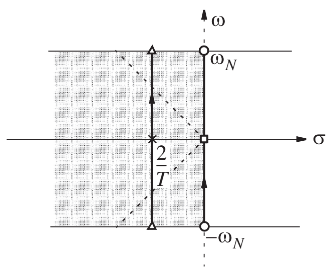
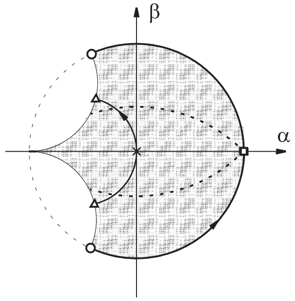
| *s*-plane | *z*-plane |
|---|---|
|  |  |

**Fig. D.18** The correspondence between s-plane and z-plane (Tustin-approx.)

***Example D.8*. Tustin Approximation of a Continuous System**

For the continuous system from Examples D.6–D.7 with the transfer function (D.41) it can immediately be found that (for $T=0.2$ s),

$$\begin{aligned} H_T(z) &= (0.1(z+1))^2 \frac{3}{(1.2z-0.8)(1.5z-0.5)} \\ &= \frac{0.01667z^2 + 0.03333z + 0.01667}{z^2 - z + 0.2222}. \end{aligned} \tag{D.53}$$

The poles are

$$z_p = \begin{cases} 0.6667 \\ 0.3333 \end{cases}$$

which could also have found directly from (D.49).

The transient behaviour of $H_T(z)$ is discussed in Example D.9. ❐

## D.10.2 Matched-Pole-Zero Approximation (MPZ)

This approximation method allows generation of discrete transfer function equivalent in a very simple way. The method can be said to be intermediate between the exact *z*-transform and the Tustin approximation.

The MPZ equivalent to a continuous transfer function *G(s)* with *m* zeros and *n* poles is constructed according to the following set of heuristic rules.

1. A continuous pole $s_p$ is mapped into the discrete pole $z_p = e^{Ts_p}$.
2. A finite zero $s_z$ is also mapped into the zero $z_p = e^{Ts_z}$.
3. The *r* infinite zeros of *G(s)* are mapped into zeros in $z=-1$.

Sometimes the number of zeros in $z=-1$ is reduced to $r-1$. This means that the discrete transfer function will have a *pole surplus* of one. The effect is that the difference equation will only contain input terms at time $k-1$ and earlier. This can be advantageous from a computation time point of view. See Example D.9.

4. The static gain of the discrete transfer function *H(z)* is selected equal to that of the continuous transfer function, i.e., such that

$$H_{MPZ}(Z)|_{z=1} = G(S)|_{s=0} \tag{D.54}$$

❐

***Example D.9*. Matched Pole and Zero (MPZ) Approximation**

The poles of the transfer function (D.41) are

$$s_p = \begin{cases} -2 \\ -5 \end{cases} .$$

The poles of the MPZ-equivalent found from rule 1 above are

$$z_p = \begin{cases} 0.6703 \\ 0.3679 \end{cases} .$$

*G(s)* has no finite zeros and consequently $H_{MPZ}(z)$ should have two zeros in $z{-}1$. This leads to

$$H_{MPZ}(z) = K\frac{(z+1)^2}{(z-0.6703)(z-0.3679)} . \tag{D.55}$$

Equation (D.54) is used for calculation of *K*,

$$K\frac{4}{0.3297 \cdot 0.6321} = \frac{3}{2 \cdot 5}$$

$$\Rightarrow K = 0.01563,$$

and finally:

$$H_{MPZ}(z) = \frac{0.1563z^2 + 0.03126z + 0.01563}{z^2 - 1.0382z + 0.2466} . \tag{D.56}$$

If the unit step response is computed for the three discrete transfer functions (D.47), (D.53) and (D.56), the plots on Fig. D.19 are obtained. The Tustin and

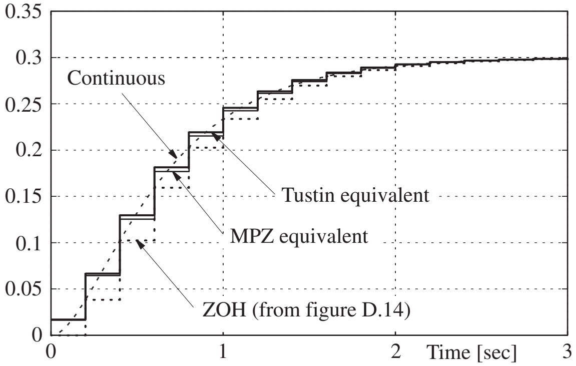


**Fig. D.19** Unit step responses for discrete equivalents

the MPZ equivalents only deviate very little from each other but they are both somewhat different from the ZOH response. As a matter of fact, they are a closer approximation to the continuous response (the thin dashed curve) than the ZOH response.

If one of the zeros in $z=-1$ is omitted in $H_{MPZ}(z)$ a modified version of the MPZ-equivalent is obtained,

$$H_{MMPZ}(z) = K\frac{z+1}{(z-0.6703)(z-0.3679)} = \frac{0.03126z+0.03126}{z^2-1.0382z+0.2466}. \qquad \text{(D.57)}$$

Note that $K$ has been recalculated accordingly. The step responses for $H_{MPZ}(z)$ and $H_{MMPZ}(z)$ are shown on Fig. D.20.

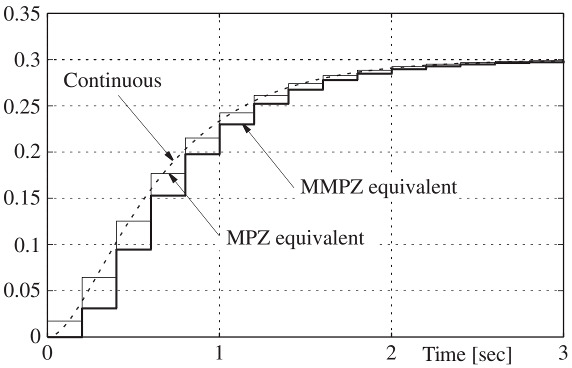


**Fig. D.20** Unit step responses for $H_{MPZ}(z)$ and $H_{MMPZ}(z)$

It is clearly seen that the output $y(0)$ is zero for the latter transfer function but not for $H_{MPZ}(z)$.

If the responses $y(k)$ are found by inverse z-transformation from (D.56) and (D.57) the following results are obtained:

MPZ (nominator and denominator have the same order, i.e., no pole surplus):

$$\begin{aligned} y(k) &= 0.01563x(k) + 0.03126z(k-1) + 0.01563x(k-2) \\ &\quad + 1.0382y(k-1) - 0.2466y(k-2). \end{aligned} \qquad \text{(D.58)}$$

MMPZ (pole surplus):

$$\begin{aligned} y(k) &= 0.03126x(k-1) + 0.03126x(k-2) \\ &\quad + 1.0382y(k-1) - 0.2466y(k-2). \end{aligned} \qquad \text{(D.59)}$$

If the transfer functions should be implemented as an algorithm in a computer these difference equations show exactly how it should be done. The significant difference is that in the case of MPZ the input $x(k)$ at the *current sampling instant kT* is needed for the calculation of $y(k)$. This means that the calculation time must be very short compared to the sampling period.

For the MMPZ only $x$-values up to time $k-1$ are necessary for the calculation of $y(k)$ and therefore the entire sampling period is available for the calculation. This is an advantage if the sampling period must be short compared to the computer's calculation speed. That will be the case if the transfer function is some kind of a discrete filter for a fast process. ❒

***Example D.10.* Tustin and MPZ Approximation Comparison**

The frequency response for (D.41) and its discrete time equivalents are seen on Fig. D.21. As expected the Tustin and MPZ equivalents have quite similar frequency responses.

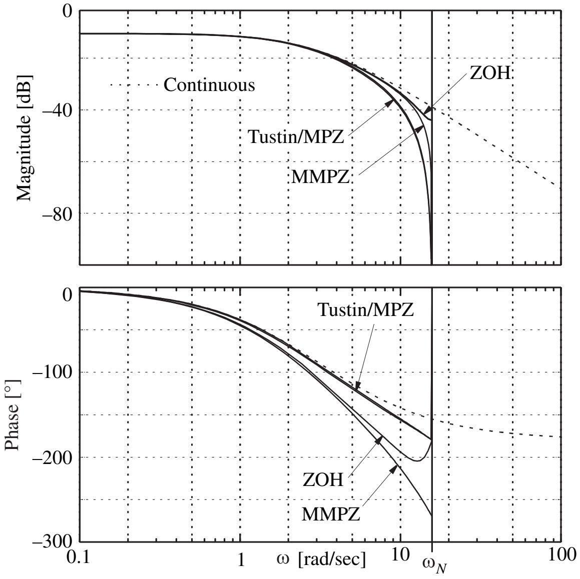


**Fig. D.21** Bode plots for the three discrete time equivalents

The ZOH transfer function and the MMPZ equivalent exhibit more phase lag than the Tustin and MPZ equivalents. This could also be anticipated having Figs. D.19 and D.20 in mind. So the suitable computation properties of ZOH and MMPZ are paid for by excess phase lag. ❒

## D.11 Discrete Equivalents to Continuous Controllers

It is simple to compute discrete equivalents to the well known standard continuous controllers (series compensators) for SISO feedback systems.

The most important controllers are:

1. PI-controller:

$$G_{PI}(s) = \frac{m(s)}{e(s)} = K_p \frac{\tau_i s + 1}{\tau_i s}. \tag{D.60}$$

2. Lead-compensator:

$$G_{lead}(s) = K_p \frac{\tau_d s + 1}{\alpha \tau_d s + 1}. \tag{D.61}$$

3. PID-controller:

$$G_{PID}(s) = K_p \frac{\tau_i s + 1}{\tau_i s} \cdot \frac{\tau_d s + 1}{\alpha \tau_d s + 1} \tag{D.62}$$

or

$$G_{PID}(s) = K_p \left( 1 + \frac{1}{\tau_i s} + \frac{\tau_d s}{\gamma \tau_d s + 1} \right). \tag{D.63}$$

4. Miscellaneous filters (low pass, high pass, notch etc.):

$$G_f(s) = \frac{B(s)}{A(s)}, \tag{D.64}$$

where $B(s)$ and $A(s)$ are appropriate polynomials in $s$.

***Example D.11.* Use of Matlab for System Discretization**

Consider the PID-controller,

$$G_{PID}(s) = 5 \frac{8s + 1}{8s} \cdot \frac{4.2s + 1}{0.96s + 1}, \quad \alpha = 0.23. \tag{D.65}$$

The discrete equivalents can be found by application of Eq. (D.50) or the MPZ-rules on p. 536. In this case the rather tedious insertion work has been avoided by use of MATLAB's `c2d`-function (continuous-to-discrete conversion).

The Tustin equivalent is (for $T = 0.2$ s),

$$\begin{aligned} H_T(z) &= 5.063 \frac{z - 0.9753}{z - 1} \cdot \frac{4.057z - 3.868}{z - 0.8113} \\ &= 20.54 \frac{z^2 - 1.9288z + 0.9299}{z^2 - 1.8113z + 0.8113}. \end{aligned} \tag{D.66}$$

The step and the frequency responses are shown on Figs. D.22 and D.23. The discrete as well as the continuous responses are shown.

The discrete step response is seen to follow the continuous response nicely and the Bode plots hardly differ from each other except that the discrete response is limited by the Nyquist frequency $\omega_N$. ❒

***Example D.12.* Discrete SISO PID Controller Responses**

Figure D.24 shows a SISO control system with the continuous PID-controller (D.65) from Example D.11. As it will often be the case the lead-compensator is

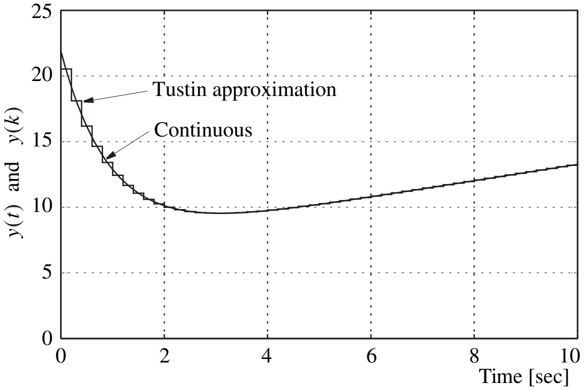


**Fig. D.22** Unit step responses for the PID-controllers

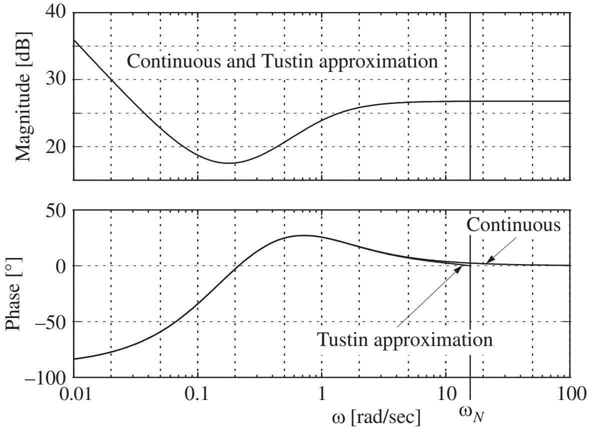


**Fig. D.23** Bode plots for the PID-controllers

placed in the feedback branch of the system. The system is fitted with a low pass prefilter to improve the step response. The controller is designed for a 48° phase margin. The crossover frequency $\omega_c = 0.43$ rad/s was obtained.

Figure D.25 shows the same system but with the discrete controller (D.66). The prefilter discrete Tustin equivalent can be found to be

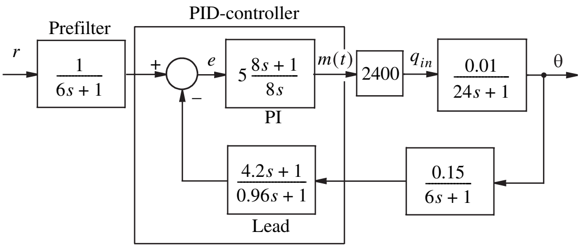


**Fig. D.24** Continuous control system

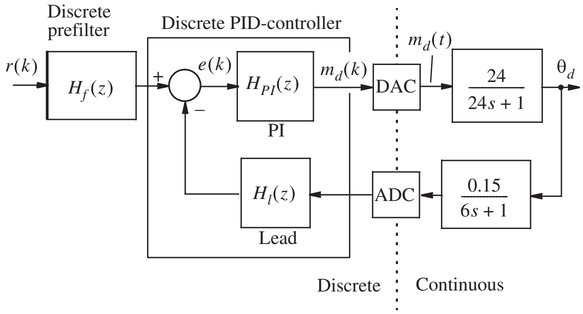


**Fig. D.25** Discrete version of the system on Fig. D.24

$$H_f(z) = 0.01693 \frac{z+1}{z-0.9672}.$$

Step responses and control signals for the two systems are found on Fig. D.26. The differences are minor. The main difference is that the control signal from the discrete controller is clearly a staircase signal.

This good result is partly due to the size of the sample period. The control object has the two time constants 24 s and 6 s. In this context the sample period $T = 0.2$ s is small in comparison. The consequence is that the control computer is unnecessarily burdened by the task of controlling this rather slow system. If the procedure is repeated with a more sensible sample period $T = 2$ s, the responses in Fig. D.27 are obtained.

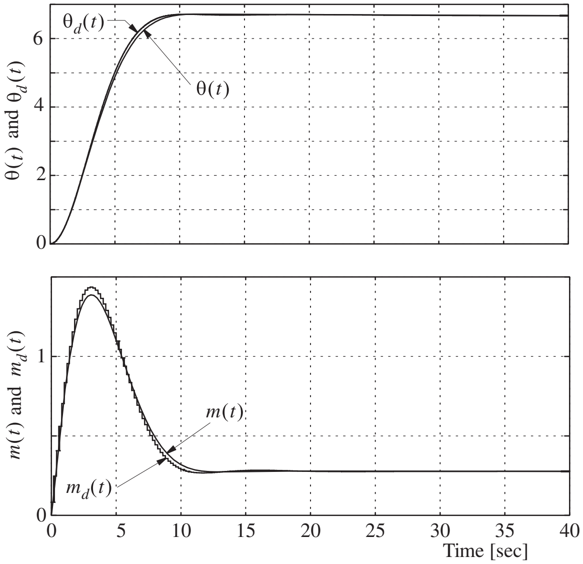


**Fig. D.26** Responses for continuous and discrete system ($T = 0.2$ s)

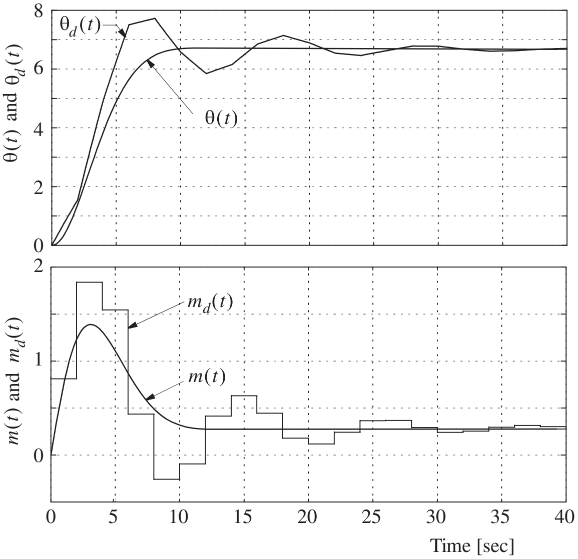


**Fig. D.27** Responses for continuous and discrete system ($T = 2\,\mathrm{s}$)

This result is obviously not good. With such a small sample frequency the effect of the sampling process is very severe. As it was pointed out previously (see Example D.7) the sampling causes a time delay and this is known to influence the stability margins negatively. By calculating discrete equivalents to continuous controllers such effects are ignored. ❐

It is not difficult to take the sampling process effects into account. It would be a reasonable first step to include the zero-order-hold into the continuous system. The ZOH has the transfer function (D.9) where a time delay is directly represented by the term $e^{-Ts}$. It is usually preferred to insert a rational function instead of the non-rational exponential function. Just like the case of the Tustin equivalent, a first order Padé approximant can be used. This leads to the transfer function:

$$G_h(s) = \frac{1 - e^{-Ts}}{s} \cong \frac{1 - \frac{1 - Ts/2}{1 + Ts/2}}{s} = \frac{T}{Ts/2 + 1}. \tag{D.67}$$

This transfer function has a stationary gain $T$. Thus if one desires to insert the function into the continuous system without changing the stationary loop gain, the factor $1/T$ must be included. The following transfer function is the overall result:

$$G_{ZOH}(s) = \frac{T}{Ts/2 + 1}. \tag{D.68}$$

With this transfer function as a part of the open loop continuous system it is possible to redesign the continuous PID-controller and obtain new controller parameters more suitable for the discrete time conversion by the Tustin or the MPZ method.

If an anti-aliasing filter is necessary, it will also be wise to include this as a new transfer function in the open loop before the redesign. The same could be the case if a substantial computation delay must be anticipated. Such a delay can be approximated by a new Padé approximant. ❒

***Example D.13.* Zeroth Order Hold Redesign of a SISO PI Controller**

If the ZOH approximation derived above is added to the continuous system in Example D.12, the feed forward branch will be as seen on Fig. D.28.

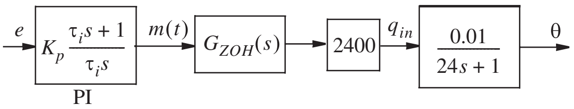


**Fig. D.28** Insertion of $G_{ZOH}(s)$ into continuous system

The redesign is carried out in the same way as the original design and the following continuous controller is obtained,

$$G_c(s) = 1.8\frac{11s+1}{11s}\cdot\frac{7s+1}{1.12s+1}. \tag{D.69}$$

The phase margin is considerably higher than before: 66°. The crossover frequency is smaller: 0.3 rad/s.

The Tustin equivalent for $T = 2$ s is:

$$\begin{aligned} H_T(s) &= 1.9636\frac{z-0.8334}{z-1}\cdot\frac{3.774z-2.8302}{z-0.0566} \\ &= 7.41\frac{z^2-5833z+0.625}{z^2-1.0566z+0.05661} \end{aligned} \tag{D.70}$$

The best prefilter is in this case:

$$G_f(s) = \frac{1}{3.5s+1} \Leftrightarrow H_f(z) = 0.2222\frac{z+1}{z-0.5556} \tag{D.71}$$

The responses are shown on Fig. D.29. The oscillations seen on the previous figure have disappeared and the step response has only a minor overshoot. The settling time is slightly larger than for the system with the continuous controller, an effect of the smaller crossover frequency and bandwidth.

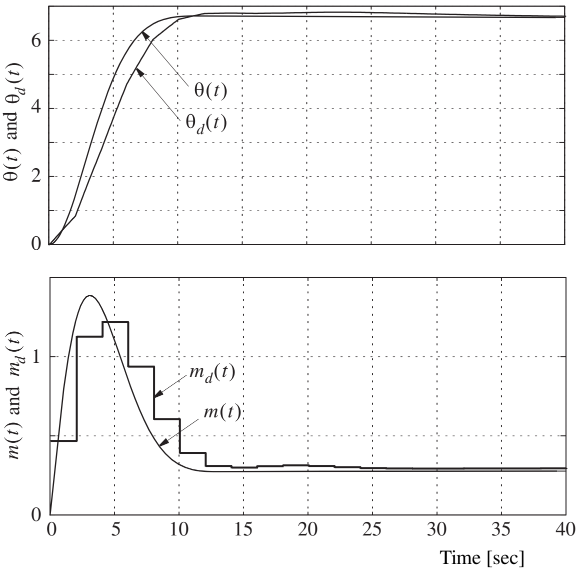


**Fig. D.29** Responses after redesign ($T = 2$ s)

Although the α-factor is smaller than before the control signal amplitude is actually smaller during this response. ❐

## D.11.1 Choice of Sampling Period

Although it is not readily apparent, the sampling period in most of the examples in this appendix (and, as a matter of fact, in the rest of the book) have been selected by an iterative trial-and-error method. Such methods are probably the preferable in most practical cases.

Shannon's sample theorem (see Eq. (D.7) on p. 519) provides a theoretical hint to the proper size of the sampling period but the Nyquist frequency is rarely well defined because virtually no signal is strictly band limited.

As a more useful first step in the sample period selection an empirical rule is usually used. There are several to choose from:

1. If is desired to design closed loop control for a process which has a dominant time constant $\tau_d$, then the sampling period $T$ should be chosen such that

$$T < \frac{\tau_d}{10}.$$

2. It is desired to have a closed loop settling time $t_s$, $T$ is required to satisfy

$$T < \frac{\tau_s}{10}.$$

3. Similarly, if the bandwidth $\omega_b$ is desired, $T$ should satisfy

$$T < \frac{2\pi}{10\omega_b} \text{(i.e., } \omega_s > 10\omega_b\text{)}.$$

   If the closed loop system has a dominant natural frequency $\omega_n$, this frequency can be inserted in the formula above instead of $\omega_b$.
4. If the process has a dominating time delay $T_d$, a reasonable requirement could be

$$T < \frac{\tau_d}{4}.$$

Such rules should only be considered as coarse guidelines allowing the designer to make a reasonable initial choice of the sampling period.

# References

Åström, K. J., *Reglerteori*, Almquist og Wiksell Forlag AB, Stockholm, Sweden, 1968, pp. 14–30.

Åström, K. J., Introduction to Stochastic Control Theory, Academic Press, New York, NY, 1970.

Åström, K. J. and Wittenmark, B., *Computer Controlled Systems*, Prentice-Hall, Inc., Engelwood Cliffs, NJ, 1984.

Bell, F., Johnson, E., Whittaker, R., and Wilcox, R., "Head Positioning in a Large Disc Drive", Hewlett-Packard Jour., Vol. 35, No. 1, January, 1984, pp. 14–20.

Bellman, R. and Kalaba, R., *Selected Papers on Mathmatical Trends in Control Theory*, Dover Publications, Inc., New York, NY, 1964.

Bennett, S., *A History of Control Engineering 1930–1955*, IEE Control Engineering Series 47, Peter Peregrinus Ltd., London, UK, 1993.

Berstein, D., "Feedback Control: An Invisible Thread in the History of Technology", IEEE Control Systems Magazine, Vol. 22, No. 2, April, 2002, pp. 53–68.

Brauer, F. and Nohel, J. A., *The Qualitative Theory of Ordinary Differential Equations*, W.A. Benjamin Inc., 1969.

Bryson, A., and Ho, Yu-Chi, Applied Optimal Control, Hemisphere Publishing Corporation, New York, Ny, 1975.

Cramer, H., *Mathematical Methods of Statistics*, Hugo Gebers Forlag, Uppsala, Sweden, 1945, also Princeton University Press, Princeton, NJ, 1946.

Doob, J. L., *Stochastic Processes*, John Wiley and Sons, Inc., New York, NY, 1953.

Elmquist, H., *SIMNON User's Manual*, Report 7502, Dept. of Auto. Control, Lund Inst. of Tech., Lund, Sweden, 1975.

Franklin, G. F., Powell, J. D. and Workman, M. L., *Digital Control of Dynamic Systems,* 2nd Ed.*,* Addison-Wesley, Reading, MA, 1990.

Friedland, B., *Control System Design*. McGraw-Hill Book Co., New York, NY, 1987.

Fuller, A., "The Early Development of Control Theory", ASME Jour. of Dyn. Systs., Meas., and Cont., Vol. 98 G, June, 1976a, pp. 109–118.

Fuller, A., "The Early Development of Control Theory II", ASME Jour. of Dyn. Systs., Meas., and Cont., Vol. 28 G, September, 1976b, pp. 224–235.

Gelb, A., ed., *Applied Optimal Estimation*, The M. I. T. Press, Cambridge, MA, 1974.

Golub, G. H. and Van Loan, C. F., *Matrix Computations*. The John Hopkins University Press, 1993.

Gran, R., Numerical Computing with Simulink, Vol. I, 2007.

Hendricks, E., Holst, J., and Hansen, H., "Identification of an Optical Fiber Pulling Process", Proc. 7th IFAC Symp. on Identification and System Parameter Estimation, York, UK, July, 1985.

Jacobs, O. L. R., *Introduction to Control Theory*, Oxford University Press, Oxford, UK, 1993.

Jazwinski, A. H., *Stochastic Processes and Filtering Theory*, Academic Press, London, UK, 1970.

Kailath, T., *Linear Systems*. Prentice-Hall, Inc., Engelwood Cliffs, NJ, 1980.

Kalman, R. E., "A New Approach to Linear Filtering and Prediction Theory", Trans. Amer. Soc. of Mech. Eng., Series D, Jour. of Basic Eng., Vol. 82, 1960, pp. 35–45.

Kalman, R. E. and Bertram, J. E., *General Synthesis Procedure for Computer Control of Single and Multiloop Linear Systems*. Trans. AIEE, Vol. 77, 1958.

Kalman, R. E., and Bucy, R. S., "New Results in Linear and Prediction Theory", Trans. Amer. Soc. of Mech. Eng., Series D, Jour. of Basic Eng., Vol. 83, 1961, pp. 95–108.

Kautsky, J., Nichols, N. K. and Van Dooren, P., "Robust Pole Assignment in Linear State Feedback*",* Inter. Jour. of Control, Vol. 41, No. 5, 1985.

Kolmogorov, A. N., "Interpolation und Extrapolation von Stationaren Zalligen Folgen", Bull. of the Academy of Sciences, USSR, Mathematical Series, Vol. 1, p. 3, 1941.

Kushner, H. J., "Dynamical Equations for Optimal Nonlinear Filtering", Jour. of Diff. Eqns., Vol. 3, April, 1967, pp. 179–190.

Kwakernaak, H. and Sivan, R., *Linear Optimal Control Systems,* John Wiley and Sons, New York, NY, 1992.

Lewis, F. L., *Optimal Control*. John Wiley & Sons, New York, NY, 1986.

Lewis, F. L., *Applied Optimal Control and Estimation*, Prentice Hall, Englewood Cliffs, NJ, 1992.

Luenberger, D. G., *Observing the State of a Linear System*. IEEE Trans. on Military Electronics, Vol. MIL-8, April, 1964.

Luenberger, D. G., *Observers for Multivariable System*s. IEEE Trans. on Automatic Control, Vol. AC-11, No. 2, April, 1966.

Luenberger, D. G., "An Introduction to Observers", IEEE Trans. on Automatic Control, Vol. AC-16, No. 6, December, 1971.

Maybeck, P. S., *Stochastic Models, Estimation, and Control*, Vol. 1, 2, 3, Academic Press, Inc., New York, NY, 1979.

McGarthy, T. P., *Stochastic Systems and State Estimation*, John Wiley and Sons, Inc., New York, NY, 1974.

Meditch, J. S., *Stochastic Optimal Linear Estimation and Control*, McGraw-Hill, Inc., New York, NY, 1969.

Middleton, R. H. and Goodwin, G. C., *Digital Control and Estimation*, Prentice-Hall, Engelwood Cliffs, NJ, 1990.

Papoulis, A., Signal Analysis, Mc-Graw-Hill, Inc., New York, NY, 1977.

Rugh, W. J., *Linear System Theory*, 2nd. Ed., Prentice Hall, Upper Saddle River, NJ, 1996.

Safonov, M., and Athens, M., "Robustness and Computational Aspects of Nonlinear Stochastic Estimators and Regulators", Paper No. FP7-3:00, 1977 American Control Conference (1977 ACC), New Orleans LA, December, 1977.

Safonov, M., and Athens, M., "Robustness and Computational Aspects of Nonlinear Stochastic Estimators and Regulators", IEEE Tran. on Automatic Control, Vol. AC-23, No. 4, August, 1978, pp. 717–725.

Schwartz, L., "Nonlinear Filtering and Comparison with Kalman Filtering", AGARDograph, No. 139, 1970, p. 143.

Soong, T. T., *Random Differential Equations in Science and Engineering*, Academic Press, Inc., New York, NY, 1973.

Sorenson, H. W., and Stubberud, A. R., *Linear Estimation Theory*, AGARDOgraph, No. 139, 1970, p. 1.

Stengel, R., *Stochastic Optimal Control*, John Wiley and Sons, New York, NY, 1986.

Stratonovich, R. L., "Conditional Markov Processes", Theory of Probability and Its Applications, Vol. 5, No. 2, 1960, pp. 156–178.

Swerling, P., "First Order Error Propagation in a Stage-Wise Smoothing Procedure for Satellite Observations", Jour. of Astro. Sciences, Vol. 6, p. 46, 1959.

Vidyasagar, M., *Nonlinear Systems Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

Wiener, N., *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*, Wiley, Inc., New York, NY, 1949.

# Index

Elbert Hendricks
Ole Jannerup
Paul Haase Sørensen

# Linear Systems Control

Modern control theory and in particular state space or state variable methods can be adapted to the description of many systems because they depend strongly on physical modelling and physical intuition. The laws of physics are in the form of continuous differential equations and for this reason, this book concentrates on system descriptions in this form: coupled sets of linear or nonlinear differential equations. The physical approach is emphasized in this book because it is most natural for complex systems. It also makes it easier to immediately apply the theory to the understanding and control of many different types of systems.

In line with the approach set forth above, the book first deals with system modelling in state space as well as transfer function form. The modelling methods are described with many examples. Linearization is treated in detail. Because computer control is so fundamental to modern applications, discrete time modelling of systems as difference equations is introduced immediately after the more intuitive differential and transfer function models. Many control schemes, based on linearized state space models, are treated in the deterministic as well as in the stochastic case. These control methods include Linear Quadratic (LQ) optimal control and LQG (Linear Quadratic Gaussian) control (LQ control with Kalman filters).

ISBN 978-3-540-78485-2

9 783540 784852

❯ springer.com